Brookhaven National Laboratory

Jefferson Lab

U.S. DEPARTMENT of ENERGY | Office of Science

A Unified Vision of AI/ML at the Electron-Ion Collider

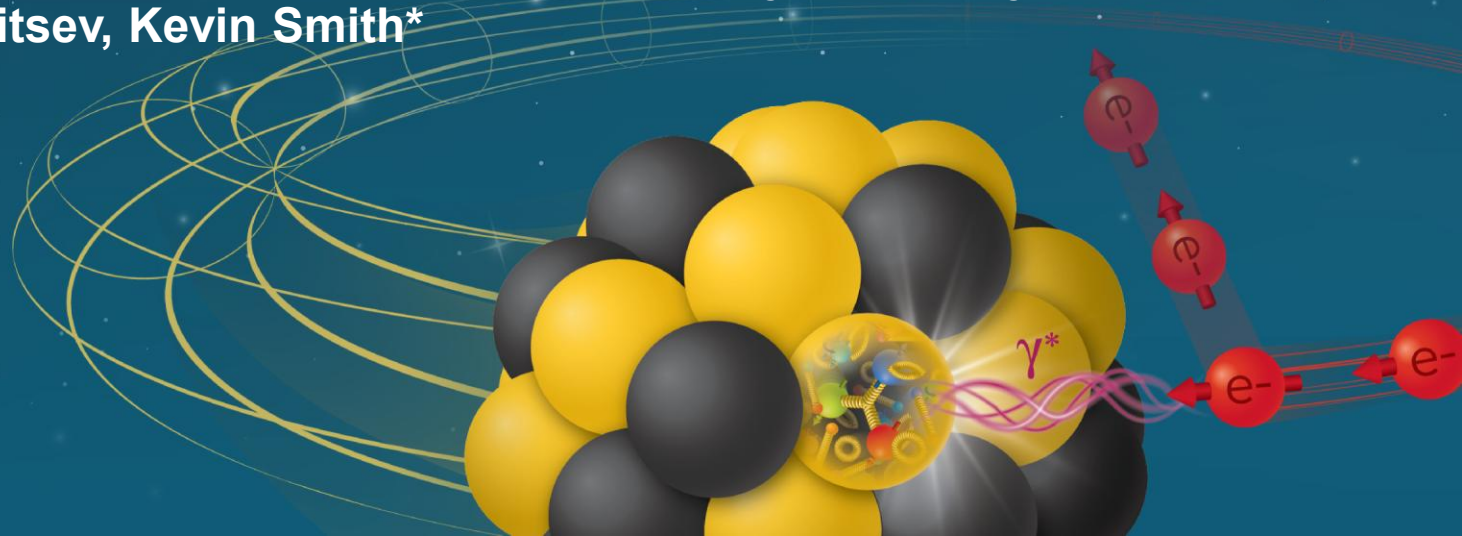
Linh Nguyen

With special thanks to: **Elke Aschenauer, Paul Bachek*, Jim Jamilkowski, Kyle Kulmatycski, Jeff Landgraf, Daniel Marx, Sergei Nagaitsev, Kevin Smith***

* Legacy contributions

AI4MuC Mini-Workshop
June 30th, 2026

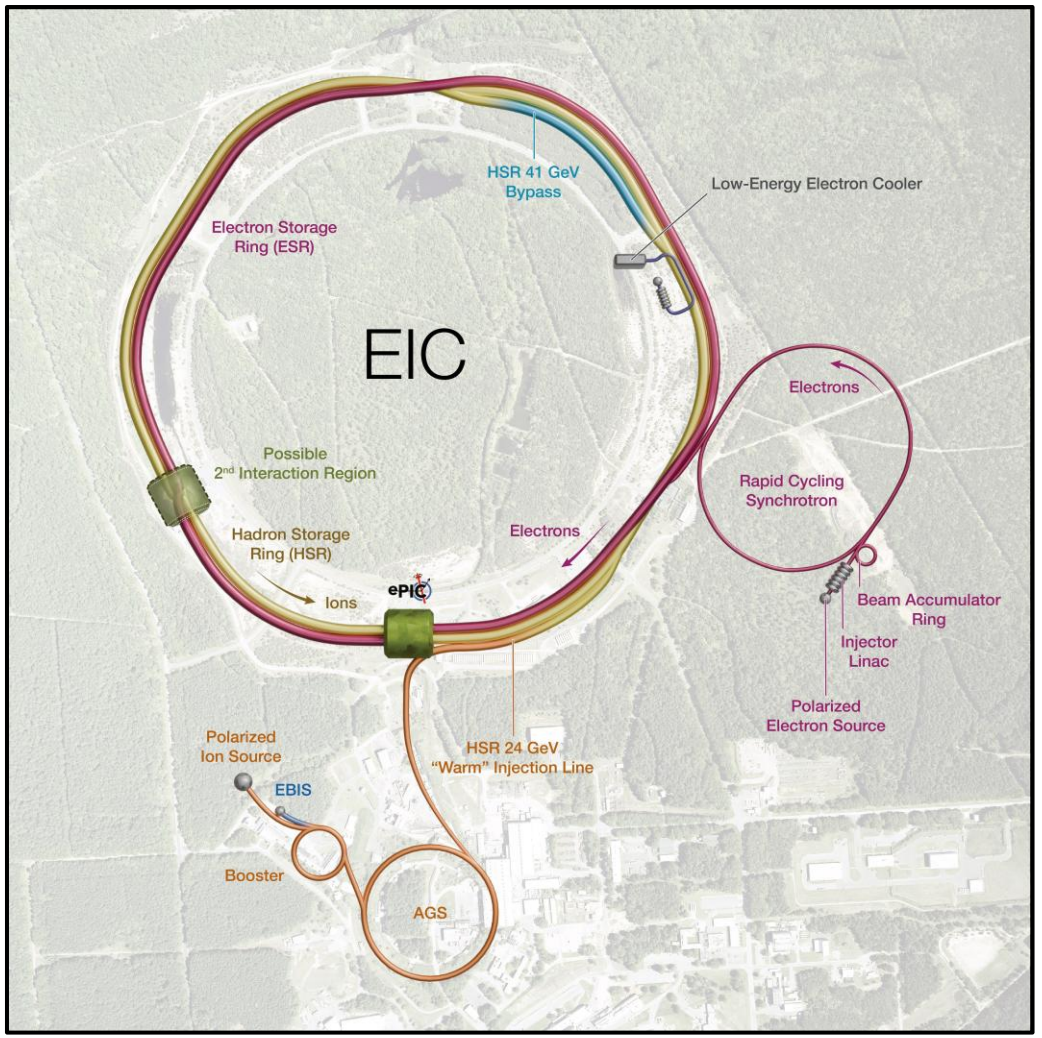
Electron-Ion Collider



The Electron-Ion Collider (EIC)

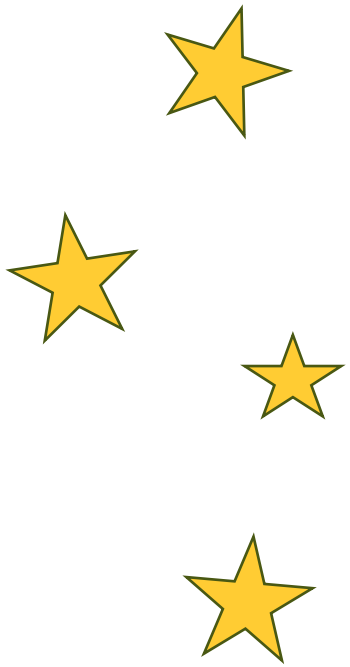
- The EIC is being built at Brookhaven National Laboratory (BNL) in partnership with Jefferson Lab and other domestic and international partners.
- It will reuse the hadron injector chain and 3.8-kilometer tunnel of the Relativistic Heavy Ion Collider (RHIC), which concluded its program on February 6th after 25 years. However, **the EIC is a major construction project**. Removals & Repurposing (R&R) of the former RHIC facility is currently ongoing.
- **Construction of the EIC is scheduled to begin in early 2027. Operations is scheduled to begin around 2036.**
- Given that it is still in the design phase, the EIC uniquely presents an entire AI/ML lifecycle for a large-scale scientific research facility. As such, it could uniquely explore and exploit the possibilities of AI/ML for science by having AI readiness deliberately incorporated into its design.

This creates a tremendous opportunity for all of us, and we are actively planning and strategizing about how to make the most of it…

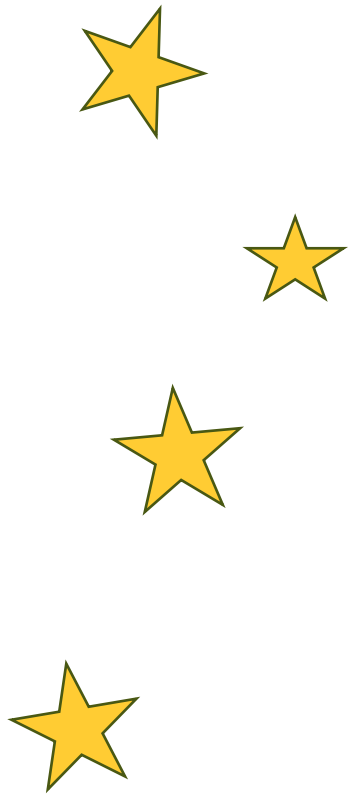


A unified vision of AI/ML at the EIC

From the EIC Global Requirements (Controls Section) document:



"The new controls system will be AI-ready, with comprehensive data acquisition across multiple time scales, sufficient computational and network resources, and standardized interfaces to support machine learning-driven diagnostics, prognostics, optimization, machine protection, collimation, and control functions. The system will also support integration with a virtual accelerator/digital twin, AI training, and safe development of control strategies."



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:

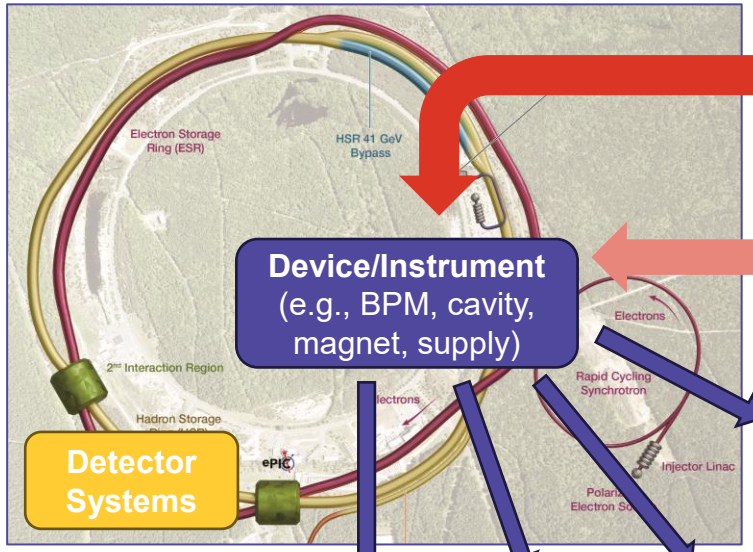
Edge AI/ML Capabilities	The EIC will have some of the most demanding AI/ML applications in the world in terms of latencies, data rates, and throughput. This requires specialized local compute resources and optimized infrastructure.
End-to-End AI/ML Capabilities	Unlike in the past, the Detector and Accelerator 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 and the organization's internal training initiative.

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 2036, with a program lasting at least 15 years. To hedge against obsolescence, it must be able to accommodate new AI/ML models and approaches 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, including cybersecurity.

Overview of EIC AI/ML infrastructure plans



EDGE (RING)

HPC (SCDF)

Device/Instrument (e.g., BPM, cavity, magnet, supply)

Detector Systems

Controls System

Fast-inference AI/ML

AI/ML MODELS

Less time-sensitive AI/ML (e.g., visualizations, slow feedbacks)

Online/Continual learning (i.e., model updating)

Tokenization

Decimation and/or post-processing (e.g., SoC or FPGA for compression, higher-level calculations)*

STORAGE
Continuous & full-rate snapshot (e.g., actionable inference, post-mortem)

Detector DAQs

Bunch-crossing-to-CP-timestamp conversion

Raw data (outbound stamped with Common Platform Timing)

Data with metadata (digital thread)

ADAPTIVE

Online learning

Peripheral systems not otherwise part of Controls System

Inference

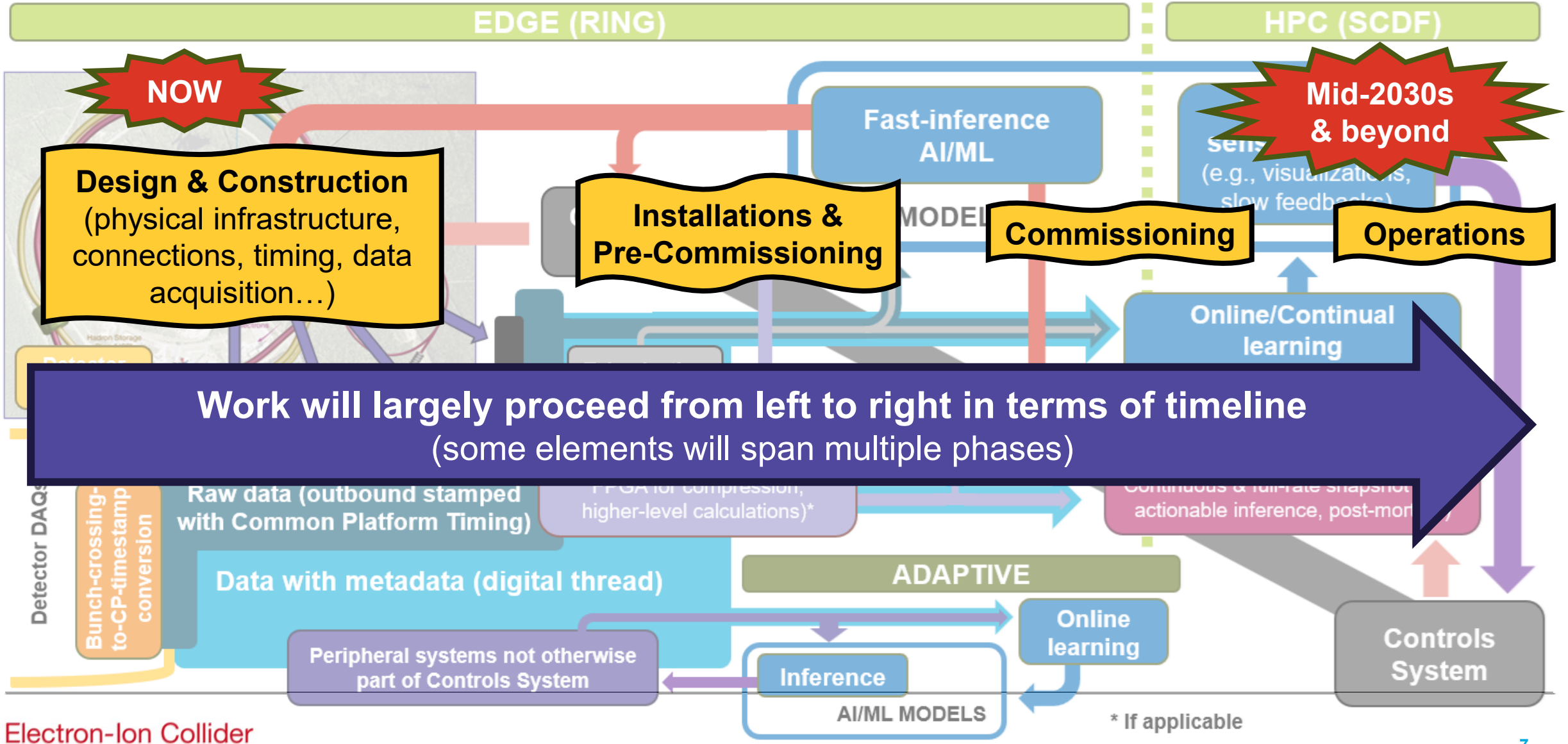
AI/ML MODELS

Controls System

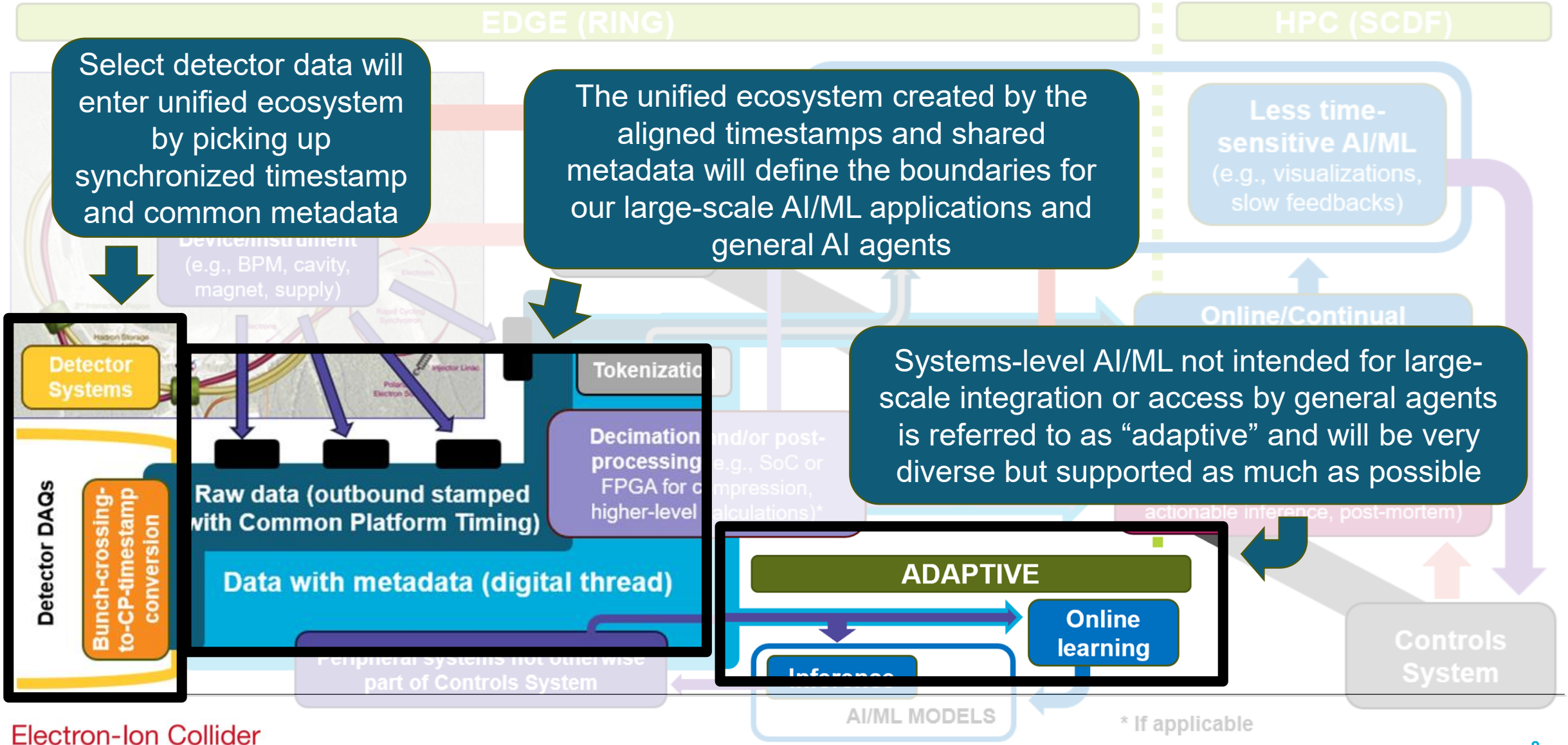
* If applicable

Electron-Ion Collider

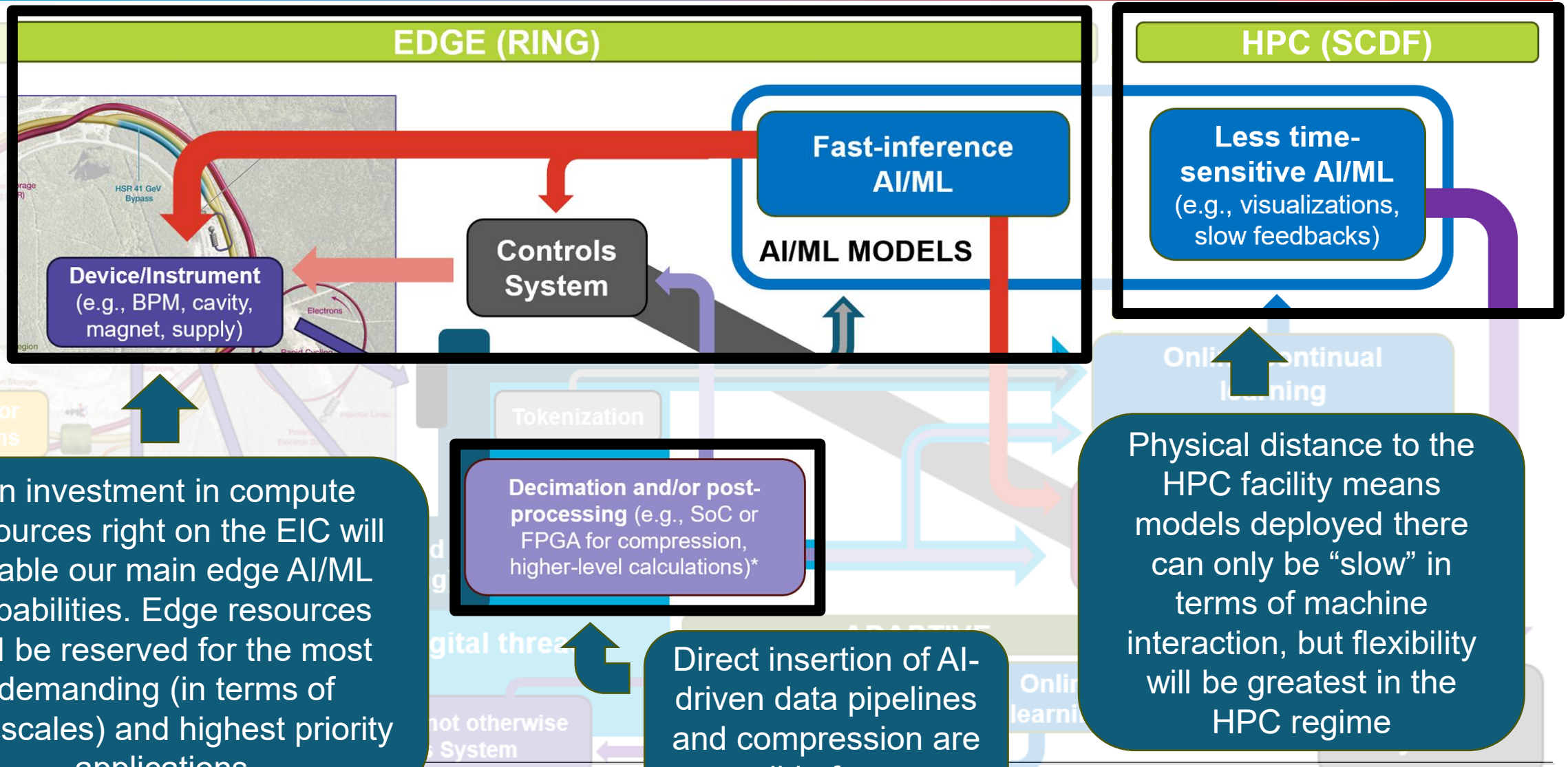
Overview of EIC AI/ML infrastructure plans



Overview of EIC AI/ML infrastructure plans

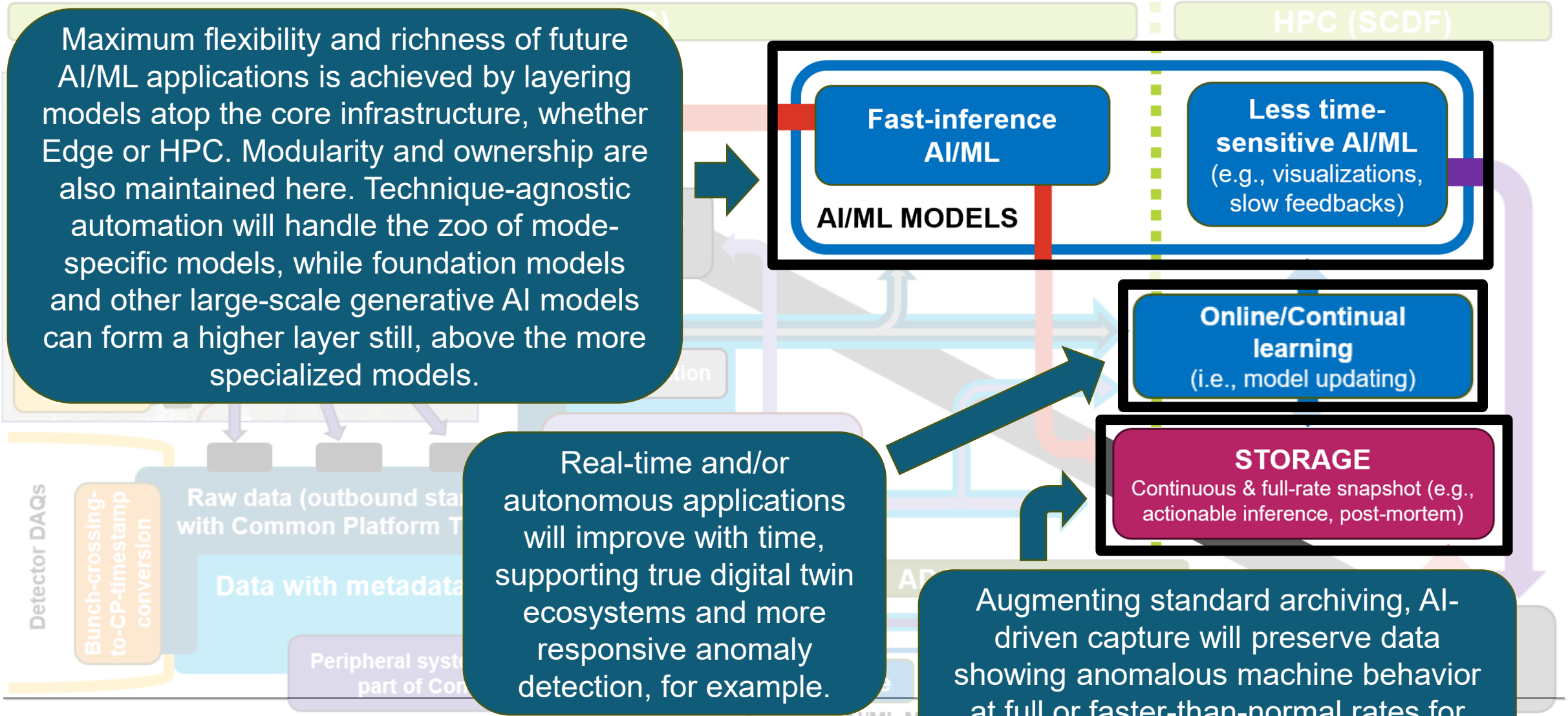


Overview of EIC AI/ML infrastructure plans



Overview of EIC AI/ML infrastructure plans

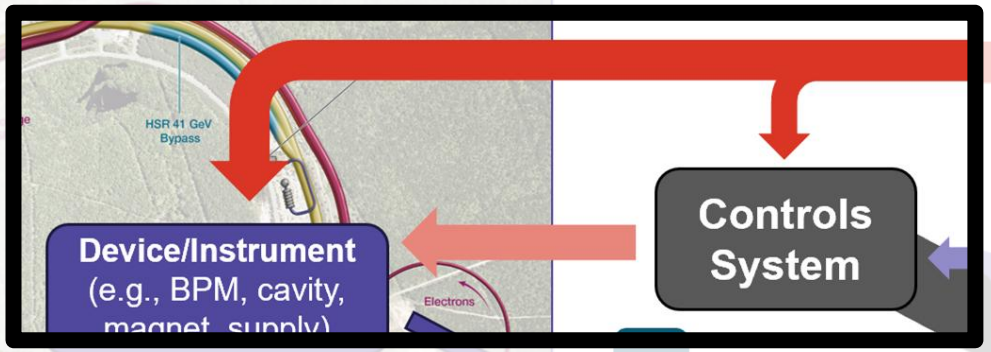
Maximum flexibility and richness of future AI/ML applications is achieved by layering models atop the core infrastructure, whether Edge or HPC. Modularity and ownership are also maintained here. Technique-agnostic automation will handle the zoo of mode-specific models, while foundation models and other 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, supporting true digital twin ecosystems and more responsive anomaly detection, for example.

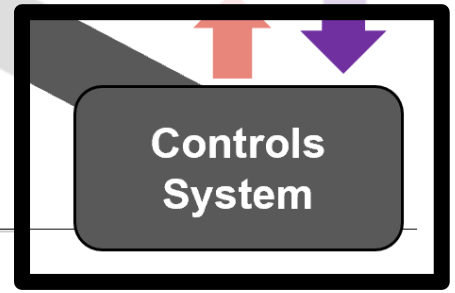
Augmenting standard archiving, AI-driven capture will preserve data showing anomalous machine behavior at full or faster-than-normal rates for expert review and AI/ML datasets

Overview of EIC AI/ML infrastructure plans



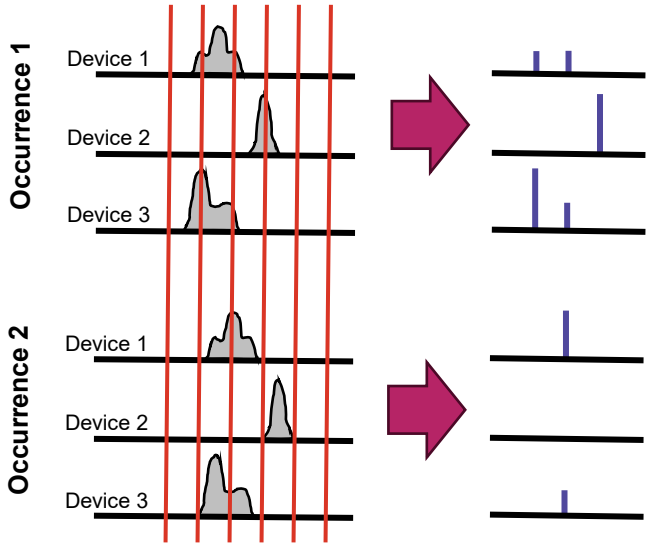
Commands sent by HPC AI/ML models will follow the same channels as operator commands, making them subject to the same safety and security controls, and provenance tracking. Edge AI/ML models are functionally identical to local feedbacks and other local controls and will follow a rigorous deployment procedure.

Cybersecurity and safety are priorities throughout the facility and throughout the AI/ML lifecycle



Spotlight on AI readiness: High-quality datasets for AI/ML

- Failure to acquire data at adequate rates tends to mask the behavior and causality of underlying phenomenon. This is especially problematic for high-performance Edge AI/ML applications. Artificially aligning data (due to asynchronous acquisition) can create a similar problem.



Despite the same underlying phenomenon, an inadequate data rate results in a poor training dataset. For example, the role of Device 2's signal or a correlation based on the relative strength of Device 1's signal and Device 3's signal could be hard to learn. Much more training data would be needed for such patterns to reveal themselves, and inference would likely still be faulty.

An adequate global base rate with globally aligned timestamps is therefore required for high-quality datasets for use in model training.

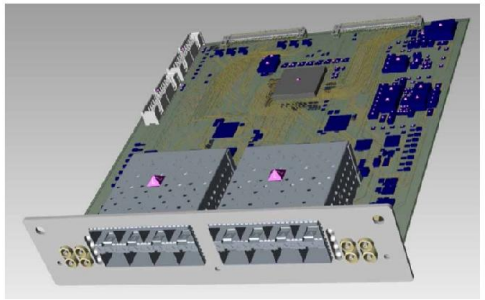
- Post-mortem capture (where high-rate data are captured around a recorded trip) is better than low base-rate capture but results in high-rate data only involving actual positives. This is problematic because AI/ML applications (e.g., anomaly detection) have different tolerances towards false negatives vs. false positives, etc.

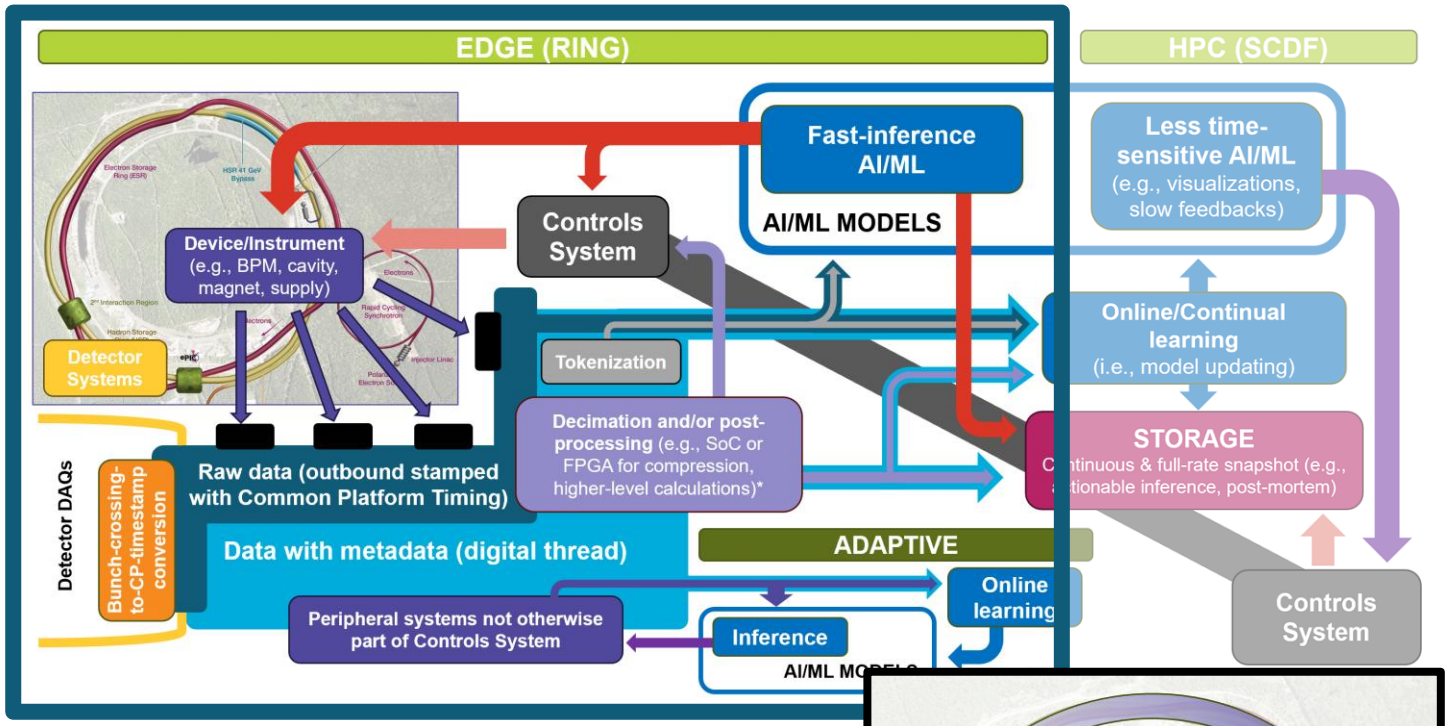
		Predicted class: Positive	Predicted class: Negative	
Actual class	Positive	True Positive (TP)	False Negative (FN)	**Recall/Sensitivity** $\frac{TP}{TP+FN}$
Actual class	Negative	False Positive (FP)	True Negative (TN)	**Specificity** $\frac{TN}{TN+FP}$
		Precision $\frac{TP}{TP+FP}$	**Negative Predictive Value** $\frac{TN}{TN+FN}$	**Accuracy** $\frac{TP+TN}{TP+TN+FP+FN}$

Datasets need to be balanced to mitigate bias and allow for options in model metrics and tuning.

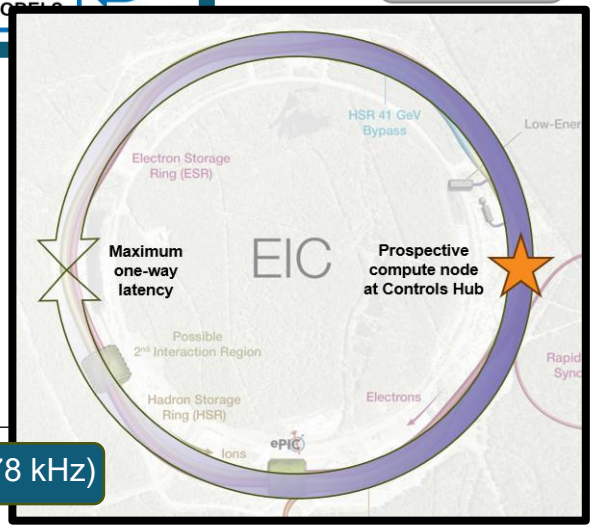
Spotlight on AI readiness: Architecture for Edge AI/ML

- We use "**Edge**" to refer to low-latency pipelines where the data remain at the physical EIC and do not require interhost communication. This is in contrast with "HPC".
- Edge capabilities are required for AI/ML to address use cases and operational challenges involving very fast timescales. **Often, these are issues for which there are poor or no traditional options.**
- Edge AI/ML capabilities will be predominantly 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. Ports on the CHP carrier can be used for point-to-point direct connections, enabling streaming AI/ML applications. ↓





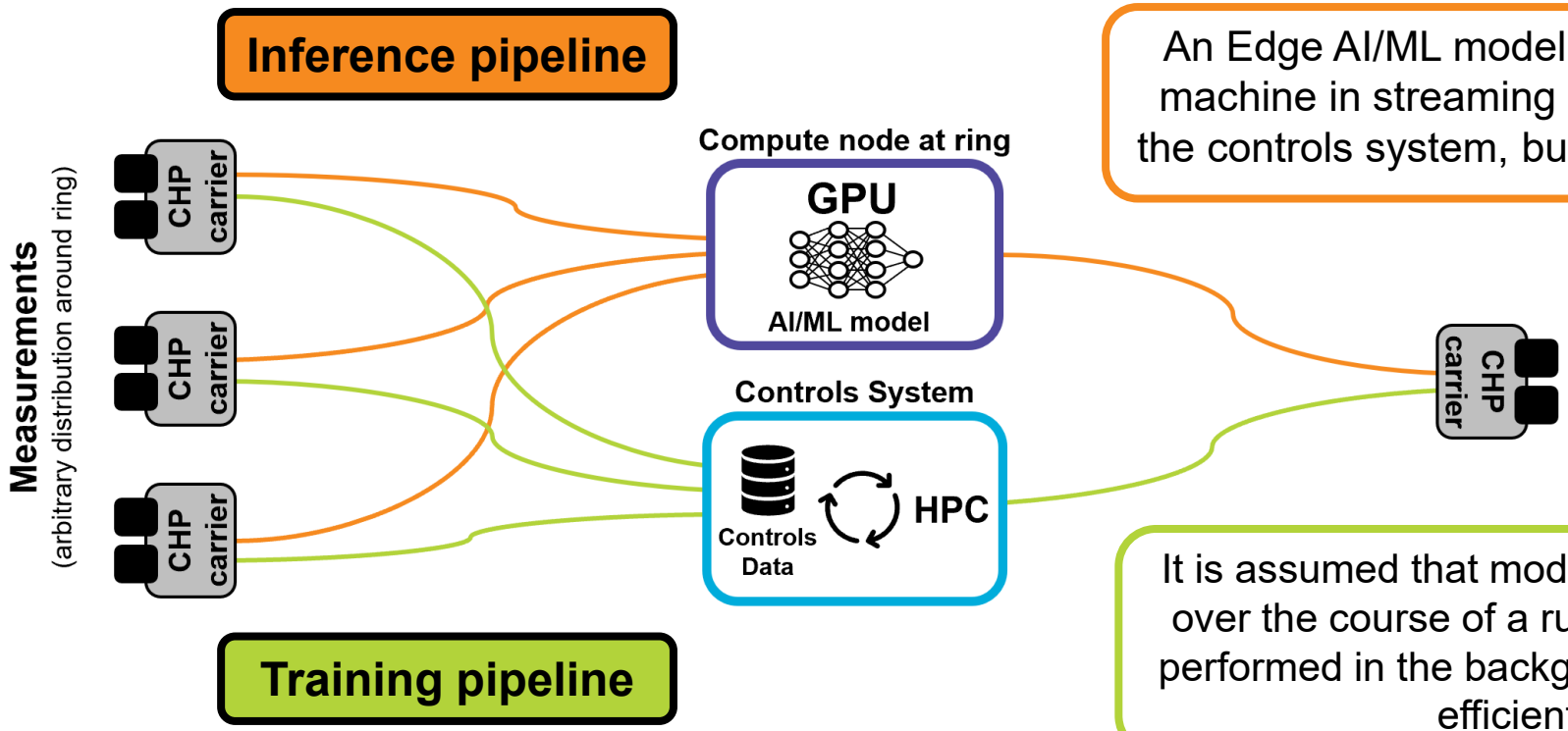
For an Edge AI/ML application to be feasible, the entire pipeline must be completed within the window dictated by any synchronized event(s). Latencies will depend on compute node location(s), so the consequences of infrastructure decisions on future potential use cases must be well understood. Compute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



Turn-by-turn interaction time is about 12.8 μs (f_0 = 78 kHz)

Spotlight on AI readiness: Architecture for Edge AI/ML

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 Edge AI/ML model will have the potential to interact with the machine in streaming mode without waiting for a response from the controls system, but event synchronization may be necessary.

Both pipelines involve mode-specific models to address several operational concerns.

It is assumed that models will need to be updated **multiple times** over the course of a run and between runs. (Re)training must be performed in the background. Software infrastructure will facilitate efficient single-ecosystem training.

See “Plans and strategy for edge AI/ML at the Electron-Ion Collider at Brookhaven National Laboratory” in ICALEPCS’25 proceedings: https://proceedings.jacow.org/icalepcs2025/pdf/WEMR003.pdf

Spotlight on AI readiness: Cybersecurity in the Age of AI

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 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.

Paper & slides available as part of ICALEPCS'25 proceedings:
https://proceedings.jacow.org/icalepcs2025/pdf/WEBG004.pdf

This work has directly led to an effort to create an **IEEE TIPPSS-based standard for scientific research infrastructure**, which would be an **international independent standard**. I am leading the Study Group for the Project Authorization Request (PAR) and will transition to Working Group Chair (with Florence Hudson as Working Group Vice Chair) of the standard upon approval of the PAR by the IEEE Standards Association. This is meant to be a resource **for all**.

Key points from the initial TIPPSS assessment

TRUST

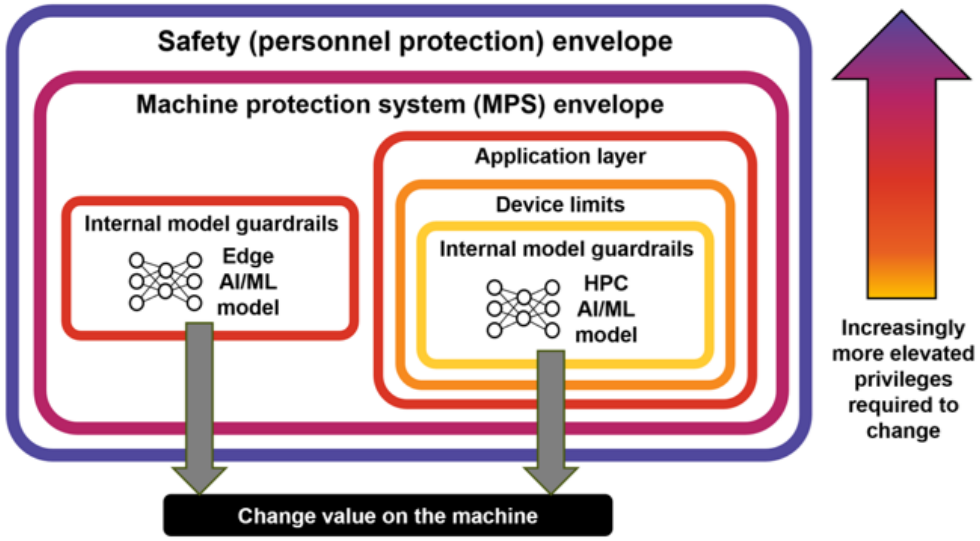
- AI/ML layers will be treated as though they are not there when it comes to ownership and responsibility—regardless of the origin of the model. Trusted humans confer some of their trusted status to AI/ML models/agents.
- A deployed model or agent must be elevated to the status of a “trusted service”.
- 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.

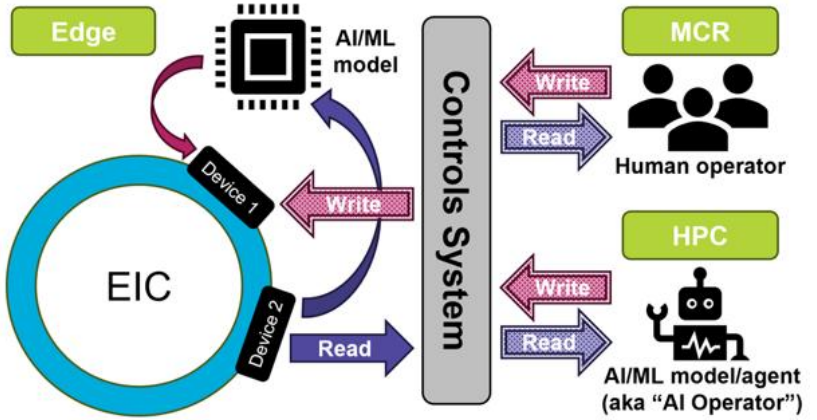
IDENTITY

- 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.

PRIVACY

- 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.

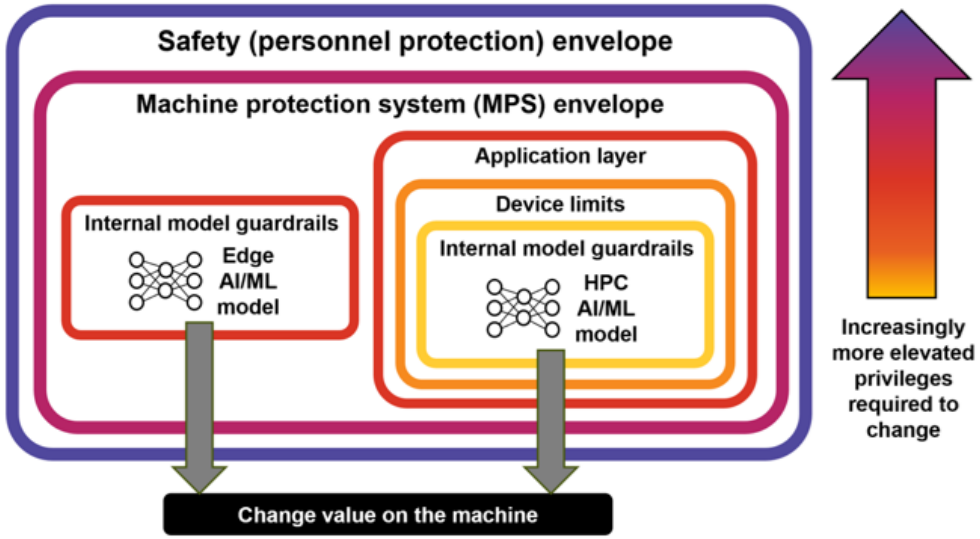


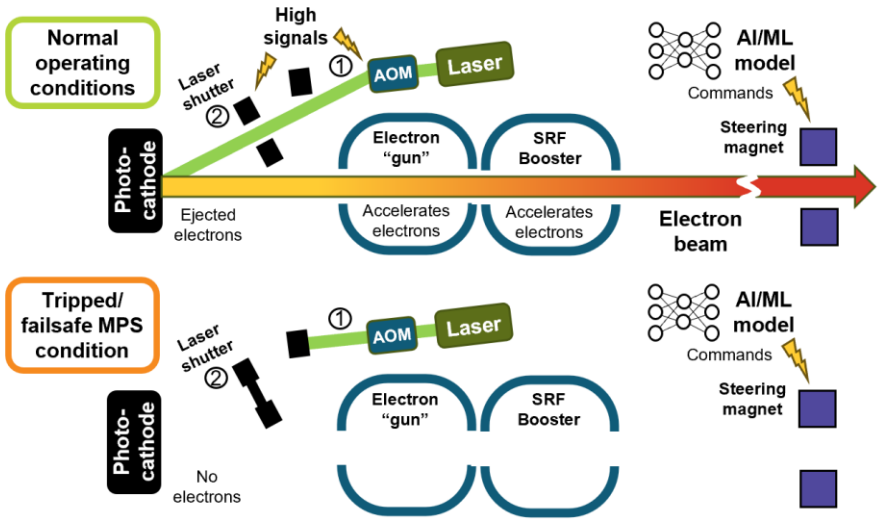


Key points from the initial TIPPSS assessment

PROTECTION

- 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.
- Automatic deactivation in response to unexpected or out-of-design conditions, followed by an alert with diagnostic information to an expert, should be leveraged whenever possible to safeguard against unusual AI/ML behavior.





SAFETY

- Under no circumstances will an AI/ML model or agent be allowed to affect a safety system. **Human safety is the strongest, most encompassing envelope (see diagram above)**.

SECURITY

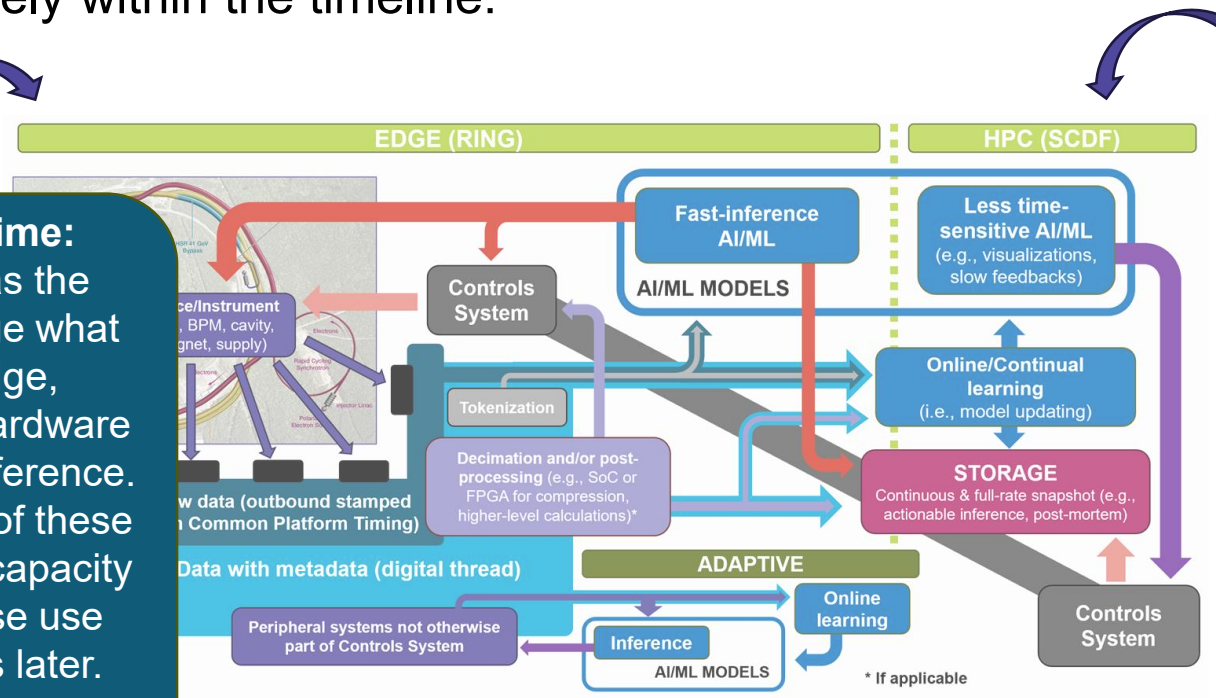
- The machine must always be able to work without AI/ML in case an AI/ML model or agent must be disconnected for troubleshooting or in response to a cybersecurity threat.

See paper in ICALEPCS'25 proceedings for more complete report: https://proceedings.jacow.org/icalepcs2025/pdf/WEBG004.pdf

The unified vision at work: A possible roadmap into Quantum

Lastly, the field of AI is bracing for the collision with quantum computing (QC) – and so are we. The unified vision allows us to more readily identify where potential intersections might be accommodated for growth into the space, what the impacts to infrastructure might be, and how to place our response appropriately within the timeline:

QC's affect on Edge regime: Quantum parallelization has the potential to drastically change what gets to be considered Edge, especially when used as a hardware accelerator for low-latency inference. However, to take advantage of these developments, we will need capacity at the ring – otherwise, these use cases will not be open to us later. We therefore need to keep an eye on these related developments *now* and possibly plan for capacity.



QC affecting HPC regime: This physical infrastructure exists away from the ring where the restrictions on major site work post-EIC Construction are not as strict. Work here can therefore much better afford to wait out developments and respond during EIC Operations.

Like so much at the EIC, stay tuned!

Conclusion

- There is a high-level vision and strategy unifying internal EIC AI/ML-related efforts. The engineered flexibility is intended to allow for the possible integration of external AI/ML-related efforts later, as well as advances in models and approaches.
- 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.
- The long timeframe of new large scientific research facilities set against a much more rapidly changing technological landscape than in the past makes a unified vision of AI/ML indispensable for nimbleness, but few resources exist for this side of the AI/ML integration and deployment problem. Broader participation and knowledge transfer will need to occur over the coming decades. The EIC hopes to assist other endeavors, such as a possible muon collider, facing similar challenges wherever it can.

Thank you for your attention!

Feel free to reach out to me at lnguyen@bnl.gov

Questions?