ePIC Mini Workshop on Heterogeneous Computing: Introduction and Goals

M. Diefenthaler (Jefferson Lab)

What is Heterogeneous Computing?

- **Heterogeneous computing** combines multiple types of processors:

Processor Type	Strength
CPU	Flexible, general-purpose processing
GPU	Massive parallel throughput
FPGA	Deterministic low-latency processing
AI Accelerators	Optimized ML inference and training
ASICs	Maximum efficiency for dedicated tasks
DPU	Data movement and networking acceleration
DSP	Fast signal and waveform processing

CPU Heterogeneity: x86 and ARM.

- **Goal**: Match tasks to the hardware best suited for them.
- **Today**, this approach is widely used in devices ranging from smartphones to supercomputers.
- **Future** computing platforms will become more specialized, intelligent, and tightly integrated.

ePIC in the Era of Heterogeneous Computing

- ePIC will operate in this emerging era of heterogeneous computing.
- This workshop series aims to define a coherent approach within ePIC.
- The goal is to address key questions connecting software, workflows, and infrastructure.

Compute-Detector Integration to Accelerate Science

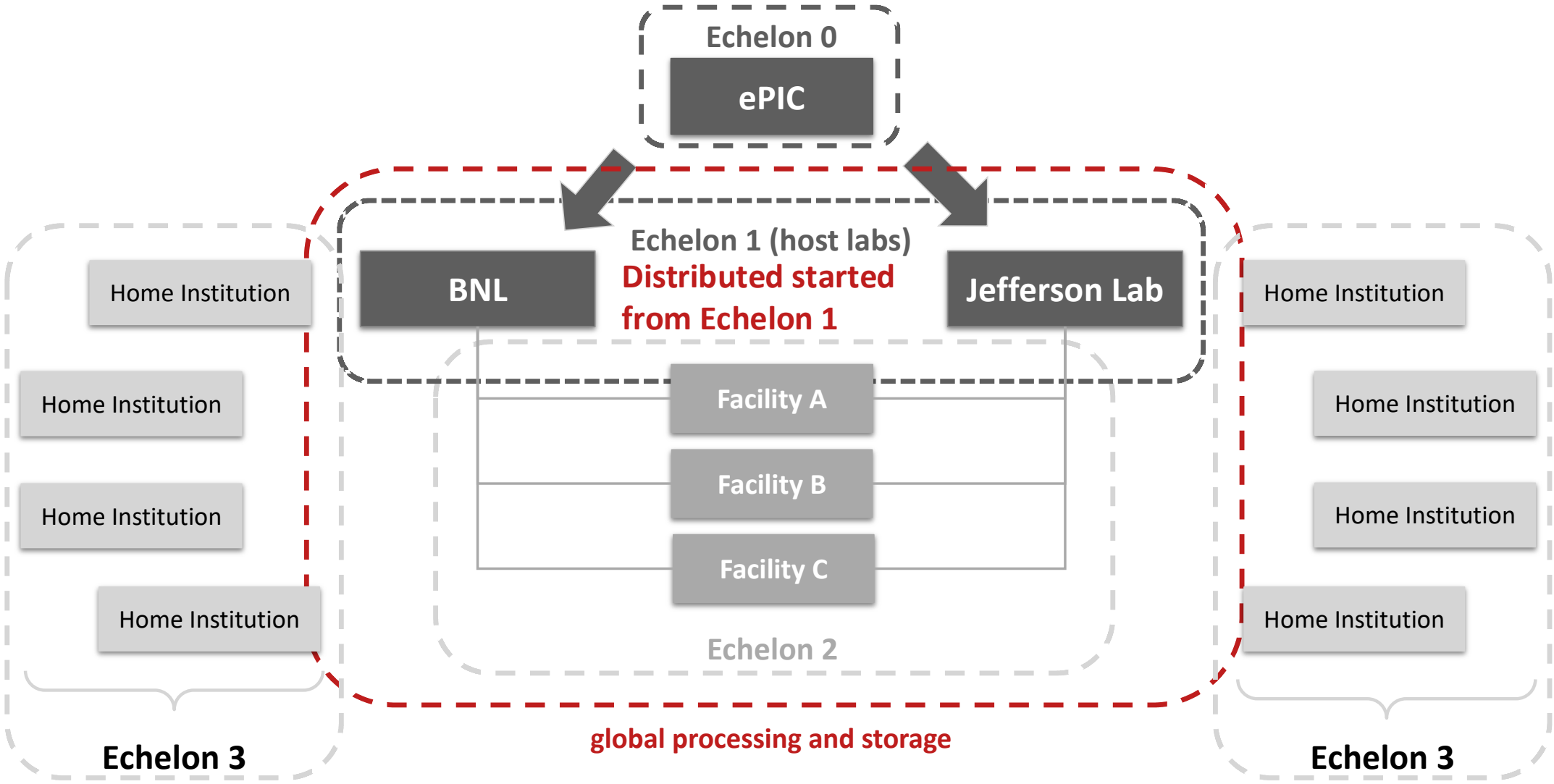
Streaming readout for continuous data flow of the full detector information.

AI for autonomous alignment and calibration, autonomous validation for rapid processing, expanded scientific insight.

Heterogeneous computing for acceleration.

- **FPGA**-based aggregation and timing systems.
- **CPUs** and **GPUs** at Echelon 1 and Echelon 2 for data processing and analysis.

ePIC Distributed Computing Fabric



Design Principles: Compute-Detector Integration

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.

Design Principles: Heterogeneous Computing

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.

Heterogeneity in ePIC

- **Streaming DAQ**:
  - Timing distribution and aggregation rely on FPGA technologies.
- **Streaming Computing**:
  - Echelon 1 and Echelon 2 sites may include GPUs and accelerator hardware.
  - AI techniques are already being explored for event identification and filtering.
  - Simulation and reconstruction workflows will span multiple architectures.
  - Digital twins and real-time detector modeling are opportunities.
  - Beyond Echelon 1 and 2: Possibly cloud/opportunistic based resources tailored to the problem (specialized hardware, e.g., AI/ML processors).

Where Is Heterogeneous Computing Beneficial?

- Identify workflows where heterogeneous computing provides clear benefit.
- Benefit means **measurable gains** that **justify the investment** in software development or hardware.
- **Gains**: improved throughput, resource utilization, or physics performance, as well as reduced latency.
- **Prototype projects** are essential for evaluation.

Example: Simulations

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**

Resource estimates assume both full and accelerated simulations.

11:05 AM **Accelerating Geant4 detector simulations on GPUs with AdePT**
Speaker: Severin Diederichs (CERN)

11:25 AM **Discussion**

11:35 AM **Geant4 Sub-Event Parallelism**
Speaker: Makoto Asai (JLab)

11:50 AM **Discussion**

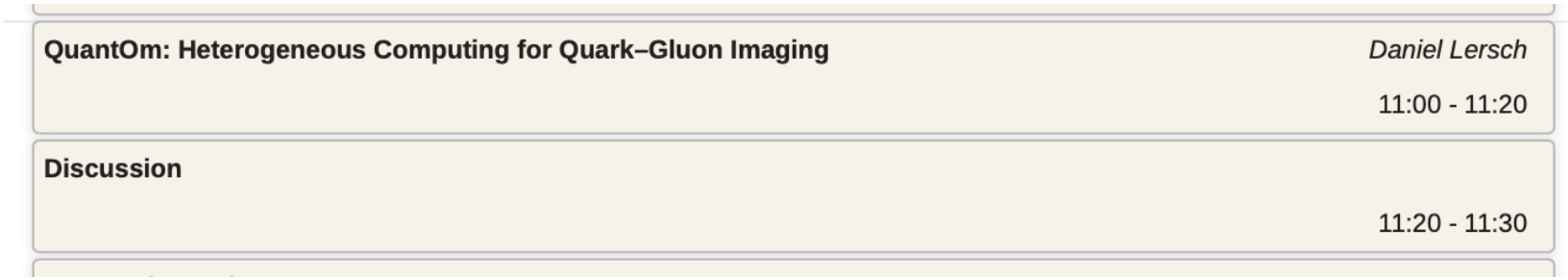
12:00 PM **EIC-Opticks**
Speaker: Gabor Galgoczi (BNL)

Accelerated Simulations:

- GPU-based detector simulation.
- AI-based approaches for fast simulations with high fidelity.
- Heterogenous computing for simulations will be discussed further on June 3 (see agenda on the left).

Example: HPC Systems

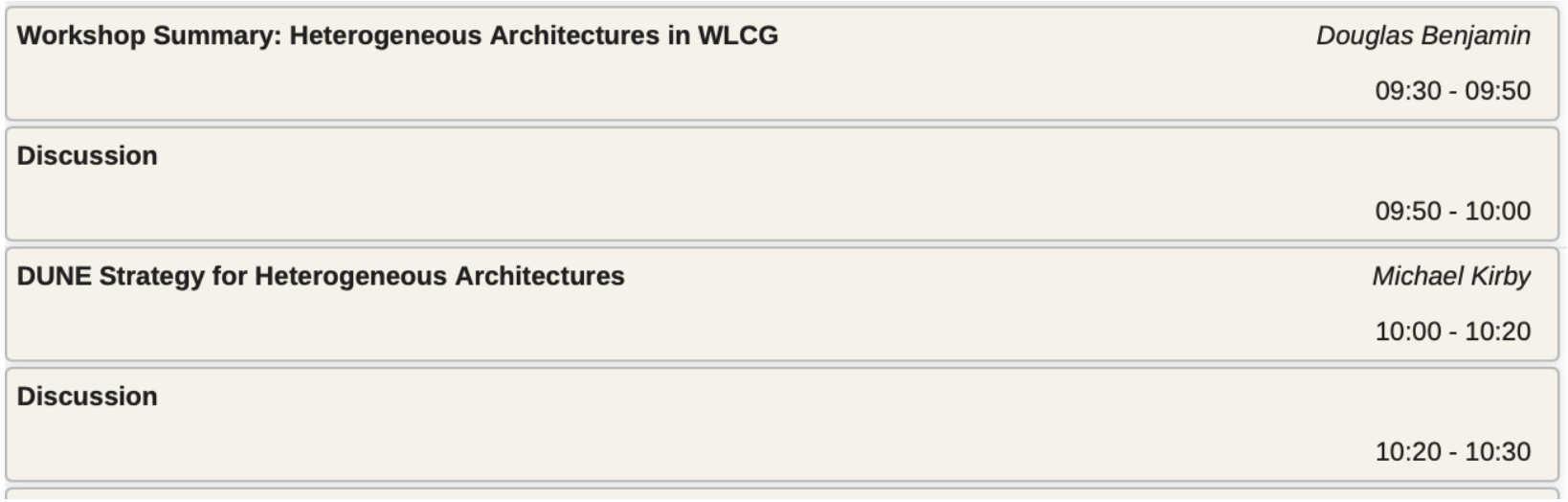
- HPC systems offer unprecedented parallelism, but efficient execution remains challenging, particularly for HENP communities with a long tradition of HTC workflows.
- Taking advantage of HPC often requires substantial investments. At the same time, these systems create opportunities to **rethink** and **redesign research workflows**.
- An example is **QuantOm**, which brings experiment and theory together in a joint workflow for imaging quarks and gluons and their interactions.



- Another example is Perlmutter at NERSC, which is already used for our monthly simulation campaigns.
- AI/ML growth will help with convergence of HPC capability and HENP applicability.

Working with the Community

- ePIC will engage with the broader community on heterogeneous computing efforts.
- The workshop includes **perspectives from WLCG and DUNE**:



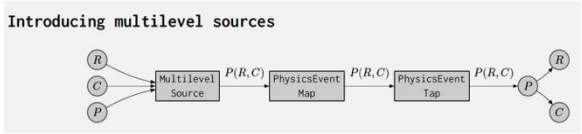
Workshop Summary: Heterogeneous Architectures in WLCG	*Douglas Benjamin* 09:30 - 09:50
Discussion	09:50 - 10:00
DUNE Strategy for Heterogeneous Architectures	*Michael Kirby* 10:00 - 10:20
Discussion	10:20 - 10:30

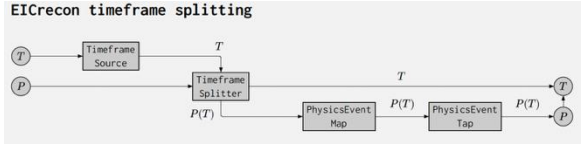
- **Collaboration with the broader community** is essential for:
  - Lessons learned and
  - Sustainable strategies.

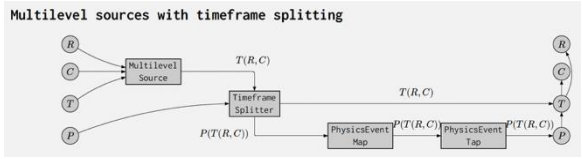
Framework and Workflow Integration

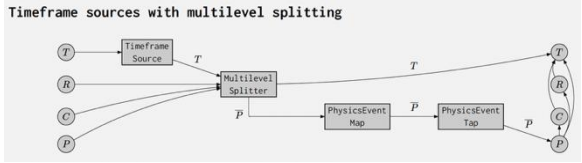
- **Integration with JANA2 and distributed workflows**:
  - Task-based and asynchronous execution models,
  - Efficient data movement between heterogeneous devices.
- **Hierarchical execution models in JANA2**:
  - Multithreaded JANA2 framework provides a component-level hierarchical decomposition of data boundaries into **Run**, **Timeframe**, **PhysicsEvent**, and **Subevent** levels.
  - The **Folder** and **Unfolder** component interfaces enable traversal of this hierarchy by supporting operations such as splitting and merging data streams.
  - What else is required for heterogeneous computing?

Enabling Heterogeneous Architectures in JANA2	*Nathan Brei* 10:30 - 10:50
Discussion	10:50 - 11:00









Discussion Points and Summary

ePIC is being designed as a streaming, distributed, heterogeneous computing system from the beginning.

- Which workflows benefit most from heterogenous architectures?
- Which examples are worth prototyping now? What is the timeline?
- How should ePIC balance performance, portability, and infrastructure costs?

Timeline