

Software Infrastructure Plans for End-to-End and Bottom-Up AI/ML Capabilities at the Electron-Ion Collider (EIC)

Linh Nguyen

Co-authors: Elke Aschenauer, Paul Bachek, Jim Jamilkowski, Kyle Kulmatycki, Jeff Landgraf, Sergei Nagaitsev, Kevin Smith

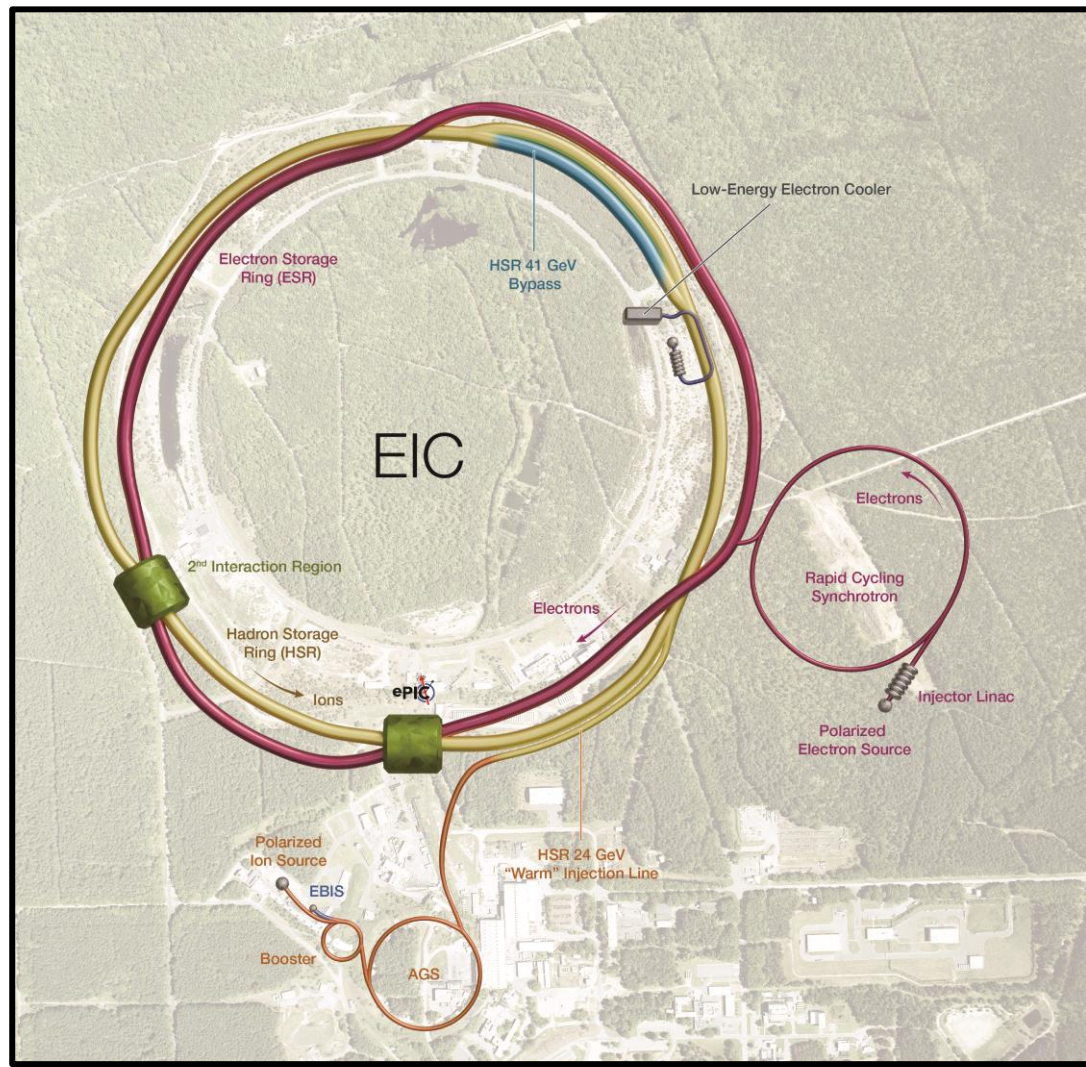
5th ICFA Beam Dynamics Mini-Workshop on Machine Learning for Particle Accelerators

April 9th, 2025

Electron-Ion Collider



The Electron-Ion Collider (EIC)



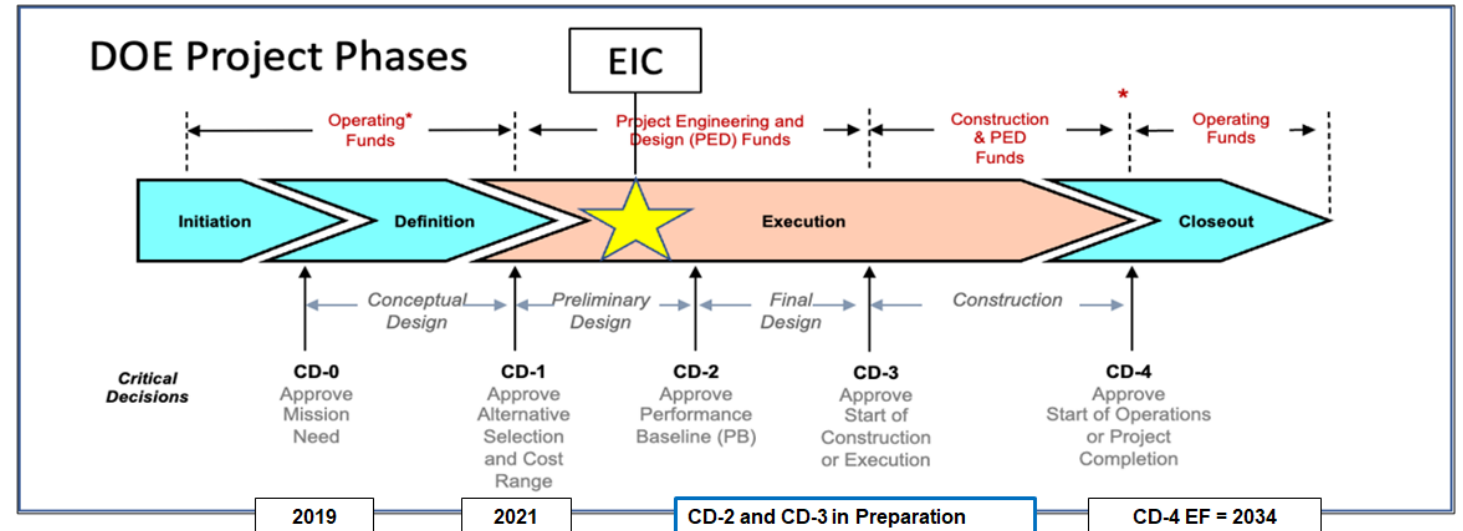
- The EIC is being built at Brookhaven National Laboratory (BNL), in partnership with Jefferson Lab (and other partners, both domestic and international)
- The EIC's physics mission is to unlock the secrets of the strong nuclear force and revolutionize our understanding of the fundamental structure of visible matter

The EIC will be the only operating particle collider in the United States and possibly the only large collider to be built in the world in the next 20-30 years, during the "Age of AI".

We recognize the unique opportunity with regards to AI/ML here. However, the 20 years leading up to EIC Mission Approval didn't occur in the Age of AI. Consequently...

Status of AI/ML at the EIC

- There is currently no formal AI/ML scope in the EIC Construction Project, but **this effort has widespread EIC leadership support.**
- We are now in the mature stages of finalizing the unified AI/ML vision for project approval, and this is the effort that is being presented.
- Having project approval and being on-project means we will be able to affect hardware and infrastructure decisions *and hence costs.* This is crucial to realizing our plans for high-performance, end-to-end, and bottom-up AI/ML capabilities.



Creating a Unified Vision of AI/ML at the EIC

The EIC is being envisioned as a **large-scale AI-ready state-of-the-art facility**. This means plans to support:

Edge AI/ML Capabilities

The EIC will have some of the most demanding AI/ML applications in the world in terms of latencies, data rates, and throughput. This requires specialized local compute resources.

End-to-End AI/ML Capabilities

Unlike in the past, the Detector and Accelerator will form a single AI/ML ecosystem to maximize possibilities and prepare for a human-in-the-loop operations environment. High-quality data will be available across the EIC.

Bottom-Up AI/ML Capabilities

System experts will be enabled to leverage AI/ML for maintaining and optimizing their local systems, for AI/ML deployment at a diversity of scales. This also reflects broad alignment with a modern engineering paradigm.

Creating a Unified Vision of AI/ML at the EIC

In addition, our AI/ML infrastructure must address the following operational needs specific to the EIC:

Long-Term Flexibility

The EIC is scheduled to begin operations around 2035, with a program lasting at least 15 years. To hedge against obsolescence, it must be able to accommodate new AI/ML techniques and models across its lifetime.

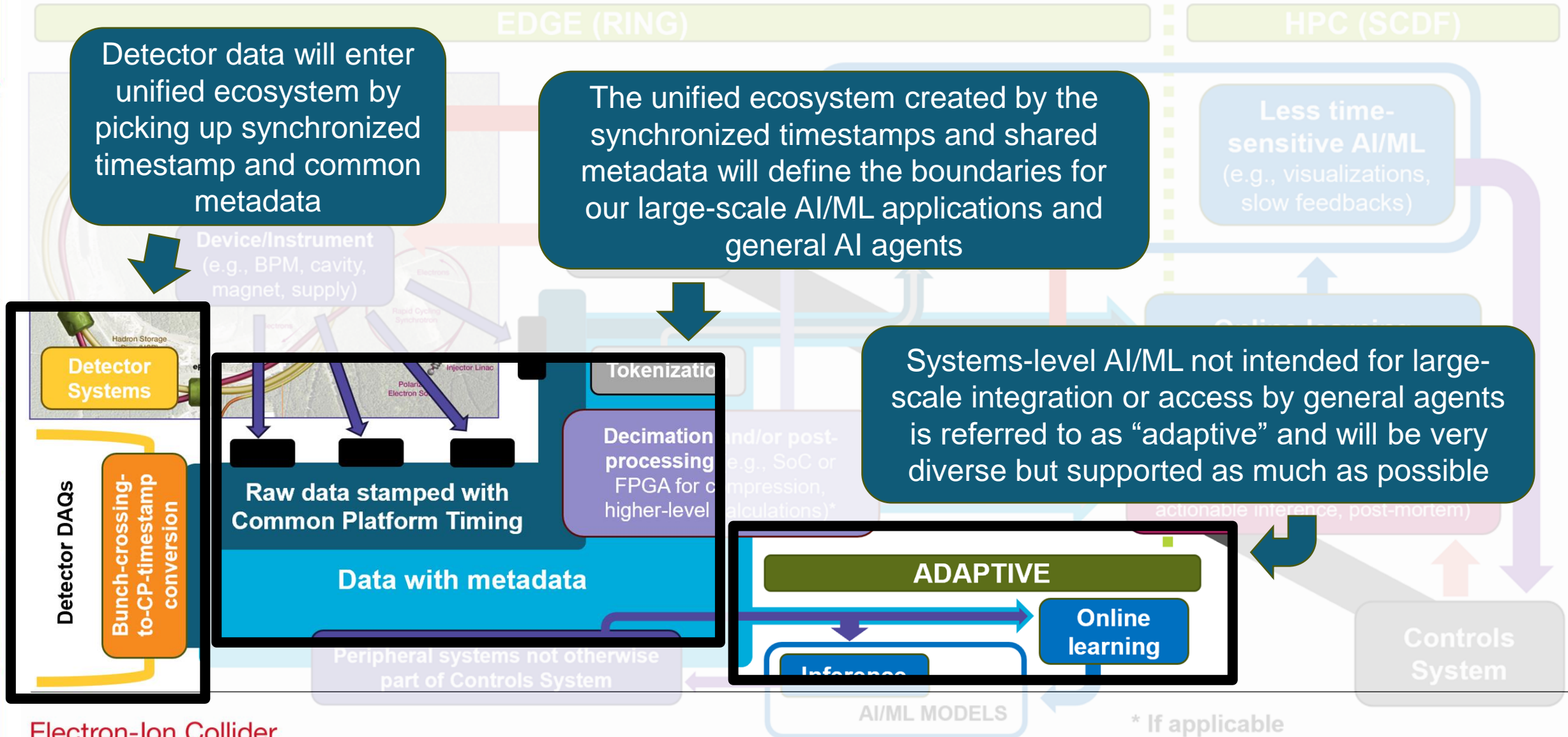
Integration & Deployment

The EIC will be an extremely complex machine. The unified AI/ML ecosystem must be modular in nature to facilitate ease of integration and deployment. Increased automation will help offset inefficiencies.

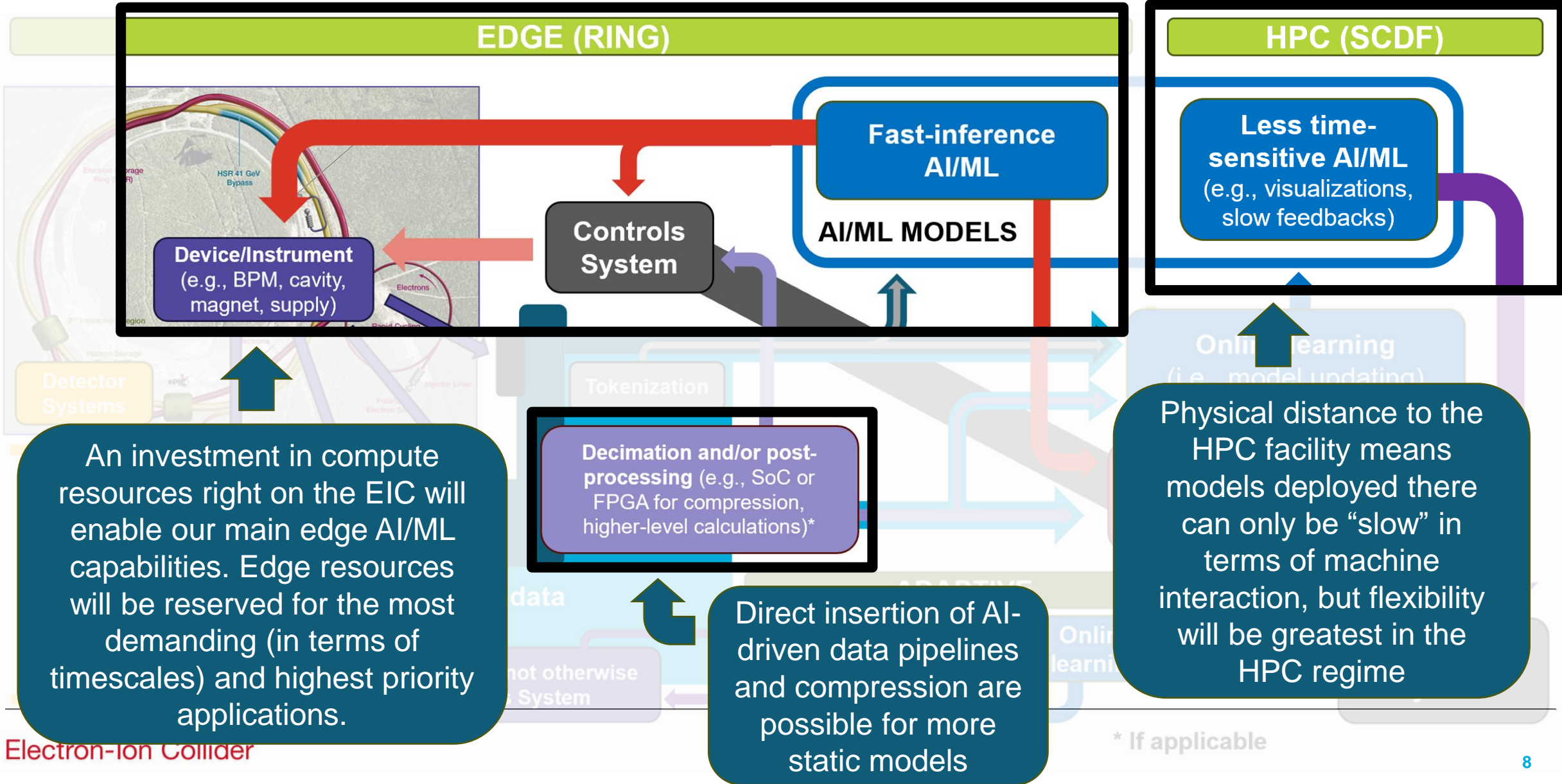
24/7 Operations & Call-In Support

System experts must be able to retain system responsibilities and exercise system expertise independent of AI/ML ecosystems. Nondisruptive cut points must be identified and integrated into the ecosystem for troubleshooting purposes.

Overview of EIC AI/ML Infrastructure Plans



Overview of EIC AI/ML Infrastructure Plans

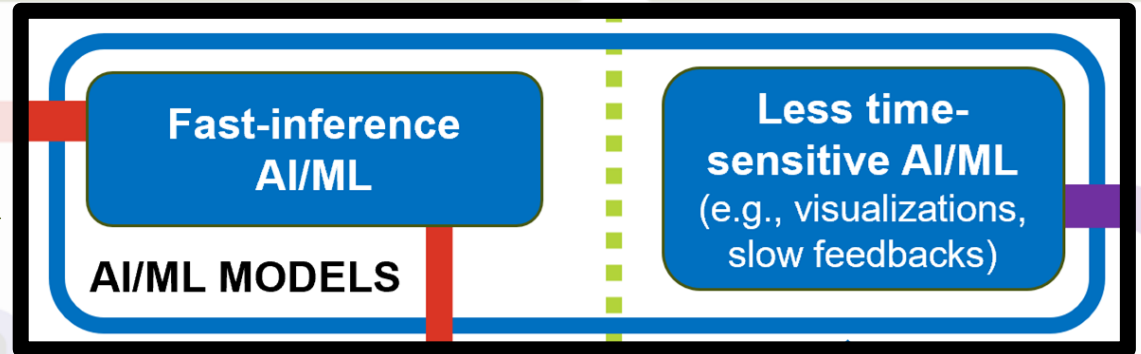


Overview of EIC AI/ML Infrastructure Plans

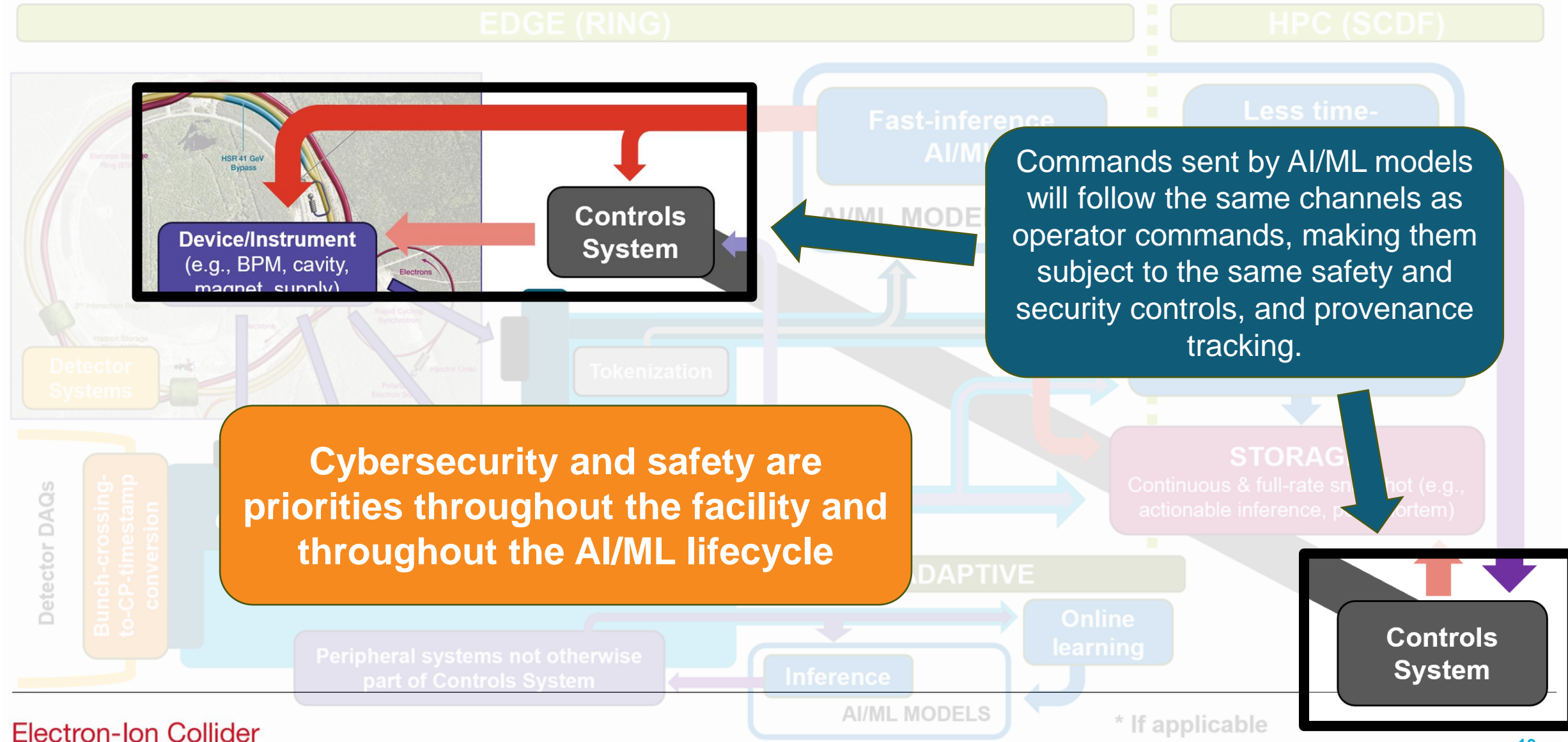
Maximum flexibility and richness of future AI/ML applications is achieved by layering models atop the core infrastructure, whether Edge or HPC. Modularity and ownership are also maintained here. Technique-agnostic automation will handle the zoo of mode-specific models, while foundation models and other large-scale generative AI models can form a higher layer still, above the more specialized models.

Real-time and/or autonomous applications will improve with time via online learning, supporting true digital twin ecosystems and more responsive anomaly detection, for example.

Augmenting standard archiving, AI-driven capture will preserve data showing anomalous machine behavior at full or faster-than-normal rates for review by experts



Overview of EIC AI/ML Infrastructure Plans



AI/ML Applications For Guiding Infrastructure

Although details about AI/ML techniques are off-project, the following applications/ use cases, generated collectively by EIC physicists and engineers and curated by leadership, are being used to help guide decisions about infrastructure:

- Increased and stabilized proton polarization
- Increased and stabilized hadron brightness
- Modeling of high intensity beam dynamics
- ML-based model predictive controls
- Electron Beam Injection Optimization and Beam Matching
- AI/ML enhanced diagnostics for EIC
- Synchrotron Radiation (SR) Shielding Optimization
- Dynamic Collimation System Optimization
- Anomaly detection & predictive maintenance
- Physics-based fast-inference applications

Current work includes establishing the associated performance requirements, with input from the cognizant EIC physicists and engineers, to determine the distribution between Edge and HPC and to ensure adequate data for the phenomena of interest.

Note that none of these applications are needed for successful Project delivery, but they will enhance EIC Operations and research capabilities.

A Vision of a Modern Controls System

On the software side, as part of a modern controls system, we are working on scoping in particular:

- Virtual diagnostics
- High-fidelity digital twins
- AI-assisted autonomous real-time monitoring

The high-fidelity digital twin environment will also serve as a platform for software development and debugging by faithfully replicating the Controls System as a whole. This will minimize time needed on the actual machine and forms a core part of our AI/ML integration and deployment strategy.

The EIC Controls System will be EPICS based, with support for C++, Python, and likely Julia (distribution to be determined). For HPC AI/ML, flexibility is the goal. To meet the performance demands of Edge AI/ML, however, a fully optimized workflow will need to be prescribed, and this investigation is ongoing.

Thank you for your attention

**If you are interested in collaborating,
please reach out to me at lnghuyen@bnl.gov**

Many thanks to my very helpful co-authors:

Elke Aschenauer, Paul Bachek, Jim Jamilkowski, Kyle Kulmatycki,
Jeff Landgraf, Sergei Nagaitsev, Kevin Smith