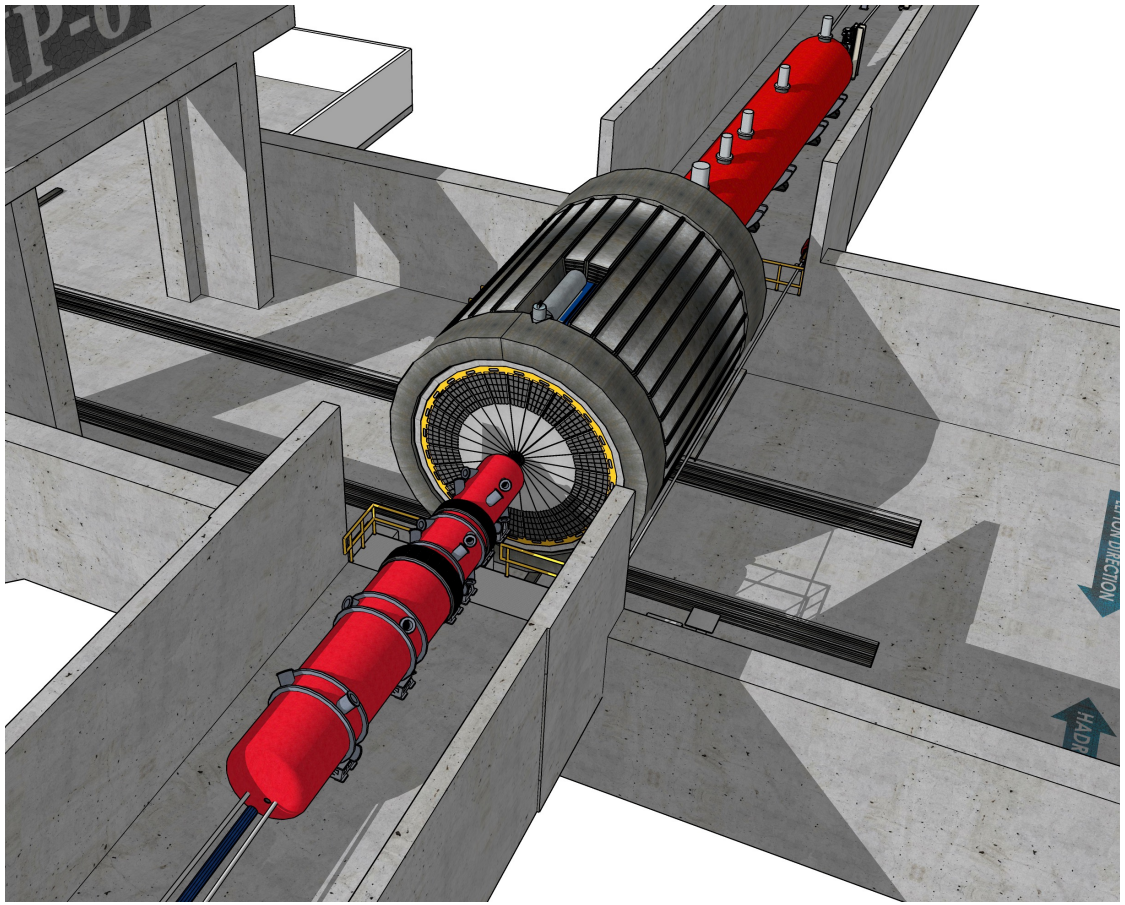
ePIC におけるストリーミングコンピューティング

Markus Diefenthaler
(ePIC Software & Computing Coordinator, Jefferson Lab)

The Highly-Integrated ePIC Experiment

Integrated Interaction and Detector Region (90 m)
Get close to full acceptance for all final state particles, and measure them with good resolution. All particles count!



Compute-Detector Integration
Seamless data processing from detector readout to analysis using streaming readout and streaming computing.

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 data processing is deferred to computing.

Advantages of Streaming Readout

- Simplification of readout (no custom trigger hardware and firmware) and increased flexibility.
- Event building from holistic detector information.
- Continuous data flow provides detailed knowledge of backgrounds and enhances control over systematics.

Compute-Detector Integration to Maximize Science

Broad ePIC Science Program:

- **Plethora of observables**, with less distinct topologies where every event is significant.
- **High-precision measurements**: **Control of systematic uncertainties of paramount importance**.

Streaming Readout Capability Due to Moderate Signal Rate:

- **Capture every collision signal**, including background.
- Event selection 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	$\vec{e}+\vec{p}, \vec{e}+A$	$\vec{p}+\vec{p}/A, A+A$	$p+p/A, A+A$
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

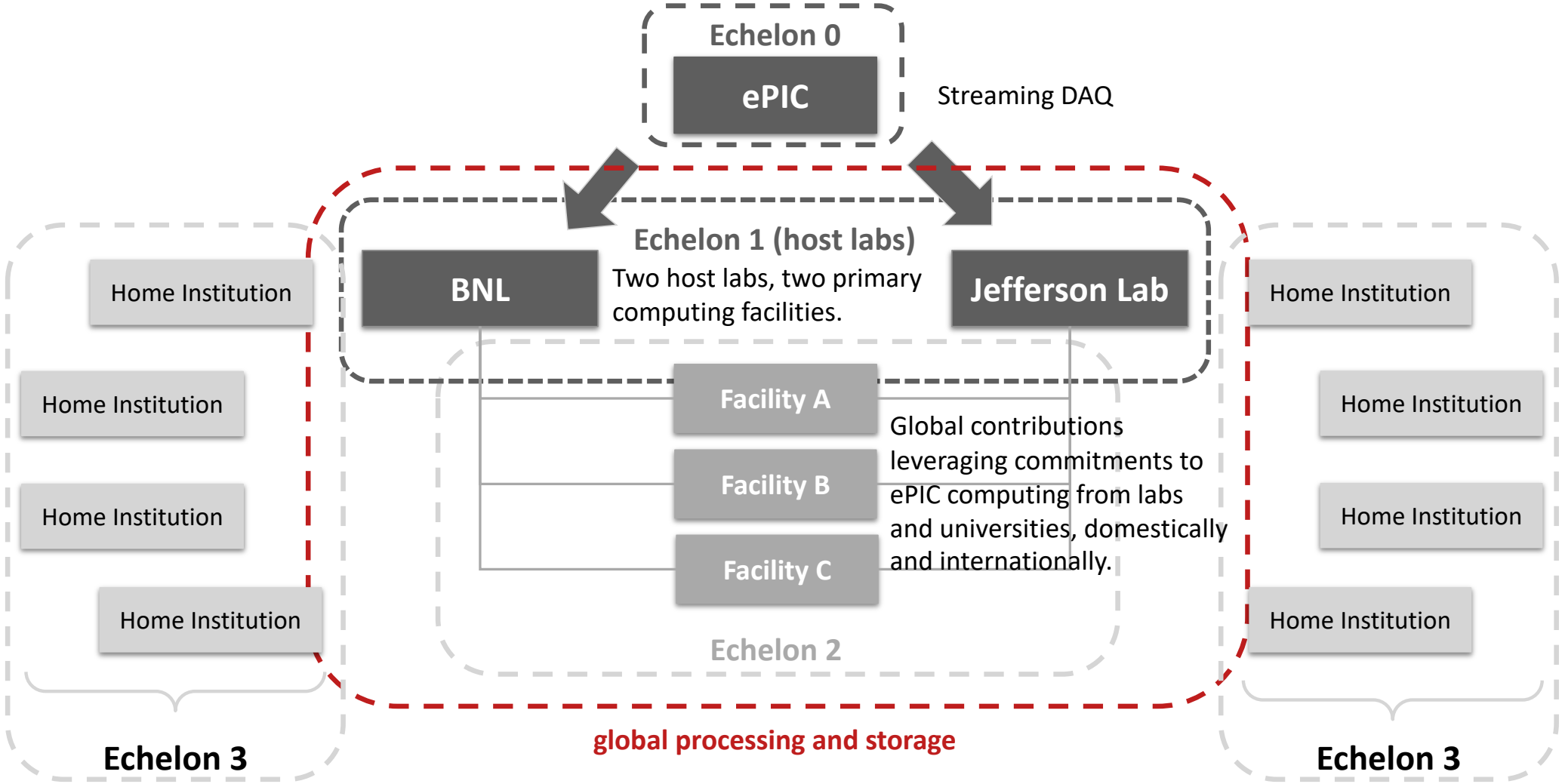
- **Problem** Data for physics analyses and the resulting publications available after O(1year) due to complexity of NP experiments (and their organization).
  - Alignment and calibration of detector as well as reconstruction and validation of events time-consuming.
- **Goal** **Rapid turnaround of 2-3 weeks for data for physics analyses**.
  - Timeline driven by alignment and calibrations.
  - Discussed alignment and calibration procedures and requirements with detector experts. Preliminary information from Detector Subsystem collaborations indicates that 2-3 weeks are realistic.
- **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 autonomous validation for rapid processing.

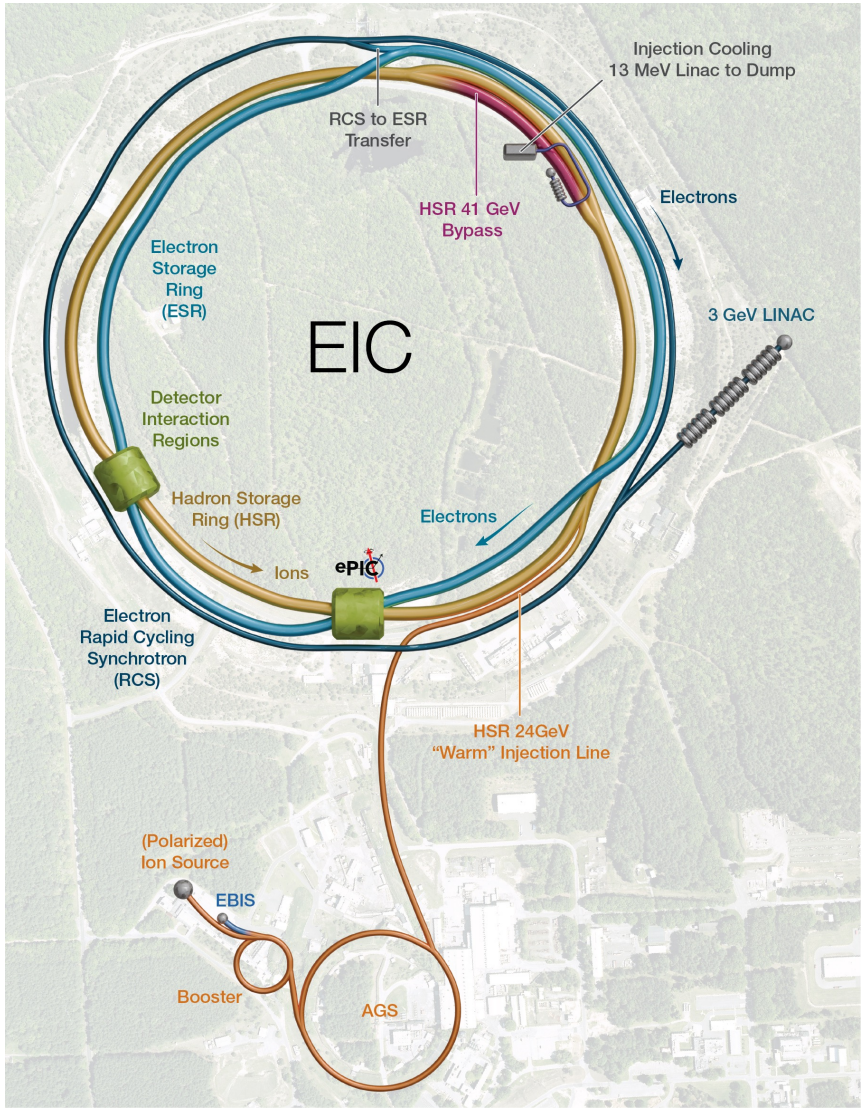
Heterogeneous computing for acceleration (CPU, GPU).

The ePIC Streaming Computing Model



Supporting the analysis community *where they are* at their home institutes, primarily via services hosted at Echelon 1 and 2.

Towards a Quantitative Computing Model: The EIC and Event Rates



- **Versatile machine**: versatile range of beam polarizations, beam species, center of mass energies.

- **High luminosity** up to L = 10^{34} cm^{-2} s^{-1} = 10 kHz/µb.
  - The e-p cross section at peak luminosity is about 50 µb. This corresponds to a signal event rate of about 500 kHZ.

- The **bunch frequency** will be **98.5MHz**, which corresponds to a **bunch spacing** of about **10ns**.
  - For e-p collisions at peak luminosity, there will be in average 200 bunches or about 2µs between collisions (98.5MHz / 500 kHz).

- The EIC Project and ePIC are currently discussing the early science program of the EIC:
  - For the computing resource estimate, we assume a luminosity scenario of L = 10^{33} cm^{-2} s^{-1} = 1 kHz/µb in 2034.

Towards a Quantitative Computing Model: Rate Estimates from Streaming DAQ

- **Event size of in average 400 kbit**,
  - Including signal and background apart from detector noise,
  - Assuming that detector noise can be substantially reduced in early stages of processing.
  - Event sizes will decrease in later stages of data taking as detector thresholds are raised.

- **Data rate of in average 30 Gbit/s**,
  - Estimate of upper limit: 10Gbit/s for detector noise + event rate * event size.
  - Event rate = 50 KHz for EIC Phase 1 luminosity and maximum e-p cross section of *50 μb.*

- **Running 60% up-time for ½ year = 9,460,800 s**:
  - Data rate of 30 Gbit/s results in 710 × 10^9 events per year.
  - The data volume of 35.5 PB per year will be replicated between Echelon 1 facilities (71 PB in total).

Computing Use Cases

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

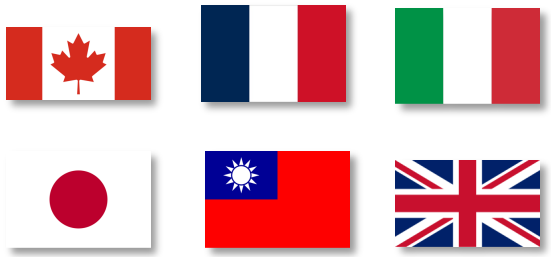
Prompt := rapid low-latency processing.

Prompt processing of newly acquired data typically begins in seconds, not tens of minutes or longer.

Global Computing

- **Echelon 2**: Global contributions leveraging commitments to ePIC computing from labs and universities, domestically and internationally.
- **Echelon 3**: Supporting the analysis community *where they are* at their home institutes, primarily via services hosted at Echelon 1 and 2.

Ongoing Discussions with:



Computing Use Cases and Their Echelon Distribution

Torre Wenaus will cover distributed computing aspects in more detail.

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

Prompt := rapid low-latency processing.

Prompt processing of newly acquired data typically begins in seconds, not tens of minutes or longer.

Assumed Fraction of Use Case Done Outside Echelon 1	
Alignment and Calibration	50%
First Full Reconstruction	40%
Reprocessing	60%
Simulation	75%

- **Echelon 1** sites uniquely perform the **low-latency streaming workflows** consuming the data stream from Echelon 0:
  - Archiving and monitoring of the streaming data, prompt reconstruction and rapid diagnostics.
- Apart from low-latency, **Echelon 2** sites fully participate in use cases and **accelerates** them:
  - Tentative resource requirements model assumes a **substantial role for Echelon 2**.
  - Capabilities and resource requirements for Echelon 2 resources developed jointly with the community.
  - Flexibility aims to leveraging opportunities at facilities for contributing computing resources.
  - The power of distributed computing lies in its flexibility to shift processing between facilities as needed.

International partners already providing resources:

Integration planned:

Computing Resource Needs and Their Implications

Processing by Use Case [cores]	Echelon 1	Echelon 2
Streaming Data Storage and Monitoring	-	-
Alignment and Calibration	6,004	6,004
Prompt Reconstruction	60,037	-
First Full Reconstruction	72,045	48,030
Reprocessing	144,089	216,134
Simulation	123,326	369,979
Total estimate processing	**405,501**	**640,147**

O(1M) core-years to process a year of data:

- Processing scale is substantial.
- Motivates attention to leveraging distributed and opportunistic resources from the beginning.

Storage Estimates by Use Case [PB]	Echelon 1	Echelon 2
Streaming Data Storage and Monitoring	71	35
Alignment and Calibration	1.8	1.8
Prompt Reconstruction	4.4	-
First Full Reconstruction	8.9	3.0
Reprocessing	9	9
Simulation	107	107
Total estimate storage	**201**	**156**

~350 PB to store data of one year.

Computing resource needs comparable to LHC experiments ATLAS and CMS.

ePIC is compute intensive experiment; must ensure ePIC is not compute-limited in its science.

Technology Choices and Evolution

- **Modularity is Key**: 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.

- **Lessons Learned from the NHEP Community** informed the ePIC Streaming Computing Model:
  - Our software is deployed via containers.
  - Our containers are distributed via CernVM-FS.
  - We run large-scale simulation campaigns on the Open Science Grid.
  - Access to our simulations is facilitated through XRootD.
  - We are in the process of deploying Rucio for distributed data management, improving access for collaborators to specific simulation files.

  Recent Review: *"ePIC Software & Computing plans well integrated with standard practices across NHEP."*

- **Software Stewardship by the NHEP Community**:
  - **ECSAC Review**: *"ePIC Software & Computing uses many common tools and are active contributors to several."*
  - ePIC software stack used for other experiment: Detector II and SoLID fixed-target experiment.

The Role of AI and Data

- **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.

- AI will **empower the data processing** at the EIC.
  - Rapid turnaround of data relies on autonomous alignment and calibration as well as autonomous validation of reconstructed event.

- AI will also **empower autonomous experimentation and control** beyond data processing:
  - Vision for a responsive, cognizant detector system, .e.g., adjusting thresholds according to background rates.
  - Enabled by access to full detector information via streaming readout.
  - The full detector data is a vast asset that can be leveraged for further AI applications (the ”retina” of the experiment).

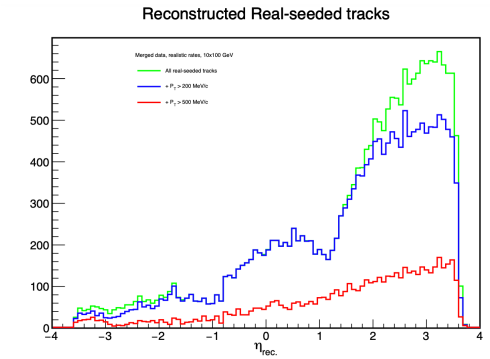
Prototype of Event Reconstruction from Streaming Data

Scope of the first prototype: Track reconstruction only. Demonstrated that we can correlate hits in a realistic time frame to the various events in the time window of the MAPS of 2µs.

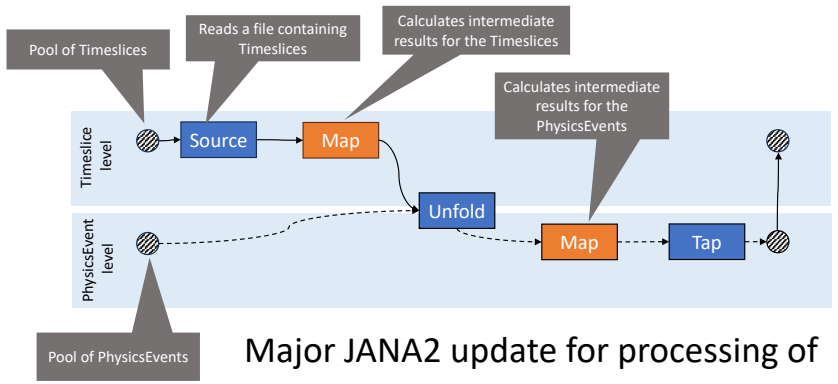
Streaming DAQ → **Timeframes** → **Streaming Computing** → **Reconstructed Events**

Continuous stream of data

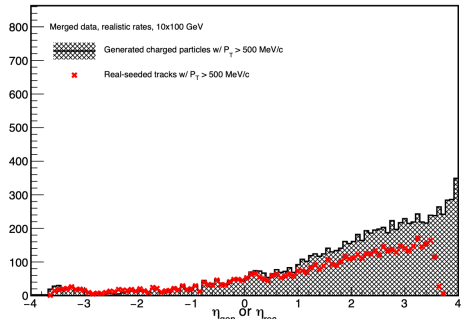
0.6ms timeframes



- Data transferred in collections called *timeframes*.
- Each timeframe includes:
  - Data read from detectors over a time window of 2^{16} cycles of the beam RF, equivalent to 0.6 ms.
  - Channel information and corresponding timing data



Major JANA2 update for processing of timeframes, events, and subevents.



Nathan Brei will cover JANA2 and streaming reconstruction in more detail.

Streaming DAQ and Computing Milestones

Japan and SPADI Alliance involvement highly welcome. Milestones guide priorities for collaboration.

Streaming DAQ Release Schedule:

PicoDAQ **FY26Q1**
- Readout test setups

MicroDAQ: **FY26Q4**
- Readout detector data in test stand using engineering articles

MiniDAQ: **FY28Q1**
- Readout detector data using full hardware and timing chain

Full DAQ-v1: **FY29Q2**
- Full functionality DAQ ready for full system integration & testing

Production DAQ: **FY31Q3**
- Ready for cosmics

Streaming Computing Milestones:

Start development of streaming orchestration, including workflow and workload management system tool.
Start streaming and processing streamed data between BNL, Jefferson, DRAC Canada, and other sites.

Support of test-beam measurements, using variety of electronics and DAQ setups:
- Digitization developments will allow detailed comparisons between simulations and test-beam data.
- Track progress of the alignment and calibration software developed for detector prototypes.
- Various JANA2 plugins for reading test-beam data required. Work started on an example.

Establish autonomous alignment and calibration workflows that allows for validation by experts.

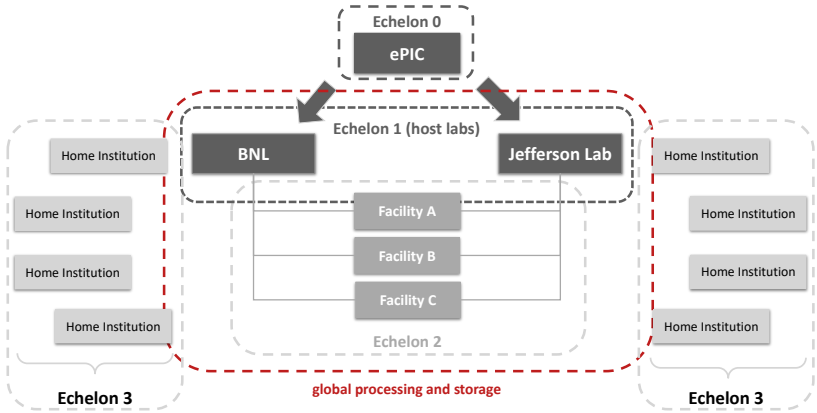
Analysis challenges exercising end-to-end workflows from (simulated) raw data.

Streaming challenges exercising the streaming workflows from DAQ through offline reconstruction, and the Echelon 0 and Echelon 1 computing and connectivity.
Analysis challenges exercising autonomous alignment and calibrations.

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 computing requirements of ePIC.

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 alignment and calibration.
- **Four tiers of the ePIC Streaming Computing Model computing fabric:**
  - **Echelon 0**: ePIC experiment and its streaming readout.
  - **Echelon 1**: Two host labs, two computing facilities.
  - **Echelon 2**: Global contributions for flexibility and to accelerate science.
  - **Echelon 3**: Full support of the analysis community.



- **ePIC** will be a compute **intensive experiment**; must ensure it is not compute-limited in its science.
- **Next steps**:
  - **Ongoing**: "**Streaming Computing XII**" workshop, December 2–4, 2024, University of Tokyo, Tokyo, Japan.
  - "**ePIC Data: From Detector Readout to Analysis**", parallel session at the ePIC Collaboration meeting, January 20–24, 2025, Frascati, Rome.
  - Continue good collaboration with **EIC Computing and Software Joint Institute**:
    - Establish **EIC International Computing Organization** and distributed computing fabric for ePIC with international partners. More details in "International Computing Model and Governance" session.
  - "**ePIC Streaming Computing Model**" publication in preparation.

Backup

Towards a Quantitative Computing Model: Reconstruction and Simulation

Reconstruction and Simulation Times	Times based on current software on modern cores
Reconstruction event processing time with background **[s]**	2
Reconstruction algorithmic speedup factor 10yrs out	1.5
Simulation event processing time with background **[s]**	15
Full simu speedup factor 10yrs out	1.5
Combined time with background, with speedup **[s]**	11

Simulation Use Cases		
Number of simulated events per event of interest	10	The canonical 10x more.
Optimized simu events per physics event	4	~40% of measured events will be signal.
Fast simulation speedup relative to full simulation	4	
Proportion of simulation events using fast simulation	70%	

Computing Resource Estimates

Actual needs in 2034.

Processing by Use Case [cores]	Echelon 1	Echelon 2
Streaming Data Storage and Monitoring	-	-
Alignment and Calibration	6,004	6,004
Prompt Reconstruction	60,037	-
First Full Reconstruction	72,045	48,030
Reprocessing	144,089	216,134
Simulation	123,326	369,979
Total estimate processing	**405,501**	**640,147**

See prompt reconstruction.

Roughly 10% of data stream.

Must keep up with data taking; assume 2x headroom.

Reprocessing includes simulation as well as data.

Simply adding together the core counts is an overestimate. Reconstruction core hours used only part time.

Storage Resource Estimates

Actual needs in 2034.

Storage Estimates by Use Case [PB]	Echelon 1	Echelon 2
Streaming Data Storage and Monitoring	71	35
Alignment and Calibration	1.8	1.8
Prompt Reconstruction	4.4	-
First Full Reconstruction	8.9	3.0
Reprocessing	9	9
Simulation	107	107
Total estimate storage	**201**	**156**

Echelon 1 sites arrive data, two copies
One copy (can and may be more) across Echelon 2 sites for alignment, calibration, and reconstruction use cases.

Networking Estimates

Echelon 0: The raw data from the ePIC Streaming DAQ (Echelon 0) will be replicated across the host labs (Echelon 1). At the highest luminosity of 1e34, the data stream from the ePIC Streaming DAQ is estimated at 100 Gbit/s. Consequently, Echelon 0 requires an outgoing network connection of at least 200 Gbit/s.

Echelon 1: Each Echelon 1 facility has similar requirements, as it will receive up to 100 Gbit/s of raw data and will share this data with Echelon 2. In addition, Echelon 1 will send a small amount of monitoring data, approximately 1 Gbit/s, back to Echelon 0. Echelon 1 will also receive calibration and analysis data from various Echelon 2 nodes at a comparable rate of about 1 Gbit/s.

Echelon 2: The network connection requirements for Echelon 2 facilities will depend on the proportion of raw data they intend to process. For the 10% of Echelon 1 scenario, a network connection of 20 Gbit/s would be required.

Streaming Data Processing

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 data acquisition.

ePIC Streaming Data Processing Characteristics

- Quasi-continuous flow of fine-grained data.
- Dynamic flexibility to match real-time data inflow.
- Prompt processing is crucial for data quality and detector integrity.
- Processing full data set quickly to minimize time for detector calibration and deliver analysis-ready data.

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 data-taking 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.