ePIC Streaming Computing Model

Markus Diefenthaler (Jefferson Lab) for ePIC Software & Computing

ePIC Streaming Computing Model

This presentation offers a glimpse into the comprehensive presentation of the ePIC Streaming Computing Model presented during the ePIC Software & Computing Review on October 19–20.

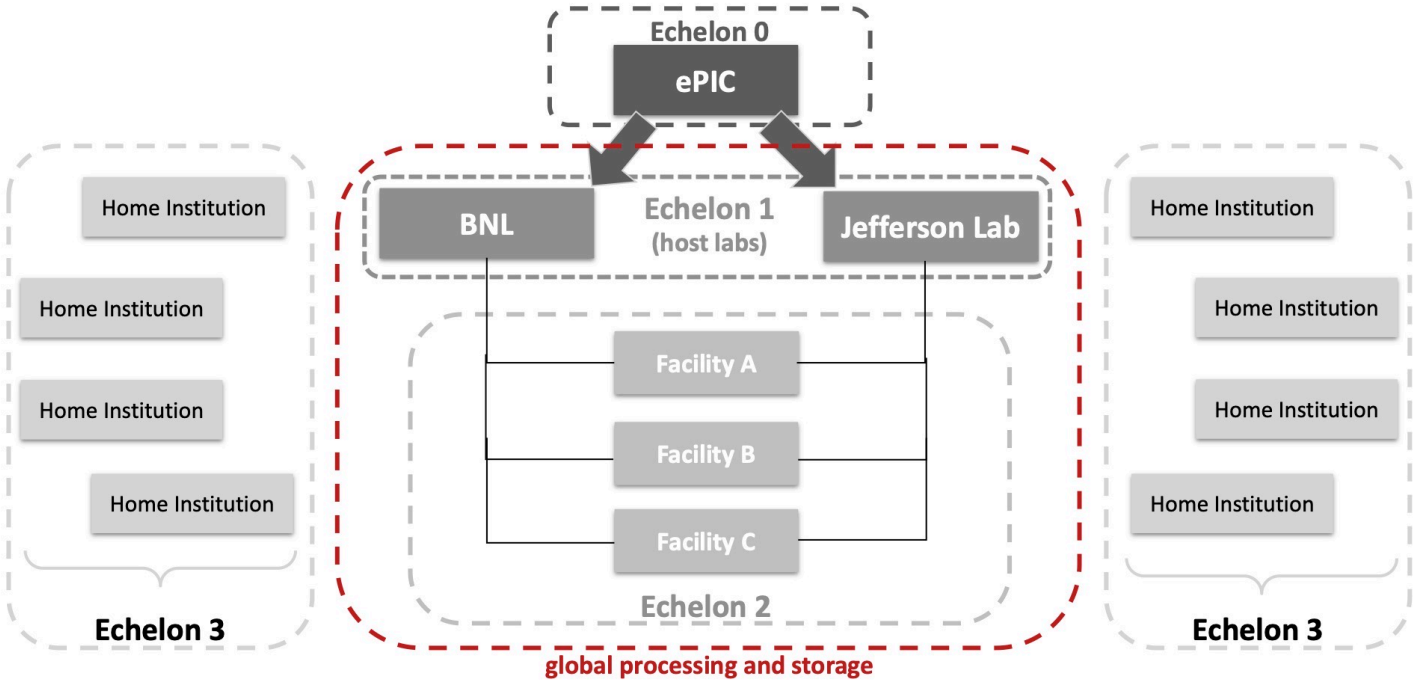
Four Tiers:

Echelon 0: ePIC Experiment

Echelon 1: Host Labs

Echelon 2: Global processing and data facilities, includes HPC and HTC resources.

Echelon 3: Home institute computing



Use Case to Echelon Mapping

Echelon 0: ePIC Experiment
Echelon 1: Host Labs
Echelon 2: Global processing and data facilities
Echelon 3: Home institute computing

Use Case	Echelon 0	Echelon 1	Echelon 2	Echelon 3
Stored Data Streaming and Monitoring	✓	✓		
Alignment and Calibration		✓	✓	
Prompt Reconstruction		✓		
First Full Reconstruction		✓	✓	
Reprocessing		✓	✓	
Simulation		✓	✓	
Analysis		✓	✓	✓
Modeling and Digital Twin		✓	✓	

Streaming Readout: Trigger-Less Data Acquisition

Definition of Streaming Readout

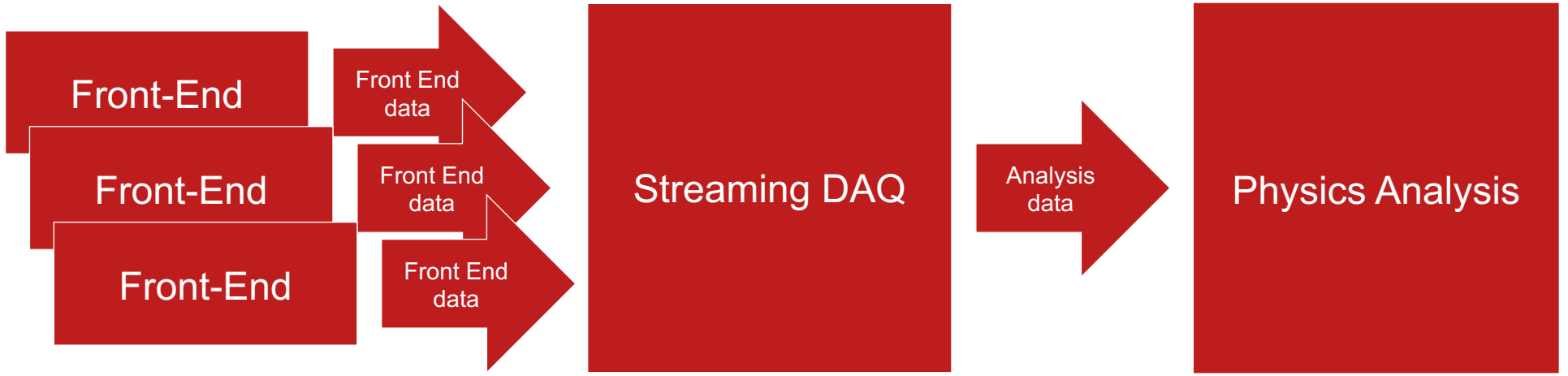
- Data is digitized at a fixed rate with thresholds and zero suppression applied locally.
- Data is read out in continuous parallel streams that are encoded with information about when and where the data was taken.
- Event building, filtering, monitoring, and other processing is deferred until the data is at rest in tiered storage.

Advantages of Streaming Readout

- Simplification of readout and increased flexibility (no custom trigger hardware and firmware).
- Continuous data flow provides detailed knowledge of backgrounds.
- Streamline workflows and take advantage of other emerging technologies:
  - AI for autonomous experimentation and control.
  - Reconstruction of physics events from holistic detector information.

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**.
- **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.



Streaming Data Processing for ePIC

Traditional Workflow Characteristics in NP and HEP Experiments:

- Data is acquired in online workflows.
- Data is stored as large files in hierarchical storage.
- Offline workflows process the data, often with substantial latency.
- Batch queue-based resource provisioning is typical.
- **Key features**: discrete, coarse-grained processing units (files and datasets) and decoupling from real-time DAQ.

Challenging Characteristics of Streaming Data Processing:

- **Time critical**, proceeding in near real time.
- **Data driven**, consuming a fine-grained and quasi-continuous data flow across parallel streams.
- **Adaptive and highly automated**, in being flexible and robust against dynamic changes in datataking patterns, resource availability and faults.
- **Inherently distributed** in its data sources and its processing resources.

Assumptions for Infrastructure:

- Existing batch-style processing likely to remain.
- Dynamic processing, e.g. Kubernetes, may displace the batch model.
- Design the system for both batch and dynamic processing to ensure resilience against technology evolution.
- Accommodate but effectively hide these underlying infrastructure characteristics.

Milestone for the TDR: Reconstruction from Frames

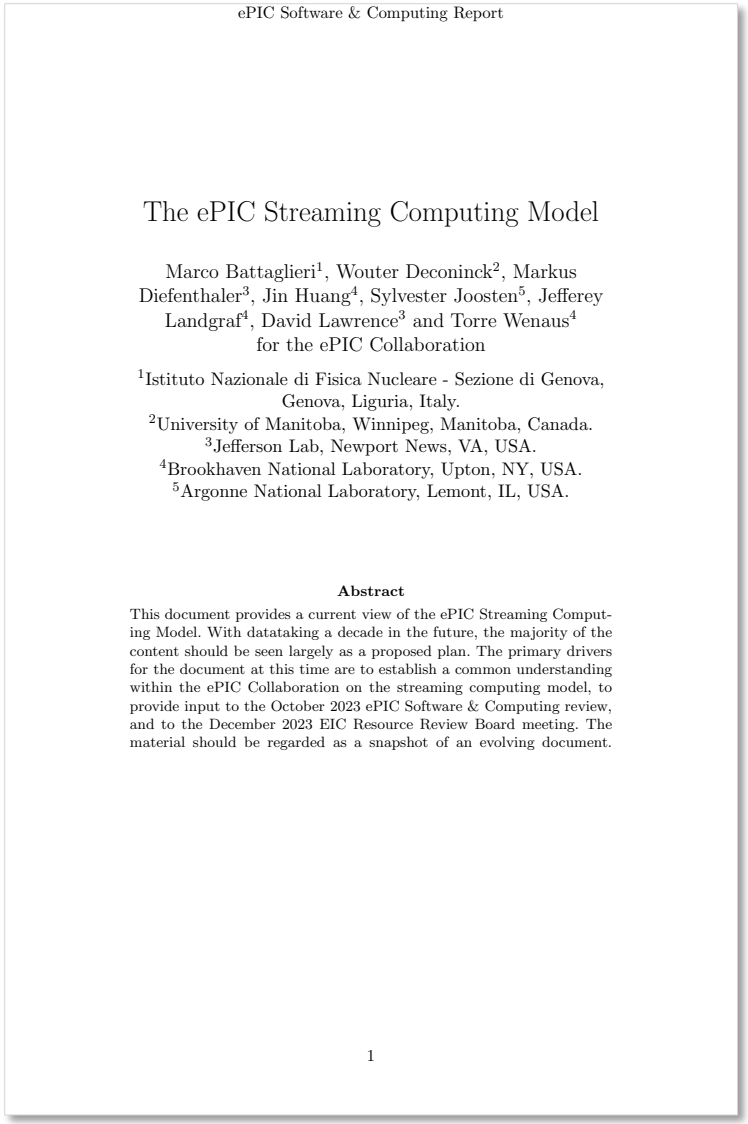
We limit the scope of the first study to the track reconstruction only.

The key is to demonstrate we can correlate hits in a realistic time frame to the various events in the time window of the MAPS.

We need to accomplish the following tasks (15 FTE months):

- **Framework Tests and Development**:
  - **01/24** Integrate Jana2's built-in workflow for supporting frames in and events out in EICrecon.
- **Simulation**:
  - **12/23** Prepare simulation productions, using detailed information on FEEs for tracking detectors, utilizing the full, wide MAPS integration window for tracking purposes.
  - **01/24** Implement and utilize the frame-building infrastructure post-Geant4 and post-digitization.
- **Reconstruction Process**:
  - **02/2024** Adapt the reconstruction process to work with frames, making it frames-aware.
  - **03/2024** Demonstrate tracking from realistic frames
  - **04/2024** Deliver first estimate of reconstruction time from frames.

Publication on “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

For the review material, we have submitted a report on the "ePIC Streaming Computing Model”.

Until November 10, you can submit comments via: https://forms.gle/RANT95trA7FZSAPB6

The link has a direct link to the report.

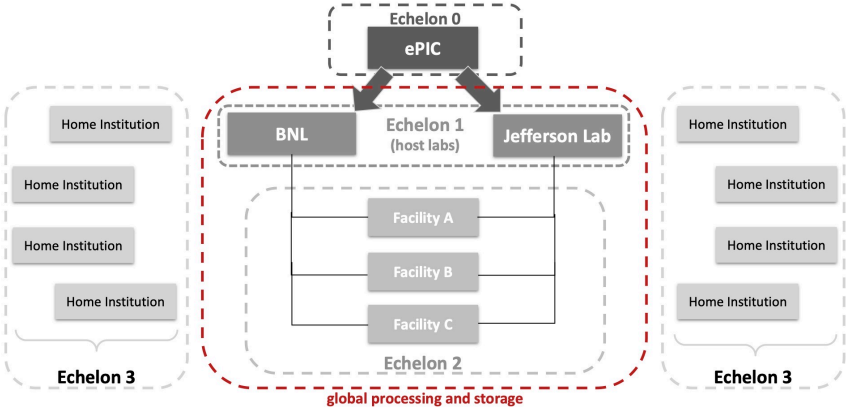
Using the findings from the review and feedback we receive from you, we will evolve the report into a publication on the "ePIC Streaming Computing Model”.

Realistic submission date:

- April or May, 2024.
- Driven by work on track reconstruction from frames.

Summary

- **Streaming Readout of the ePIC Detector to maximize and accelerate science:**
  - ePIC aims for **rapid turnaround of 2-3 weeks for data for physics analyses**.
  - Timeline driven by calibrations.
- **Four tiers of the ePIC Streaming Computing Model computing fabric:**
  - **Echelon 0**: ePIC experiment and its streaming readout.
  - **Echelon 1**: Crucial and innovative partnership between host labs.
  - **Echelon 2**: Essential global contributions.
  - **Echelon 3**: Full support of the analysis community.

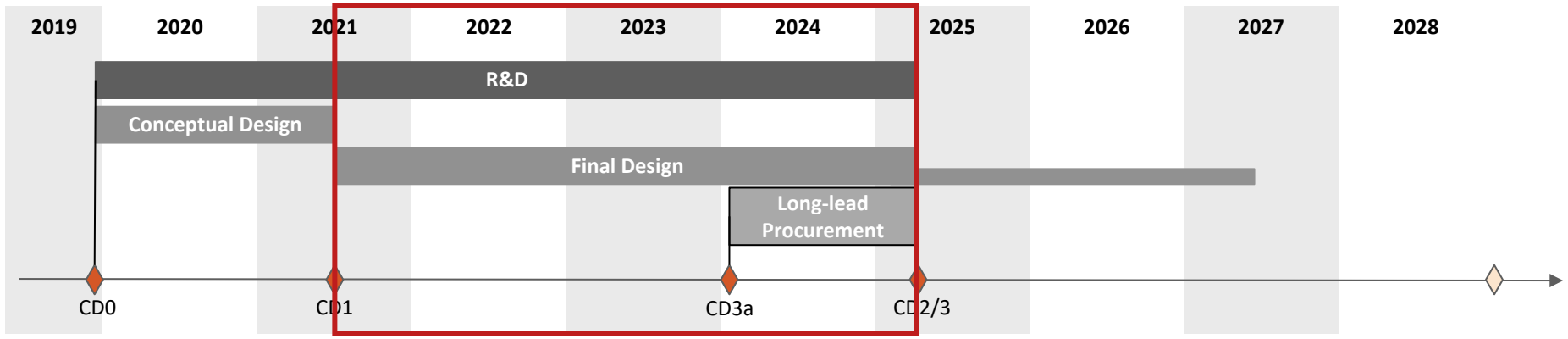


- **High level milestones** ensures that the agile development process is continuously confronted with real world exercising of the software and the developing realization of the computing model:
  - Priority always given to meeting near-term needs.
  - Longer range timeline progressively exercising 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 data taking.

Exciting opportunities to get involved in ePIC Software & Computing!

Backup

Milestones

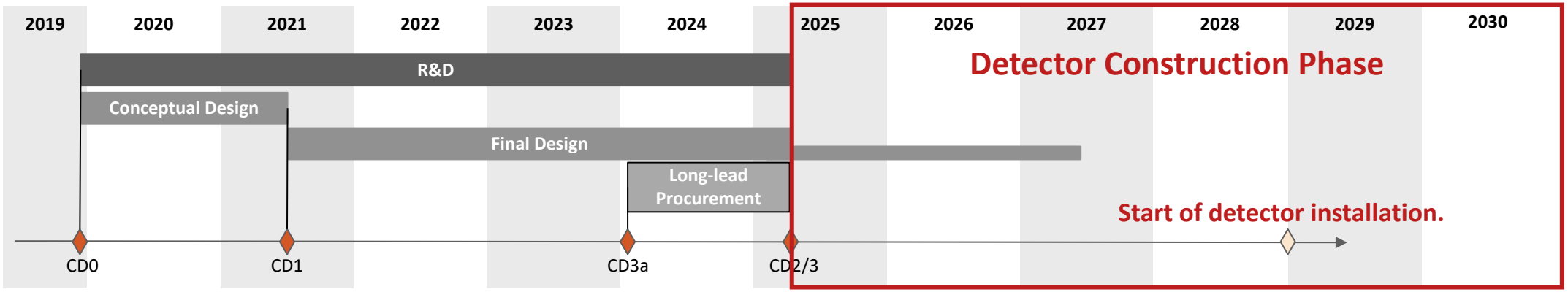


CD2/3 and TDR

Milestones Prior to CD2/3 and TDR

- Software and simulation readiness for TDR preparation (and subsequent phases of the CD process).
- Provide for each use case detailed estimates on the compute resources; update the networking and storage estimates according to format of streaming data format that is currently being defined.

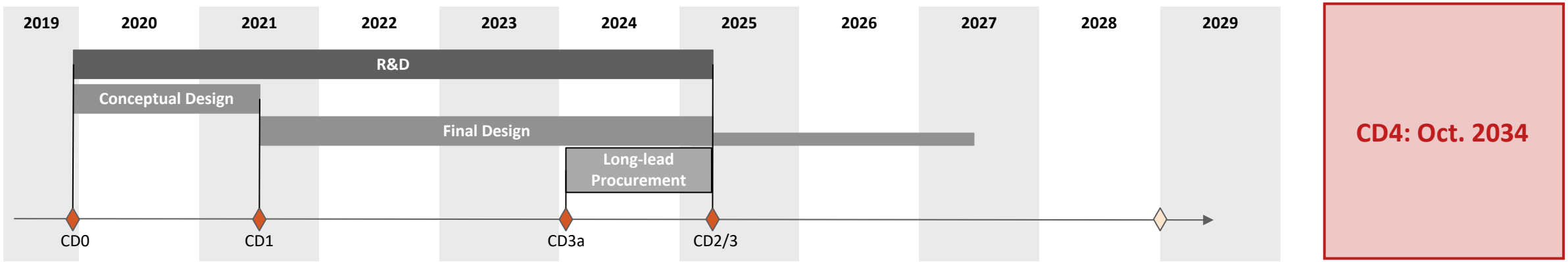
Milestones



Milestones During Detector Construction Phase

- Provisioning DAQ and software sufficient for test beams, which can serve as small scale real-world testbeds for the developing DAQ and software.
- Streaming challenges exercising the streaming workflows from DAQ through offline reconstruction, and the Echelon 0 and Echelon 1 computing and connectivity.
- Data challenges exercising scaling and capability tests as distributed ePIC computing resources at substantial scale reach the floor, including exercising the functional roles of the Echelon tiers, particularly Echelon 2, the globally distributed resources essential to meeting ePIC's computing requirements.
- Analysis challenges exercising autonomous alignment and calibrations.
- Analysis challenges exercising end-to-end workflows from (simulated) raw data to exercising the analysis model.

Milestones



Milestones During Detector Commissioning

- This phase has unique expectations and requirements compared to steady-state operation:
  - Utilization of semi-triggered data-taking modes.
  - Initial calibrations.
  - Gradual extension of first pass processing from Echelon 1 to Echelon 2.
- Careful planning of software & computing efforts and leveraging experience from data and analysis challenges during the detector construction phase essential.

Milestones during Early Datataking Phase

- Simpler and more conservative approaches will be used during initial data-taking phase.
- The ePIC Streaming Computing Model will be gradually deployed and validated.