

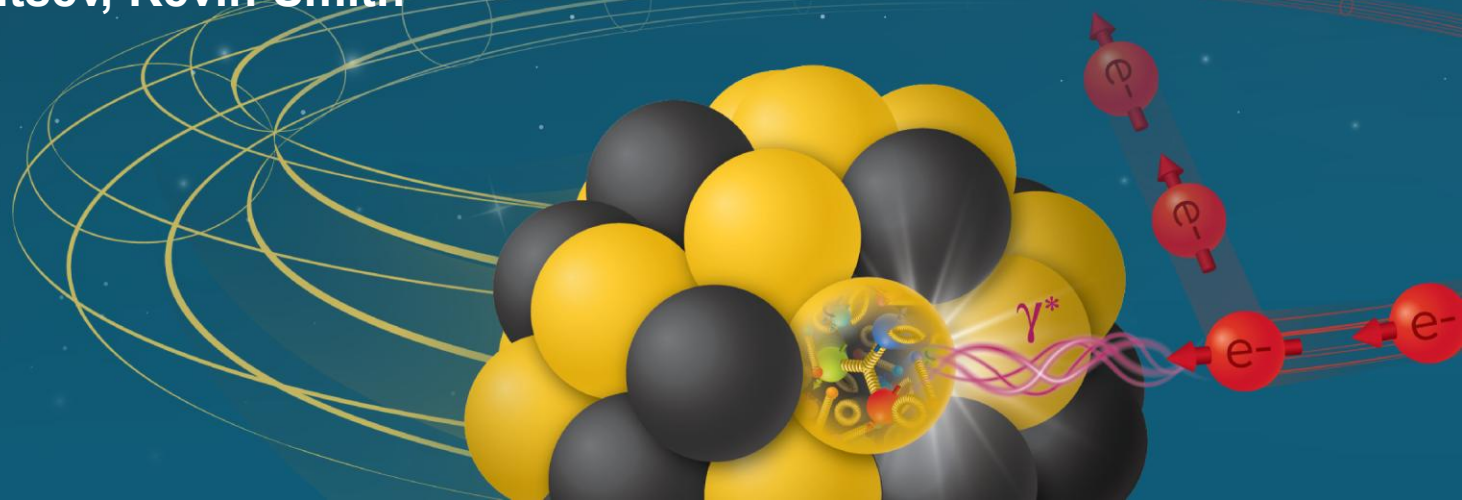
A Unified Vision of AI/ML at the Electron-Ion Collider: Infrastructure and Capabilities

Linh Nguyen

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

4th AI4EIC Workshop, 2025
October 29, 2025

Electron-Ion Collider



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 and optimized infrastructure.

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. AI-driven tools will enhance the productivity of workflows.

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.

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.

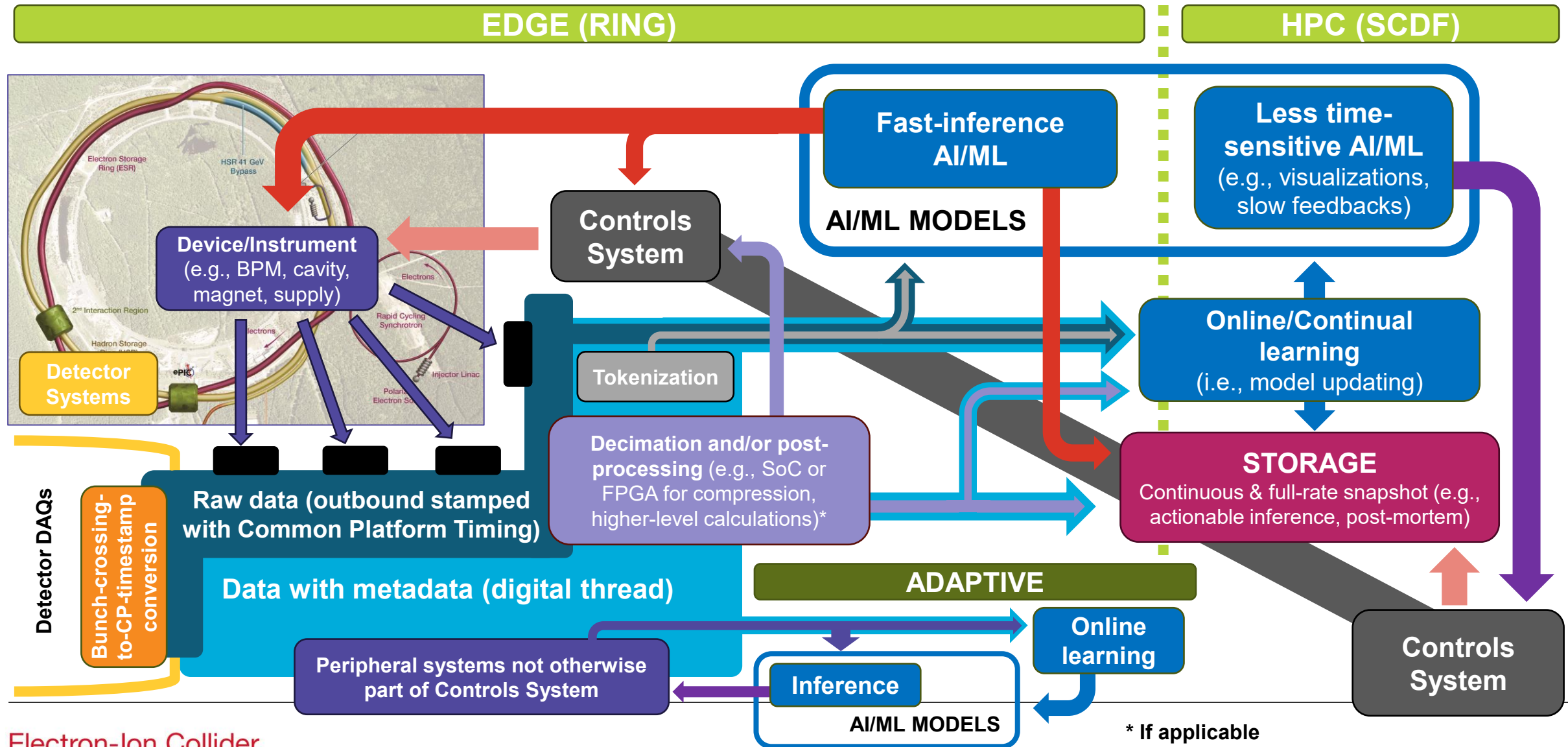
Integration & Deployment

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.

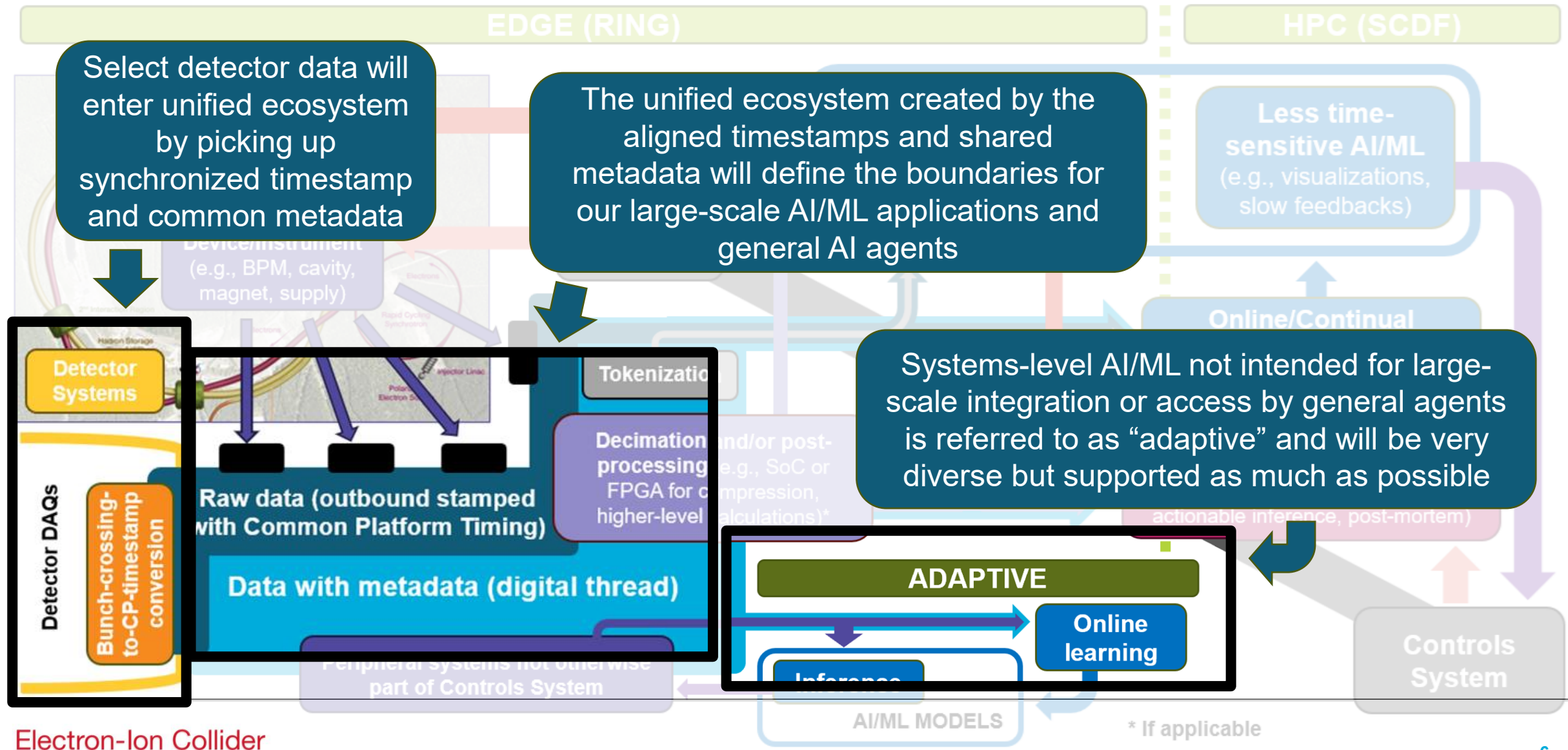
24/7 Operations & Call-In Support

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 and other operational purposes.

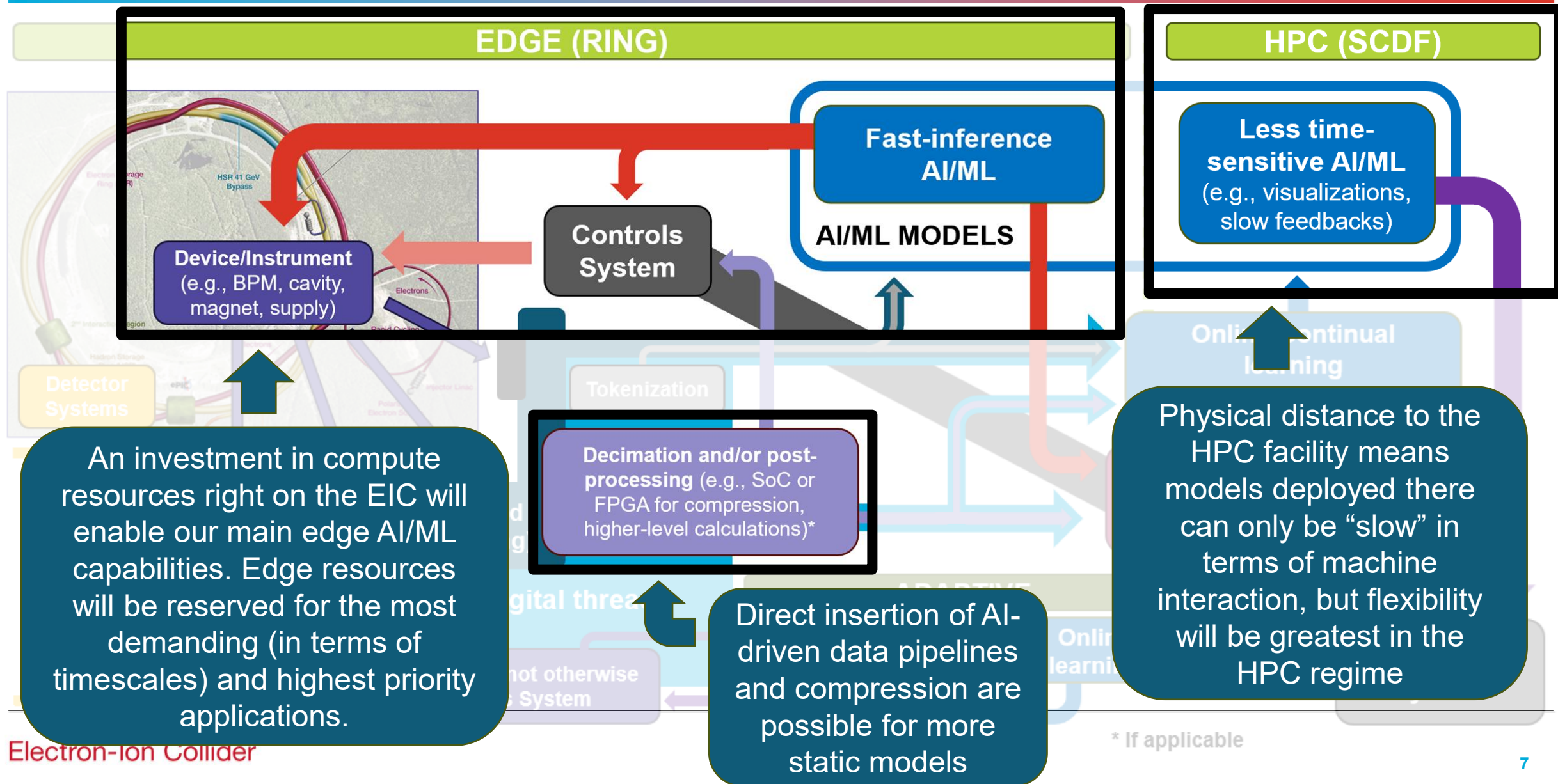
Overview of EIC AI/ML infrastructure plans



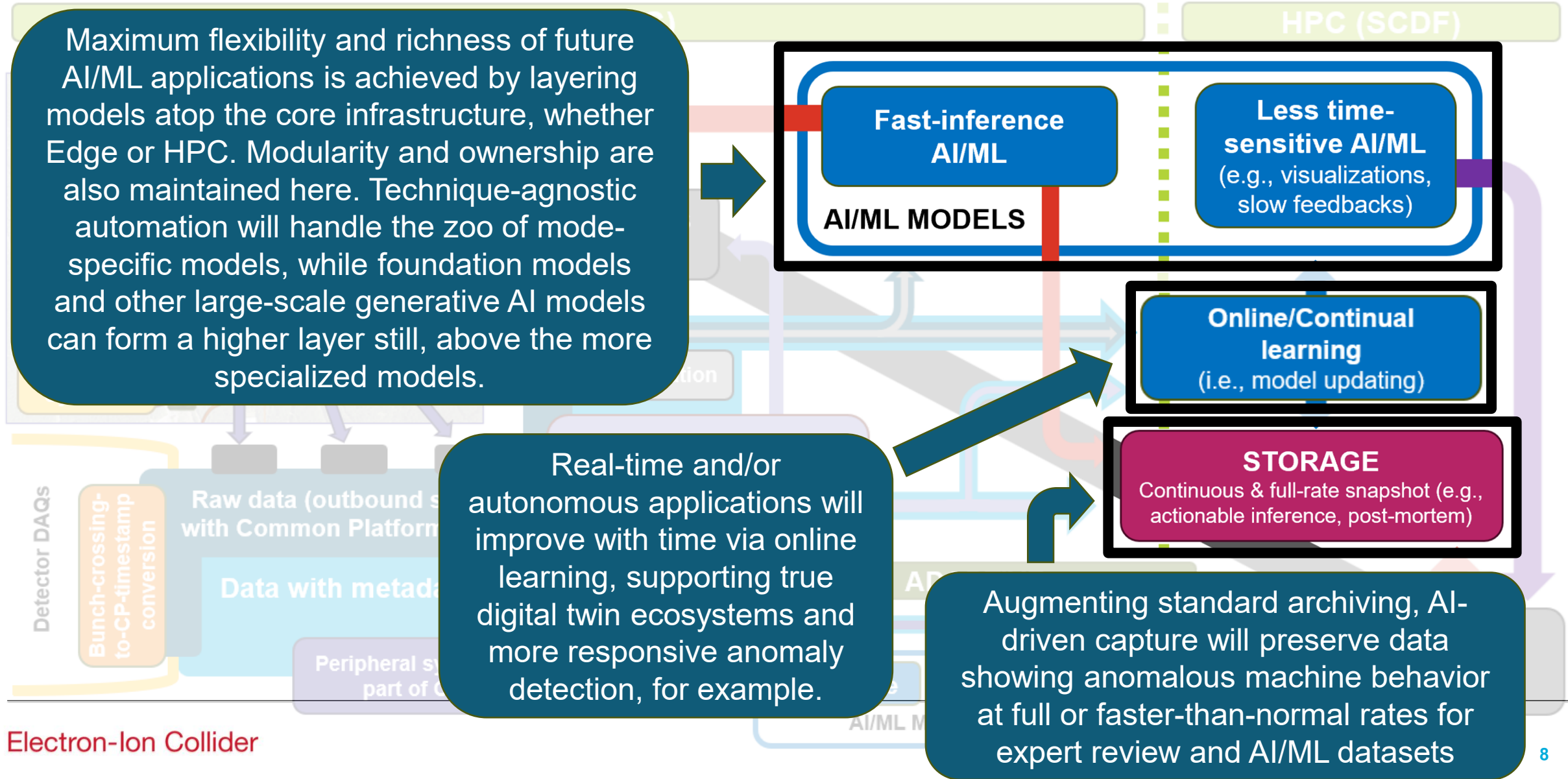
Overview of EIC AI/ML infrastructure plans



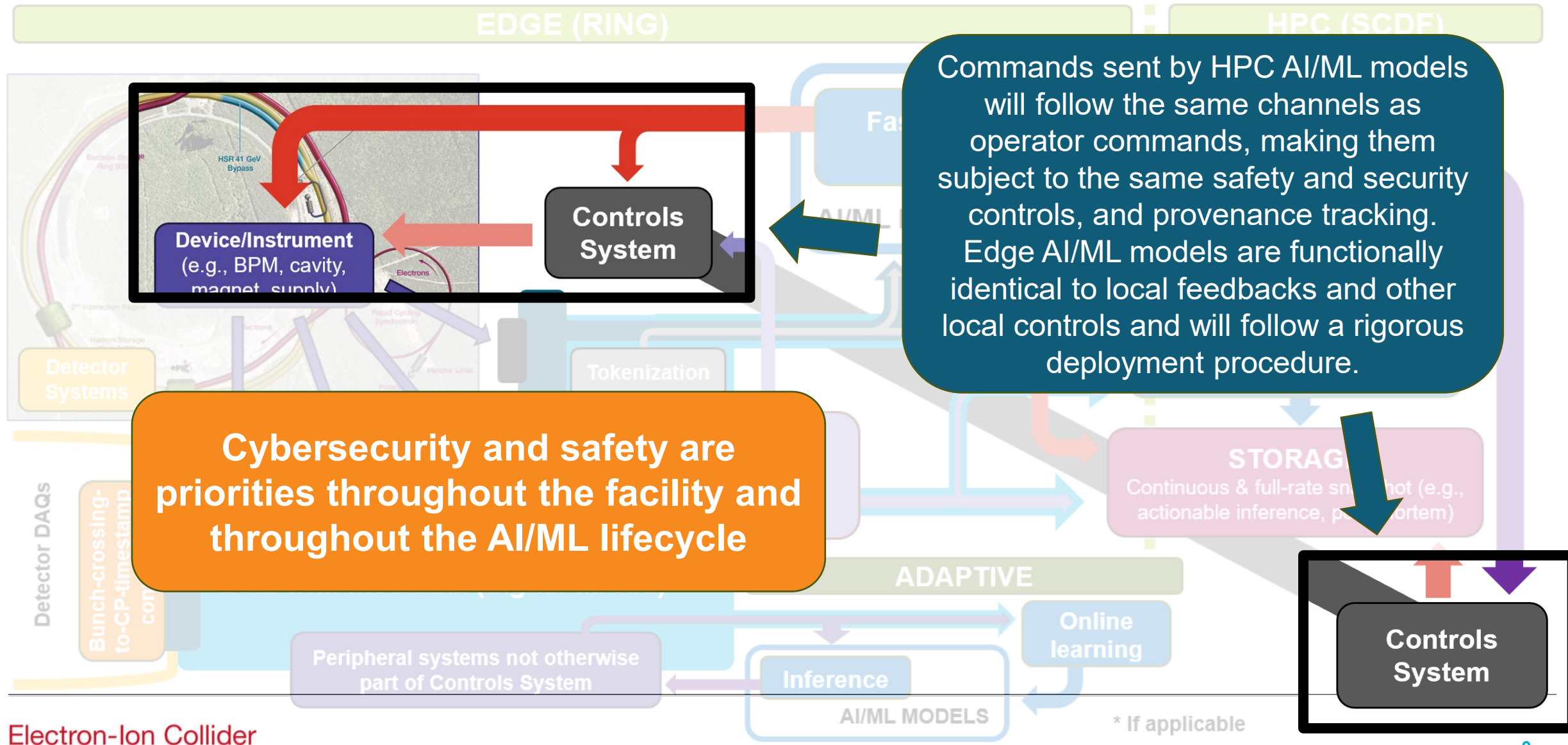
Overview of EIC AI/ML infrastructure plans



Overview of EIC AI/ML infrastructure plans



Overview of EIC AI/ML infrastructure plans

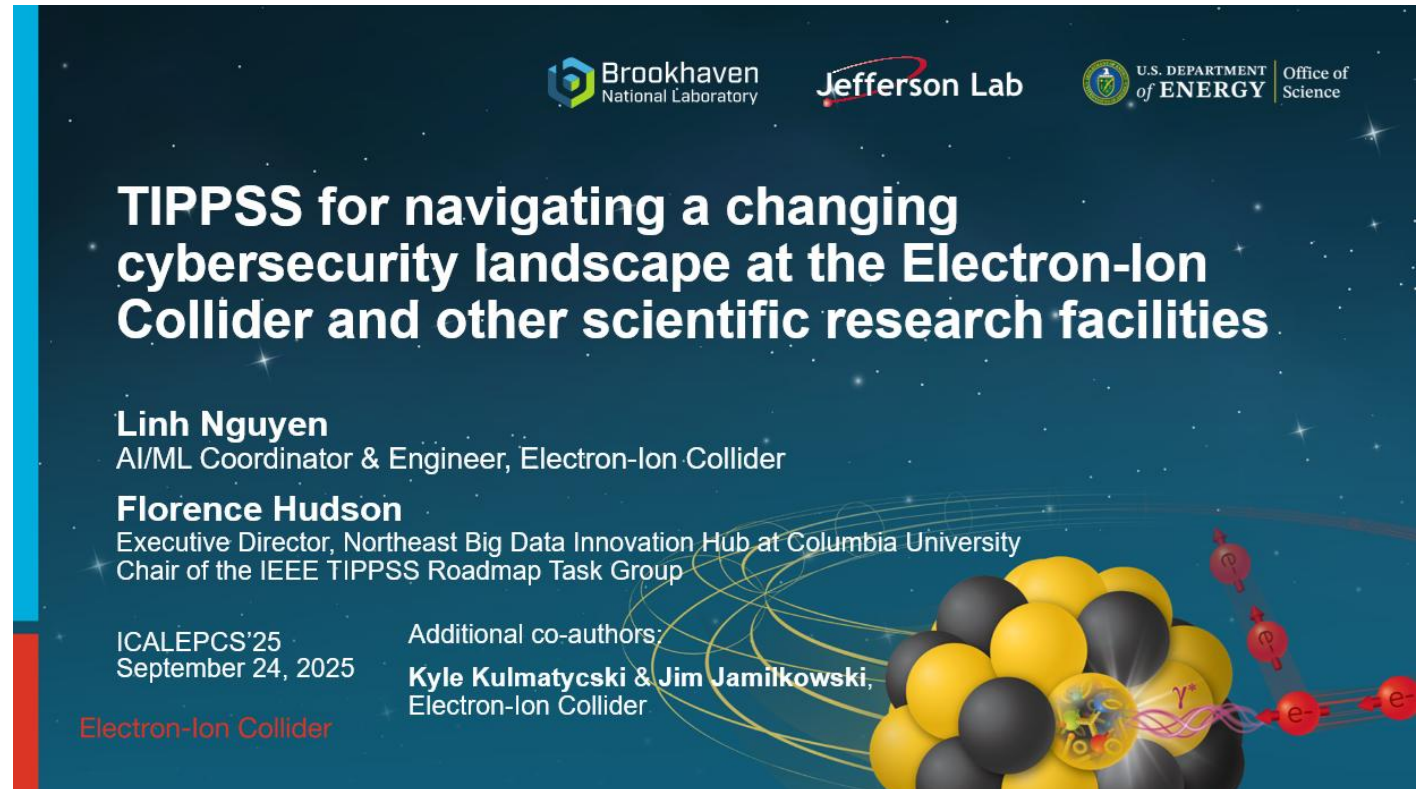


In fact, safety and security come first...

Especially because of the scale of our planned AI/ML ecosystem and the risks posed by our long timelines, the EIC has taken the lead in engineering safety and security into AI/ML ecosystems for scientific research facilities.

At ICALEPCS'25, we presented the use of the new Trust, Identity, Privacy, Protection, Safety, and Security (TIPPSS) framework from the IEEE/UL 2933 TIPPSS standard as a framework for scientific research facilities, using the EIC's planned AI/ML ecosystem as a case study.

Regular TIPPSS-based assessments will allow us to proceed in designing a safe and secure infrastructure, and robust trust and identity architecture, while we await our official AI-specific guidelines.

The slide features a dark blue background with a starry pattern. At the top right are logos for Brookhaven National Laboratory, Jefferson Lab, and the U.S. Department of Energy Office of Science. The title 'TIPPSS for navigating a changing cybersecurity landscape at the Electron-Ion Collider and other scientific research facilities' is in large white font. Below it, the authors 'Linh Nguyen' (AI/ML Coordinator & Engineer, Electron-Ion Collider) and 'Florence Hudson' (Executive Director, Northeast Big Data Innovation Hub at Columbia University; Chair of the IEEE TIPPSS Roadmap Task Group) are listed. The event 'ICALEPCS'25 September 24, 2025' and 'Additional co-authors: Kyle Kulmatycki & Jim Jamilkowski, Electron-Ion Collider' are also mentioned. A red 'Electron-Ion Collider' label is at the bottom left. On the right, there is a 3D visualization of a particle collision with yellow and black spheres and red arrows, labeled with γ^* and e^- .

Brookhaven National Laboratory Jefferson Lab U.S. DEPARTMENT of ENERGY Office of Science

TIPPSS for navigating a changing cybersecurity landscape at the Electron-Ion Collider and other scientific research facilities

Linh Nguyen
AI/ML Coordinator & Engineer, Electron-Ion Collider

Florence Hudson
Executive Director, Northeast Big Data Innovation Hub at Columbia University
Chair of the IEEE TIPPSS Roadmap Task Group

ICALEPCS'25
September 24, 2025

Additional co-authors:
Kyle Kulmatycki & Jim Jamilkowski,
Electron-Ion Collider

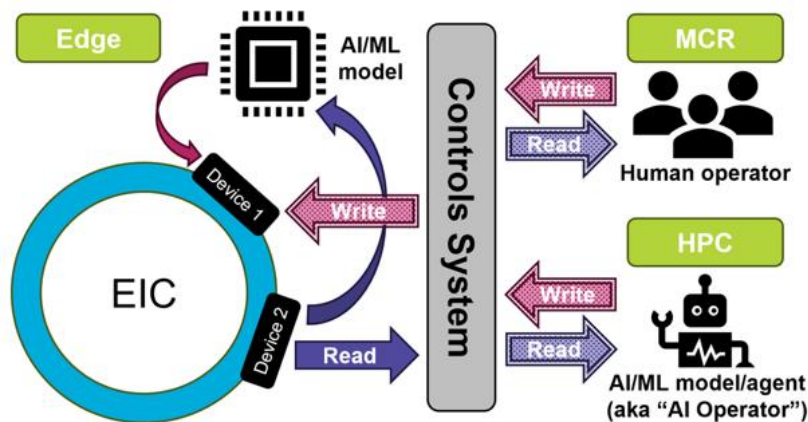
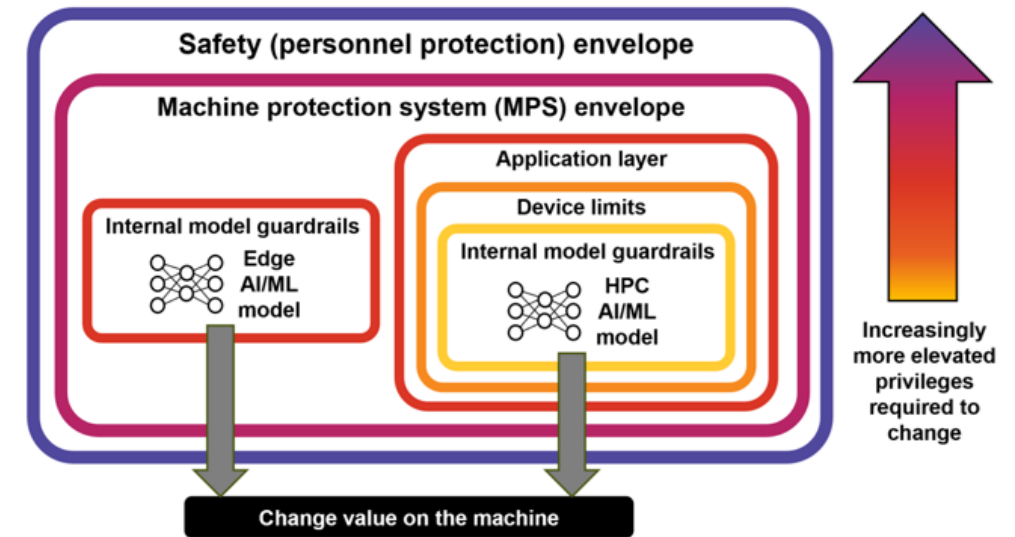
Electron-Ion Collider

Presented during the 20th International Conference on Accelerator and Large Experimental Physics Control Systems (ICALEPCS'25),
Sept. 20 - 26, 2025, Chicago, IL

Paper is currently available as part of the pre-press proceedings:
<https://epics.anl.gov/icalepcs-2025/pdf/WEBG004.pdf>

Some key points from the presentation/paper

- AI/ML layers will be treated as though they are not there when it comes to ownership and responsibility—regardless of the origin of the model. Trusted humans confer some of their trusted status to AI/ML models/agents.
- A deployed model or agent must be elevated to the status of a “trusted service”.
- Under no circumstances will AI/ML models or agents be given elevated privileges. Instead, they operate within the safety and security envelope established by trusted humans with elevated privileges.
- Access to operational model parameters should be restricted to prevent misuse of that information (e.g., for crafting highly effective adversarial attacks).
- AI/ML-generated commands and requests should always be logged and fully transparent to users, with unique identifiers and version control.
- No pipelines can exist that bypass the controls related to physics data management plans.

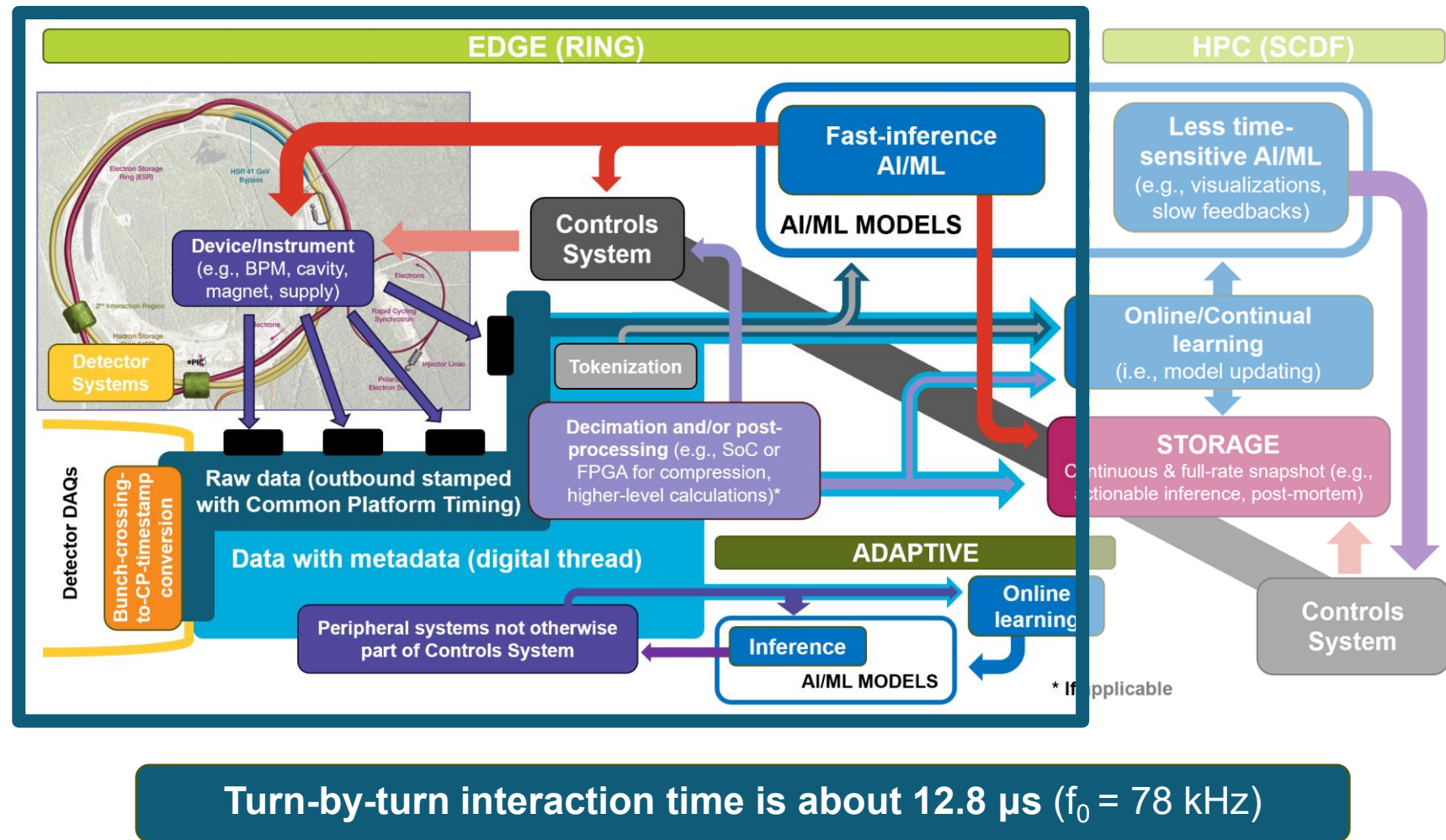


- For agentic operations, safeguards should exist to account for PII and other sensitive information inadvertently entering electronic logbooks, etc.
- Our normal machine protection system (MPS) procedures and processes already cover misbehaving AI/ML as "human error", but redundancy will be required on the AI/ML side because MPS trips never come for free.
- Under no circumstances will an AI/ML model or agent be allowed to affect a safety system.
- The machine must always be able to work without AI/ML in case an AI/ML model or agent must be disconnected for troubleshooting or in response to a cybersecurity threat.

Paper is currently available as part of the pre-press proceedings:
<https://epics.anl.gov/icalepcs-2025/pdf/WEBG004.pdf>

Edge AI/ML in the overall ecosystem

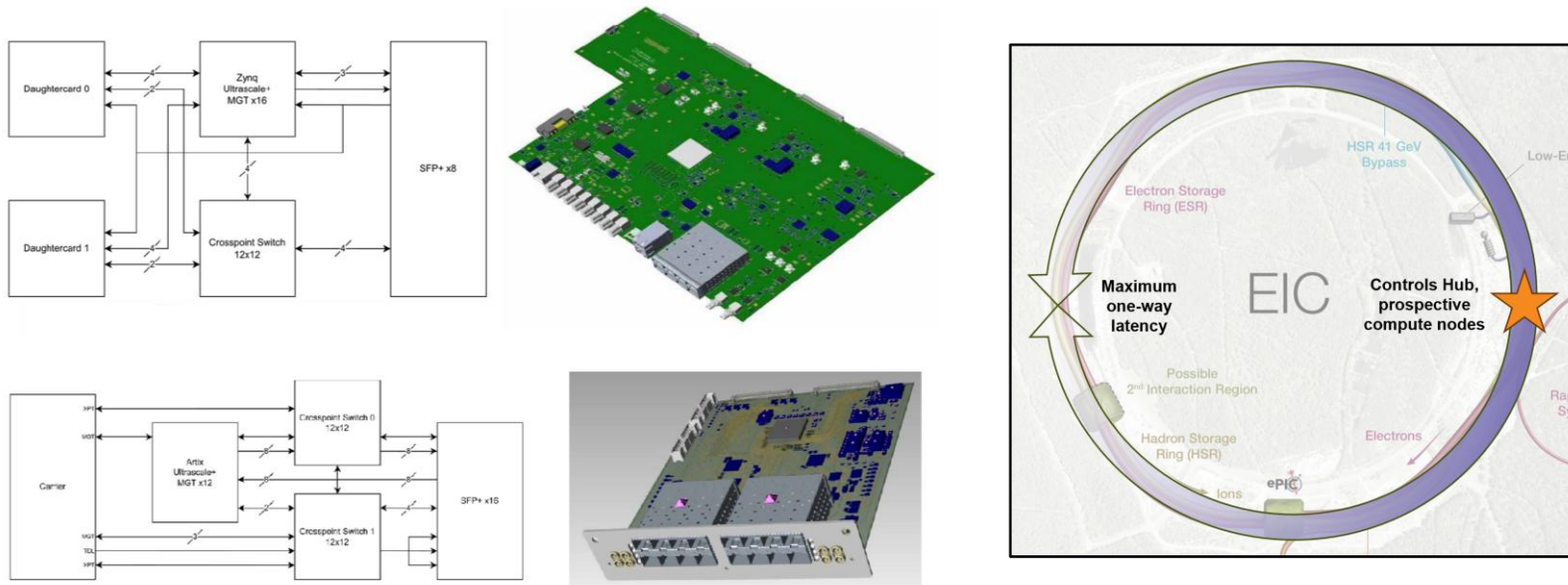
- We use “**Edge**” to refer to low-latency pipelines where the data remain at the physical EIC and never require interhost communication. This is in contrast with “HPC”.
- Edge capabilities are required for AI/ML to address use cases and operational challenges that manifest themselves on fast timescales and/or are only resolvable via fast interaction times with the machine. **Often, these are issues for which there are poor or no traditional options.**
- Work is currently being done to compile and prioritize these potential use cases.



Common Hardware Platform

Edge AI/ML capabilities will be enabled by the EIC Common Hardware Platform (CHP), which is the standardized hardware platform that is serving as the basis for the EIC controls system.

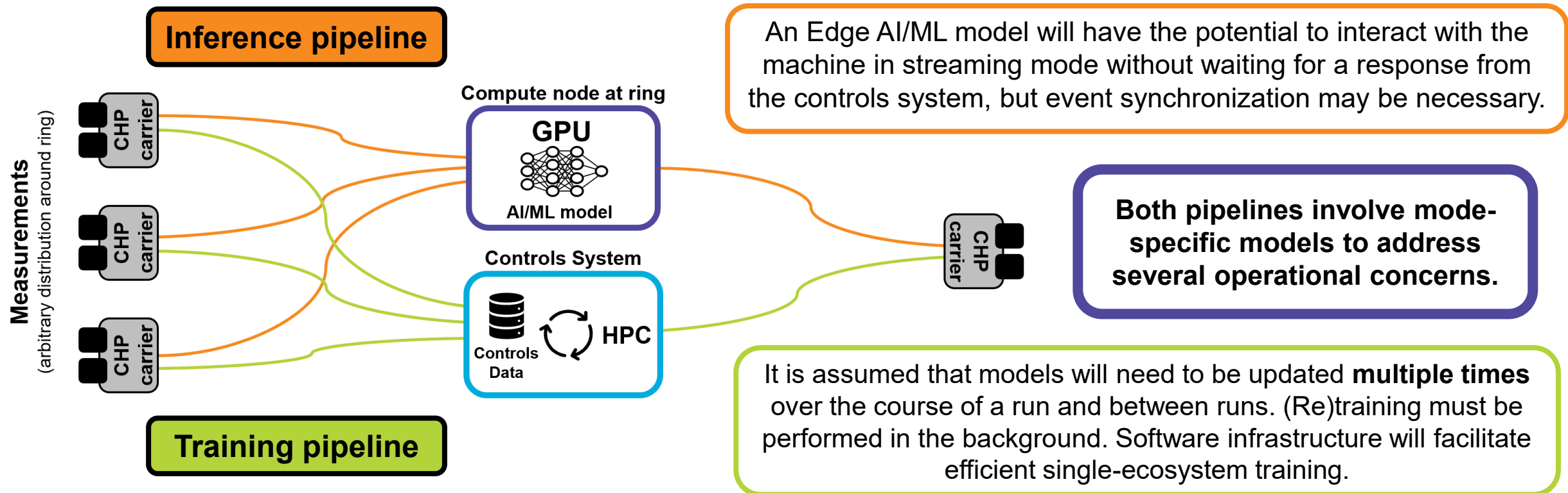
- Low-latency deterministic timing data is distributed to connected devices by the timing data link (TDL), which will help define the single ecosystem by providing the aligned timestamps. However, the TDL will often be bypassed in the EIC's most high-performance Edge AI/ML. For these cases, SFP+ ports on the CHP carrier will be used for point-to-point direct connections, enabling streaming AI/ML applications.



- **For an Edge AI/ML application to be deemed feasible, the entire pipeline must be completed within the window dictated by any synchronized event(s).** Latencies will depend on compute node location(s), so the consequences of infrastructure decisions on future potential use cases must be well understood. **This is a focus of current work.** Compute latencies must then be added to this.

Overview of Edge AI/ML architecture

Edge AI/ML models will interact with the machine locally but still involve HPC as part of their lifecycle. Specifically, they will rely on HPC during the training phase. This leads to a second distinction between pipeline types within the EIC AI/ML ecosystem: **inference pipelines** and **training pipelines**. Below is an illustrative example.



An organization-wide effort

A facility is more than its physical structures and systems—it also comprises all its people. For the EIC to be a large-scale AI-ready state-of-the-art facility therefore also means that personnel are **trained** and **empowered** to pursue potential AI/ML use cases across the organization.

Introduction to EIC
AI/ML unified vision

Survey of domain-
related use cases

Training & use case
characterization

RF Systems

Instrumentation

Cryogenic Systems

Infrastructure

Power Supplies

Accelerator Design
Accelerator Physics
Pre-Operations

**These are the
interfacing groups
established so far,
and the process is
always tailored to
the interests and
schedule of each
group**

A less EIC/group-specific training effort is also taking place among engineers and technicians, currently broken up into “Beginner” and “Non-beginner”

Defining the problem and having high-quality data come first, so the focus has been on developing broader awareness of data-driven approaches and gaining an intuition for how AI/ML can help with problems characterized by:

- Nonlinearities
- Unknown functions/ lack of analytical models
- Insufficient speed/compression
- Latent/hidden variables
- Excessive data volume
- High dimensionality
- Uncertainty
- Non-numeric parameters or decisions
- Diversity of scales
- Noise/disturbances

Conclusion

- There is a high-level vision and strategy unifying internal EIC AI/ML-related efforts. The incorporated flexibility is intended to allow for the possible integration of external AI/ML-related efforts later and advances in models and techniques.
- Especially unique to the planned EIC AI/ML ecosystem will be the possibility of large-scale AI/ML applications involving both accelerator and detector systems with high-quality data, and specialized compute resources and hardware for optimized edge applications.
- Owing to project timeline, current work is focused on physical infrastructure, hardware, the Edge AI/ML ecosystem, cybersecurity, and productivity enhancements.
- AI/ML training is also an integral part of our internal effort and will continue over the course of the project.

Thank you for your attention!
If you are interested in collaborating,
you may reach out to me at lnguyen@bnl.gov

Questions?

Back-Up Slide

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 EIC Project delivery, but they will enhance EIC Operations and research capabilities.