ePIC CERN Software & Computing Meeting Report

Community Building

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

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

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

Monday: Reviewing Status and Future Plans

DALL-E, could you please draw a picture of exciting discussions at CERN?

- May 1 marked the first anniversary of the WG structure and conveners being endorsed by the Collaboration Council.
- Over the past year, we have addressed many time-critical items.
- Now, as we approach the completion of the first year, it is crucial to consider the broader perspective:
  - Is the WG charge still accurate?
  - What about the priorities for the second year?
- Therefore, we dedicated the first day of the meeting to reviewing the status and future plans of each of the WGs.

Physics and Detector Simulation WG: Year One Highlights

- **Key achievements and milestones.**
  - Geometry update and detailed DD4hep implementation for all sub-detectors - in agreement with geometry database
  - Collected github contacts for all DSCs for sign-off on these and other PRs
  - Like the thresholds from the digitization spreadsheet
  - Initial support and service structure implementation
  - Automated tracking material map generation and checks

- **Major impacts and successes.**
  - Over 150 merged PRs and 40 closed issues since 2023/07
  - Automated tracking material map generation (#677)
  - Detailed DD4Hep description of pfRICH (#664)
  - Implementation of inner barrel support cylinder (#659)
  - Support structure for Barrel Imaging Calorimeter (#710, #608)
  - Event generator campaigns, naming, workflow documentation brought into compliance with production WG requirements
  - Event + Background merging at the HepMC level implemented
    - Arbitrarily long “events” → time frames
    - Synchrotron radiation is getting a major upgrade to a standalone Geant4 version, getting rid of SynRad and allowing HepMC treatment just as the other sources
  - And many geometry updates, bug fixes, and code improvements (check github PRs)

Reconstruction Framework and Algorithms WG: Year One Highlights

Key achievements and milestones:

- Developed a full chain of ACTS-based track reconstruction with auto material map generation, realistic seed finder, combinatorial track finding/fitting, ambiguity solver, and primary vertexing.
- Implemented initial prototype of electron-finder, which includes the implementation of track-calorimeter projections
- Made progress towards the implementation of particle flow in ePIC
- Made significant improvements across EICrecon (e.g. in the jet software)
- **Major updates to JANA2**: integrating cleanly with the PODIO-based data models and other layers of the key4hep stack, enabling external configuration of existing components, and supporting timeframe splitting for streaming readout.

Major impacts and successes:

- Enabled tracking performance study with realistic event and beam background.
- Supported internal review of tracking detectors

Production WG: Year One Highlights

Highlights from Year 1

Central production on the Open Science Grid (OSG).

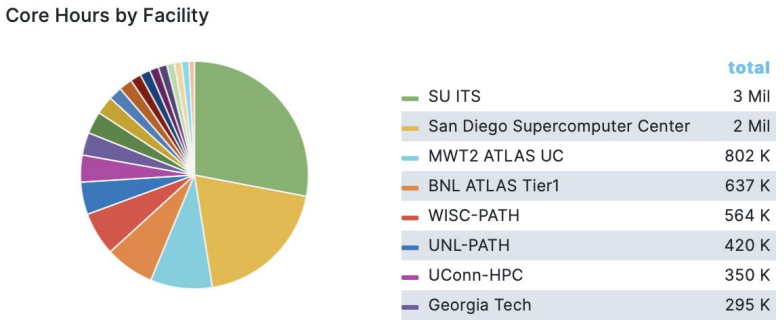
- Production workflow successfully running since May 2023. Centralized submission through JLAB submit host since August 2023.
- OSG submit host used as test bench.
- ~280 TB of data relevant to production: RECO: 196 TB, FULL: 78 TB, EVGEN: 9.8 TB.

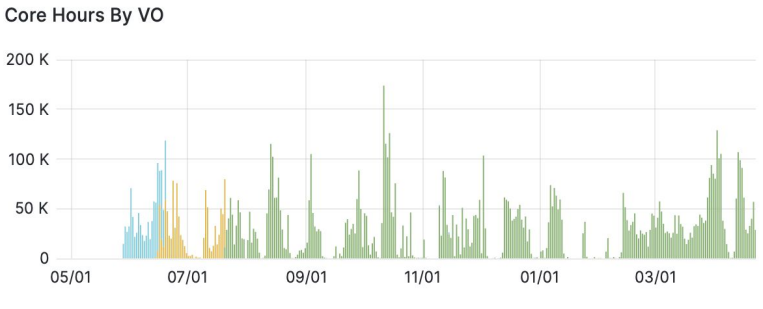
Rucio Test Instance

- Rucio instance set up at JLab
- Successful tests of cataloguing and file access
- Drafting a Rucio naming scheme proposal

Simulation Production Links

- Live campaign updates
- Input Preprocessing Policy
- Default List of Datasets (Still being updated)
- File access instructions



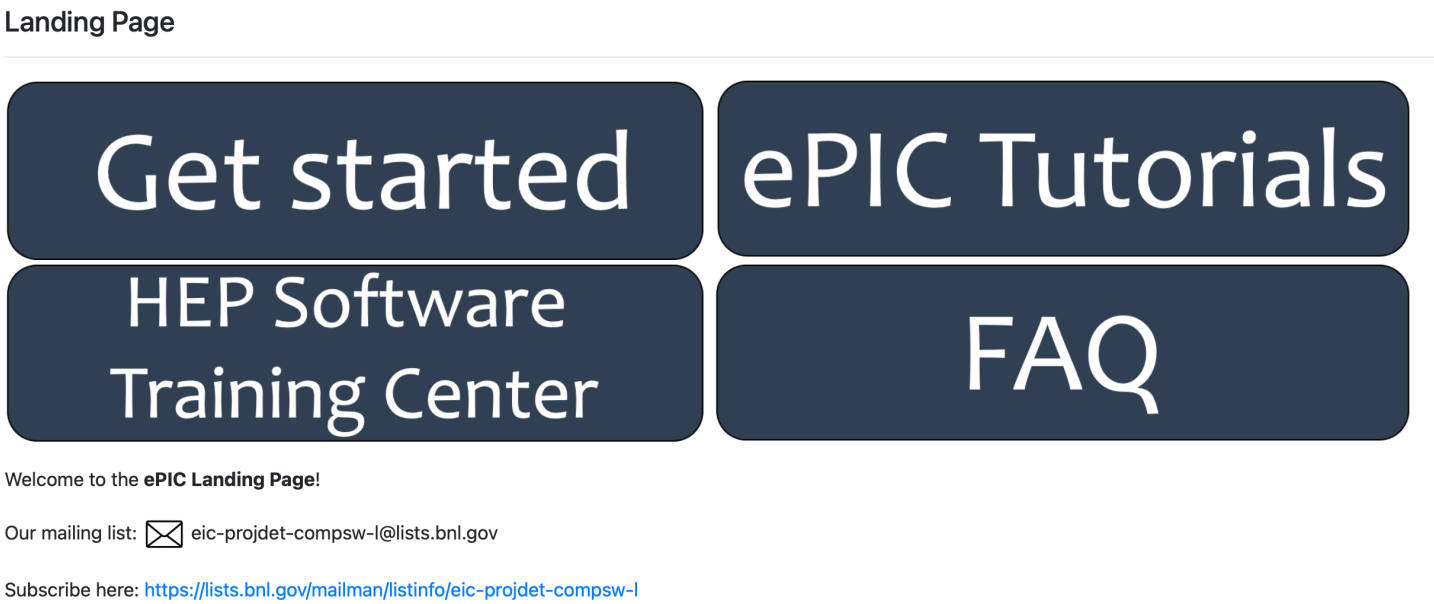


3

User Learning WG: Year One Highlights

Landing Page for Onboarding New Users:

- This page includes a continuously updated and improved list of **useful links**, **software tutorials**, and **frequently asked questions**.
- Any member of the collaboration can **directly contribute by submitting change requests**.



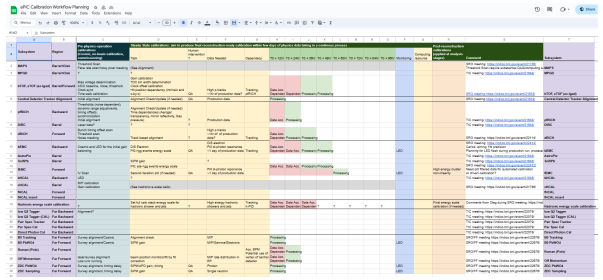
- Lots of new documentation accessible to users in mostly uniform format (tutorials and FAQ)
- HelpDesk on Mattermost continues to be an active site of information exchange and help.

Validation WG: Year One Highlights

- Partial adoption of Snakemake-based workflows showed that the software, indeed, meets our current needs
- Lots of improvements for development support by benchmarking
  - More tests running as GitHub actions for epic and EICrecon
  - Full container build is now triggered for eicweb EICrecon
  - Capybara reports are well integrated into the PR review process
- Initial tutorial on benchmark development: https://github.com/eic/tutorial-developing-benchmarks
- New detector benchmarks: “backgrounds”, “ecal_gaps”, “material_scan”
- New physics benchmarks: “diffractive_vm”

Streaming Computing WG: Year One Highlights

- May 1, 2023 – WG established
- July 11, 2023 – Kickoff Meeting
- Oct 2023 – ePIC Software and Computing Review
- Dec 2023 – EIC Resource Review Board
- Map detector calibration's, logic, and dependencies



https://docs.google.com/spreadsheets/d/1Rb5Cdjthkfh4IWWxFwHgiHNtMs5cDxcahTTUN394cJ8/edit#gid=0

The ePIC Streaming Computing Model

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

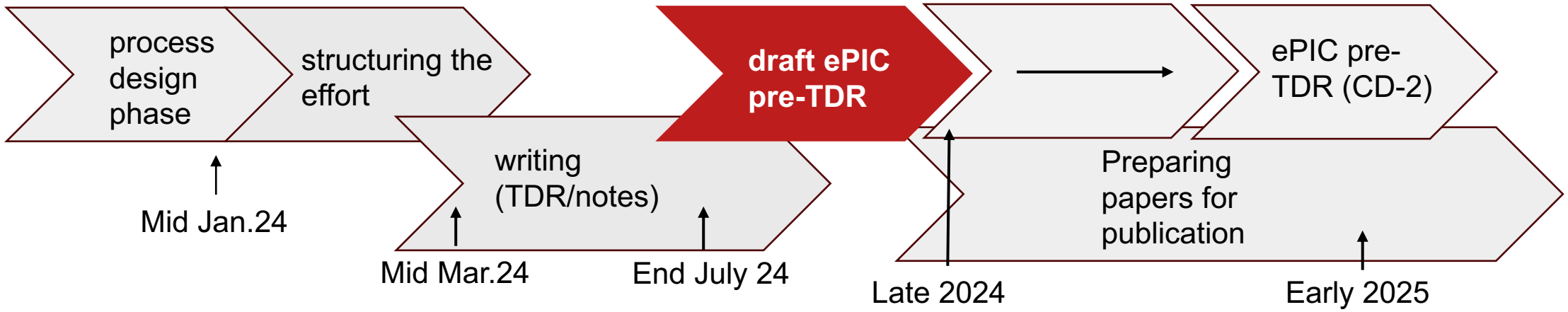
Abstract

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

https://indico.bnl.gov/event/20481/attachments/49818/86296/ePIC-StreamingComputingModel.pdf

- Collaboration with Electronics & DAQ towards defining contents, parameters, & handling of time frames
  - Length ~<= 2^16 bunch crossings: (Max time ~ .6ms, Max size ~ 10MB)
  - Contents: (Detector data built into single buffer per time frame)
  - Infrastructure: (ex. JANA2 mechanisms for splitting timeframes to discrete events)
- Strategy for demonstrating streaming reconstruction
  - TDR – separate analysis of streaming tracking based on "merged" timeframes
  - Post TDR – incorporate streaming digitization into simulation

Tuesday: TDR Strategy and Publications



- ePIC Software & Computing is essential to the TDR, providing **advanced software** and **simulation productions** that are the **input for detector and physics studies**:
  - ***"Software and Simulation Readiness for TDR"*** parallel session and plenary discussion at the ANL collaboration meeting, where we extensively defined the remaining development tasks, drawing on significant input from the collaboration at large.
  - Good progress since collaboration meeting, improving the accuracy of the simulations and building up the reconstruction in shared priorities with DSCs and the PWGs.

- A **publication** on the ***"ePIC Streaming Computing Model"*** is planned for CSBS, based on the existing report.

Priorities for Discussion Identified in April 10 Meeting

- **Data and Analysis Preservation for TDR** (follow up discussion on May 29 after meeting technical editors)
- **Closing out Data Model Review**:
  - Policy on MC Truth Associations: Proposal (follow up discussion on May 29).
  - Examples for Data Model Usage. Excellent tutorial by Wouter with examples in C++ and Phyton:

One-to-One Relations

Relations are references to one objects in another collection.

```
edm4hep::SimTrackerHit:
  OneToOneRelations:
    - edm4hep::MCParticle MCParticle
      // (‘particle’ in edm4hep v1.0)
```

E.g. `B0TrackerHits` contains `edm4hep::SimTrackerHit`, and the one-to-one relations is in:
`_B0TrackerHits_MCParticle.collectionID`
`_B0TrackerHits_MCParticle.index`

Transparent access when using `podio` tools.

C++

```
#include <edm4hep/SimTrackerHitCollection.h>
#include <podio/Frame.h>
#include <podio/ROOTFrameReader.h>

podio::ROOTFrameReader r;
r.openFile(argv[1]);
auto f =
podio::Frame(r.readNextEntry(podio::Category:
:Event));
auto& h =
f.get<edm4hep::SimTrackerHitCollection>("B0Tr
ackerHits");
std::cout <<
h.at(0).getMCParticle().getPDG();
```

One-to-Many Relations

Relations are references to multiple objects in other collections. This requires an intermediate table.

```
edm4hep::MCParticle:
    OneToManyRelations:
      - edm4hep::MCParticle daughters
```

All different `daughters` are referenced in `_MCParticles_daughters`, as for one-to-one relations.

`MCParticles.daughters_begin` and `MCParticles.daughters_end` indicate range of references that should be used.

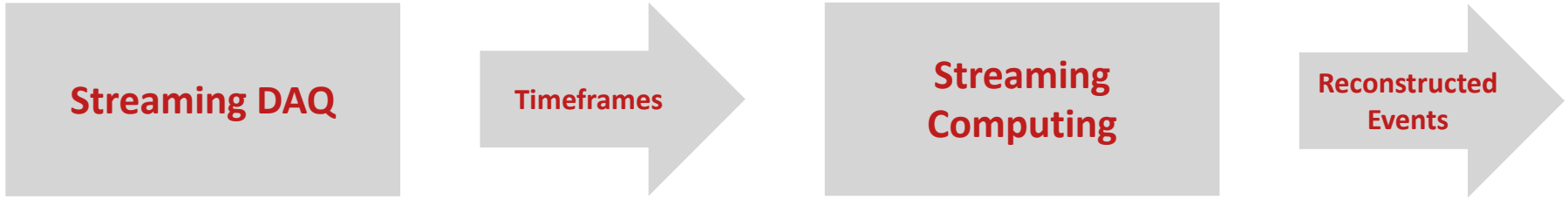
Python

```
import edm4hep
from podio import Frame
from podio.reading import get_reader

reader = get_reader("file.edm4hep.root")
frames = reader.get("events")
frame = frames[0]
p = frame.get("MCParticles")
d = p[0].getDaughters()
print(d.size())


auto& d = p.at(0).getDaughers();
std::cout << d.size();
```

- **Reconstruction**:
  - TDR Progress: Calorimeter Reconstruction
  - TDR Progress: Topical ACTS Meeting on ePIC Tracking, jointly with ACTS team.
  - Streaming Reconstruction Prototype

Priority Target for the TDR

Prototype of event reconstruction from realistic frames:



Purpose: Demonstrate that we can reconstruct events from Streaming DAQ.

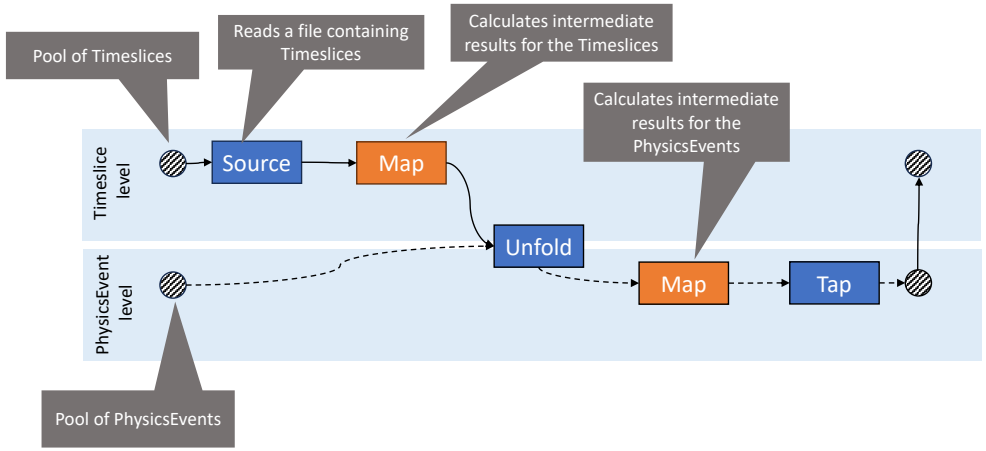
Purpose: Estimation of streaming reconstruction time for compute resource planning.

- **Follows up on ECSAC recommendation**: *"We recommend that ePIC document a first computing needs assessment by the next ECSAC review, in roughly one year."*
- We understand the event rates and data sizes but **require** a **reliable estimate for streaming data processing, a unique aspect of the ePIC experiment**.

Prototype of Event Reconstruction from Realistic Frames

Key Tasks: We limit the scope of the first prototype to the track reconstruction only. The key is to demonstrate we can correlate hits in a realistic time frame of 1ms to the various events in the time window of the MAPS of 2µs.

- Reached consensus on composition of *realistic* time frames in terms of signal and background.
- Implemented and utilized the infrastructure for building timeframes instead of events in detector simulations.
- Major update to JANA2 framework for processing of timeframes, events, and subevents:



- **Work in Progress**:
  - Adapt EICrecon to work with timeframes:
    - Test EICrecon with new JANA2 version.
    - Prototype for an Unfold algorithm.
  - Demonstrate tracking from realistic frames.

Wednesday: Software Tutorials

1. Overview of ePIC Software (Holly Szumila-Vance, Jefferson Lab)

   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.

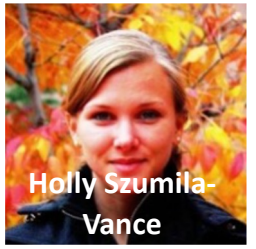
2. Working with Simulation Output (Stephen Kay, University of York)

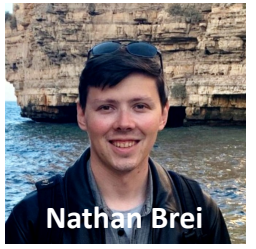
3. Simulating Detectors and Their Readout (Simon Gardner, University of Glasgow)

4. Reconstruction Algorithms (Nathan Brei, Jefferson Lab)

45 participants from Africa, Asia, and Europe.

Material available on Landing Page. Recordings available on: YouTube channel of the EICUG





Thursday: Streaming Computing

- **Goal** **Rapid turnaround of 2-3 weeks for data for physics analyses**.
- **Solution** Compute-detector integration using:

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

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

Heterogeneous computing for acceleration.

Discussion Items:

- (Near) Real-time processing for streaming data. Lessons learned from LHCb Allen.
- **Distributed computing**: Scientific data and workflow and workload management. Reports from Rucio and PanDA teams.
- **EICrecon Roadmap**: How should EICrecon evolve? What are the incremental steps needed to achieve this evolution?
- **Related to that**: Configuration management, calibrations and conditions databases.

EICrecon Evolution Roadmap

Slide from Derek

Discussion Topics

- ☑ Tracking: transition to full podio data model support
- ☑ Development experience
- ○ Timeslices [top priority]:
  - Fold/Unfold algorithms
  - Allowing both timeslices & physics events as input
- ○ Modularity:
  - ☑ Algorithms [need further offline discussion]
  - Digitization
- ○ Configuration & Orchestration [top priority]:
  - External algorithm wiring
  - Conditions database & configuration management
- ○ Performance enhancers:
  - Multithreading
  - Input batching, buffered processing [Not top priority: schedule June discussion in Reco WG]
  - GPU and other accelerator integration [cross-cutting]
  - Multi-architecture support (ARM) [cross-cutting]
- ○ AI/ML Integration [2nd highest priority]
- ○ Integration w/ object storage, remote storage

Status

- Discussed at April CERN Workshop (need to revisit)
- Discussed/touched on at 05.07.2024 Reco WG Meeting
- Will discuss in Weekly S&C Meeting
- Will be joint discussion with Simulation WG

Friday: Community Software and Common Projects

Note, many already left by Friday.

Discussion with scientists from CERN EP-SFT:

- Community software (HSF) and common projects (Stacks / key4hep, AI in production).
- How to enhance existing collaboration? First step: Write up on common work.

ePIC Software & Computing at CERN

Thank you so much to Yasemin Altinbilek and Pere Mato for making the meeting at CERN possible.

- Essential review of the **WG status and plans**, yielding numerous concrete action items that are currently being addressed.
- Critical review of **TRD tasks** and follow-up on priority items.
- **Software tutorials** with strong engagement from collaboration members in Africa, Asia, and Europe.
- Advancement of our activities in **streaming computing** and related to that distributed computing.
- Strengthen **collaboration with HEP**, specifically CERN EP-SFT.

Next in-person meeting foreseen for the fall.

DALL-E, could you please draw a picture of exciting discussions at CERN?