This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under GA No 101004730.

Task 10.6: Machine Learning Techniques for Accelerator and Target Diagnostics

I.FAST 1st Annual meeting, 2022-05-04

Irena Dolenc Kittelmann, Thomas Shea, ESS
for the Task 10.6 team

iFAST

10.6 Task Summary

- Long term mission:
  - Develop low-latency Machine Learning (ML) techniques to improve performance and availability of high-power facilities at the intensity frontier.
- Goal:
  - Identify signatures of potential errant beam conditions
- Scope:
  - Assess the predictive capabilities of selected ML models
  - Prototype: proof of principle demonstration
    - The most promising ML model to be implemented on a low-latency network of FPGAs processing signals from array of detector channels.

Collaborators and responsibilities

Beneficiaries

- ESS
  - Expertise: Diagnostic scientists, Data manager, FPGA engineers
  - Contribution:
    - Construction/Commissioning phase
    - FPGA-based diagnostics
    - Low latency communications
- RTU – Faculty of Computer Science and Information
  - Expertise: AI, ML and FPGA experts
  - Contribution
    - Full-time PhD student assigned
    - PhD student supervision (senior researcher/professor)
  - Regular bi-weekly meetings

Additional participants

- CosyLab – via subcontract from ESS
  - FPGA specialist
- SNS/ORNL – collaboration
  - > decade of operational data
  - Active machine learning project
  - MOU to exchange data with ESS in preparation

Timeline

- Model **preparation** (Q3 2021 – Q1 2022)
  - ML model exploration
  - Data format preparation
- Assess the predictive capabilities of selected ML models, trained with (Q1 2022 – Q2 2023, **milestone Q4 2022**: selection for demo):
  - simulation results
  - existing SNS operation data (13 years of operation data)
  - ESS experimental data anticipated within the time line of the project (ESS Normal Conducting linac commissioning ongoing)
- ML FPGA platform **development** (Q2 2021 - Q4 2022)
  - Low-latency network development and optimisation
  - Acceleration of ML model on FPGA platform
- Demonstration (Q1 2023 - Q2 2024)
  - In the lab
  - In real conditions
- Report (March 2024)

Low-latency applications of interest

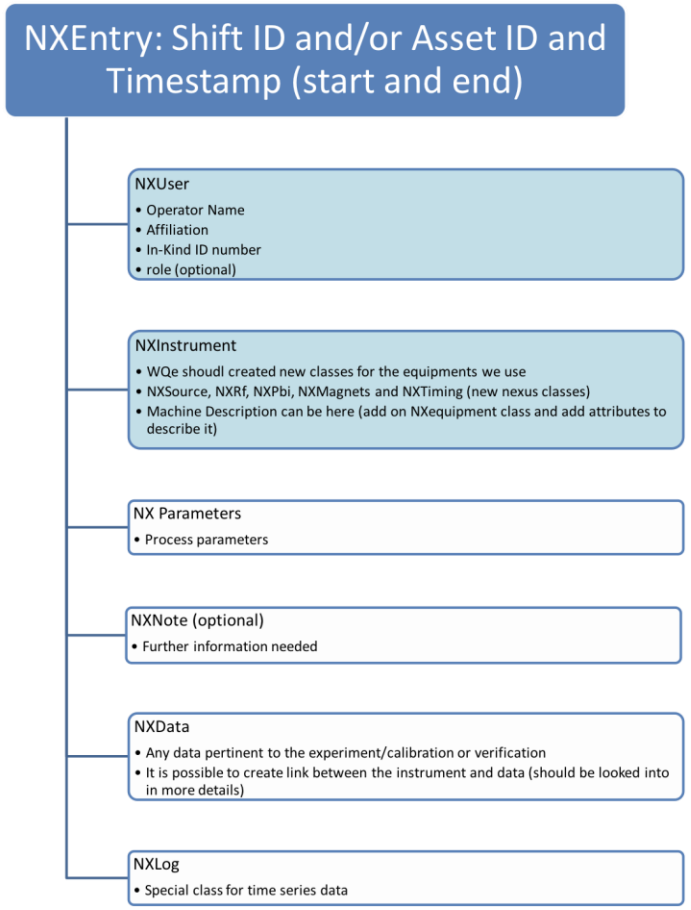
- ML based **Intelligent Trigger** for Data-On-Demand (DoD)
  - Goal:
    - Identify “off-normal/interesting” and “normal” events
    - Trigger DoD* acquisition for each “off normal” event

      Data-On-Demand (DoD) definition: partially or fully processed highly detailed data, buffered and then extracted from the FPGA level on event occurrence (on demand)
  - Relevant systems: Beam Loss, Beam Position, Beam Current measurement
- ML based **machine protection**
  - Goal: inhibit beam production when dangerous conditions are recognised/predicted
  - Relevant systems: BLM and BCM; also BPM
  - Can be considered as specific version of ”intelligent trigger” application
    - “Intelligent trigger” that is trained to recognise dangerous conditions, triggers DoD extraction (post mortem) and inhibits beam production.

Data

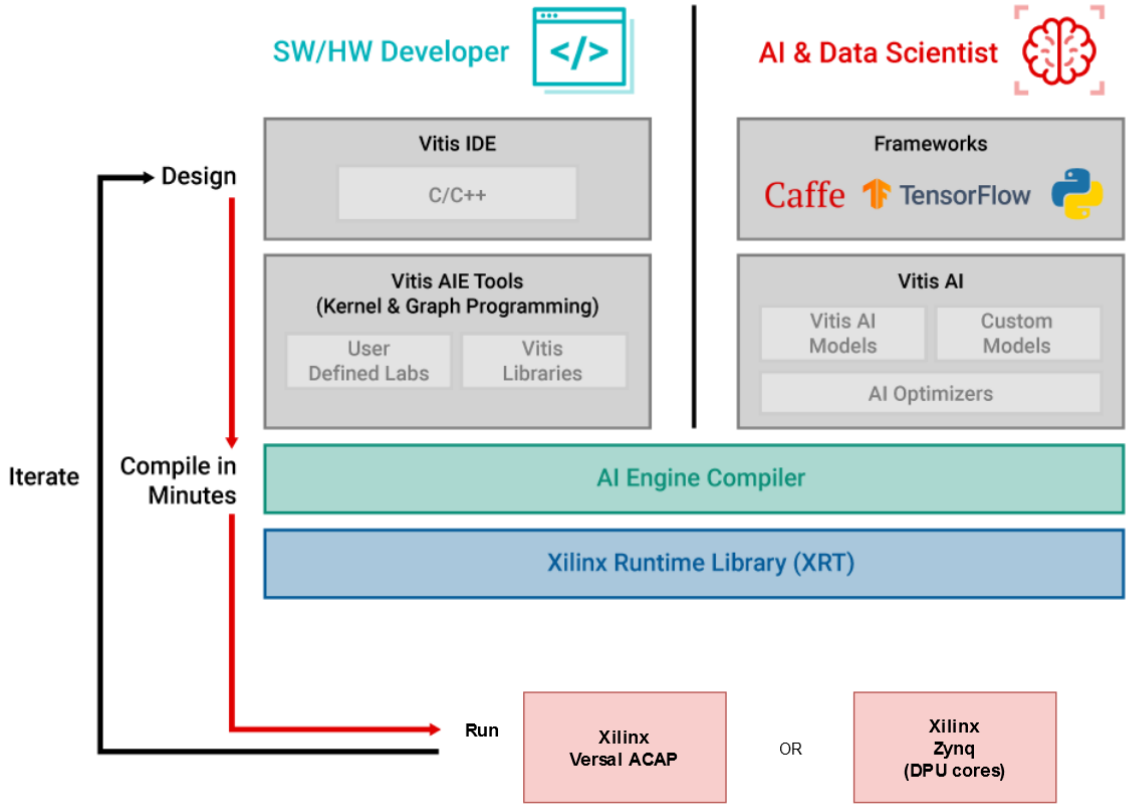
- Developing Data Exchange based on NeXus (https://www.nexusformat.org)
- Standard in the neutron and X-ray communities
- First example: agreement with ORNL on emittance data structure.
  - We will build on this to cover ML-related data that we intend to exchange
- Scope:
  - Operations data from SNS
  - Commissioning data from ESS
  - Simulation data from ESS
- Explore compatibility with SLAC data standard (standard to capture simulation results for SLAC ML team)

NeXus data format (example ESS diagnostics data)



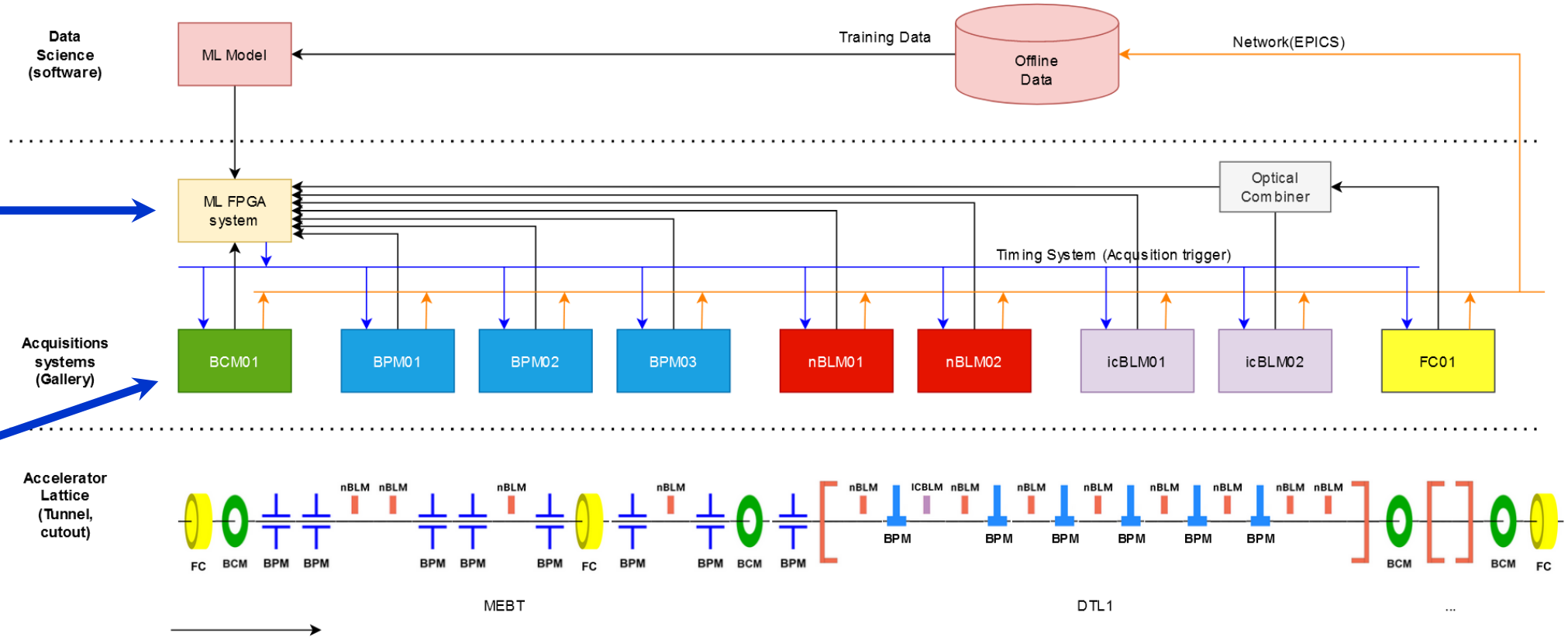
Tools

- Tools based on Xilinx AI flow accelerated on FPGA platform. Two options being investigated.
  - Xilinx Zynq: Current generation FPGA platform with AI accelerator cores in FPGA logic (Deep learning Processor Unit, DPU)
  - Xilinx Versal ACAP: Bleeding edge FPGA with dedicated AI acceleration cores
- The tool flow would provide short turn-around and flexibility when exploring ML applications
- The downside is that the solution includes compiled SW running on accelerator cores which means additional latency.
  - On Xilinx Zynq there is SW (XRT) in the loop to feed the accelerator cores.
  - On Xilinx Versal the programming model supports feeding AI cores directly from FPGA logic so should provide lower latency.



ML Prototype – ESS Normal Conducting Linac, Low Energy Section

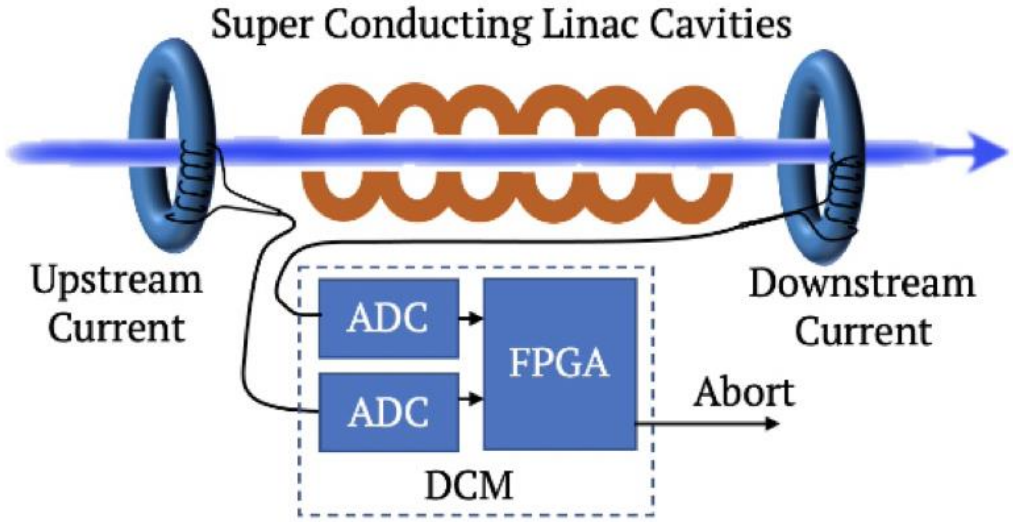
- System covering Normal Conducting Linac (NCL)
- 2nd layer: **ML FPGA System** to detect "off normal" events and request readout via timing system (DoD extraction).
- 1st layer: **network of FPGAs** to acquire and process signal from detector channels.
- System could connect to beam interlock system (not shown).



ORNL/SNS Example

- Goal: ML based prediction and mitigation of failures (errant beam conditions) to protect SCL from beam loss damage
- Approach: pulse by pulse prediction
  - Use Differential Beam Current Monitor (DBCM) data (100 Msamples/s): normal operation (GOOD), trips (BAD).
  - Use Random Forrest (RF) classifier to label the current pulse as GOOD/BAD.
    - Tried several other classifiers (k-nearest neighbours, log. Regression, etc.)
    - RF selected based on precision and speed
  - Time budget: ~16ms (1ms beam pulse, 60Hz)
- Conclusions from the offline study with full pulse:
  - Full pulse data: 120 000 samples (1.2 ms)
  - Failure signature
    - Imprinted already in the initial ¼ of pulse
    - Present in 4 previous pulses
  - Predictive power equally good if only one of the BCMs is used.
  - Trip prediction depends on acceptable failure (false positive) rate: 58% True Positive rate with 0% False Positive rate
- Currently at SNS:
  - RF implemented on the FPGA, under test
  - Part of the pulse data used (FPGA limitation): decision whether pulse is BAD/GOOD taken after 10,000 samples (100us).

- *M. Reščič, et al., NIMA, 1025 (2022), 166064, https://doi.org/10.1016/j.nima.2021.166064*
- *M. Reščič et al., NIMA 955 (2020), 163240, https://doi.org/10.1016/j.nima.2019.163240*



Current focus at ESS

- **Apply SNS example** with DBCM on ESS prototype:
  - To prepare the tool chain with realistic ML model in place
  - To explore HW platform limitations
    - How much of the beam pulse can we use? Is decimation or pulse trimming required?
    - Can we add more data channels (beam phase, beam loss), what is the limit?
- Plans for **upgrading** the SNS example
  - Add more data channels
- **ESS data**: select a few ML models and explore their predictive power and limitations
  - Model performance dependence on coverage along the linac (which set of channels gives adequate performance, can some channels be skipped?)
  - Time window and position inside the beam pulse (how many samples inside beam pulse, where inside beam pulse?)
- Toy data: **simulation** of beam envelope under normal conditions and faults
  - To explore sensitivity to errant conditions

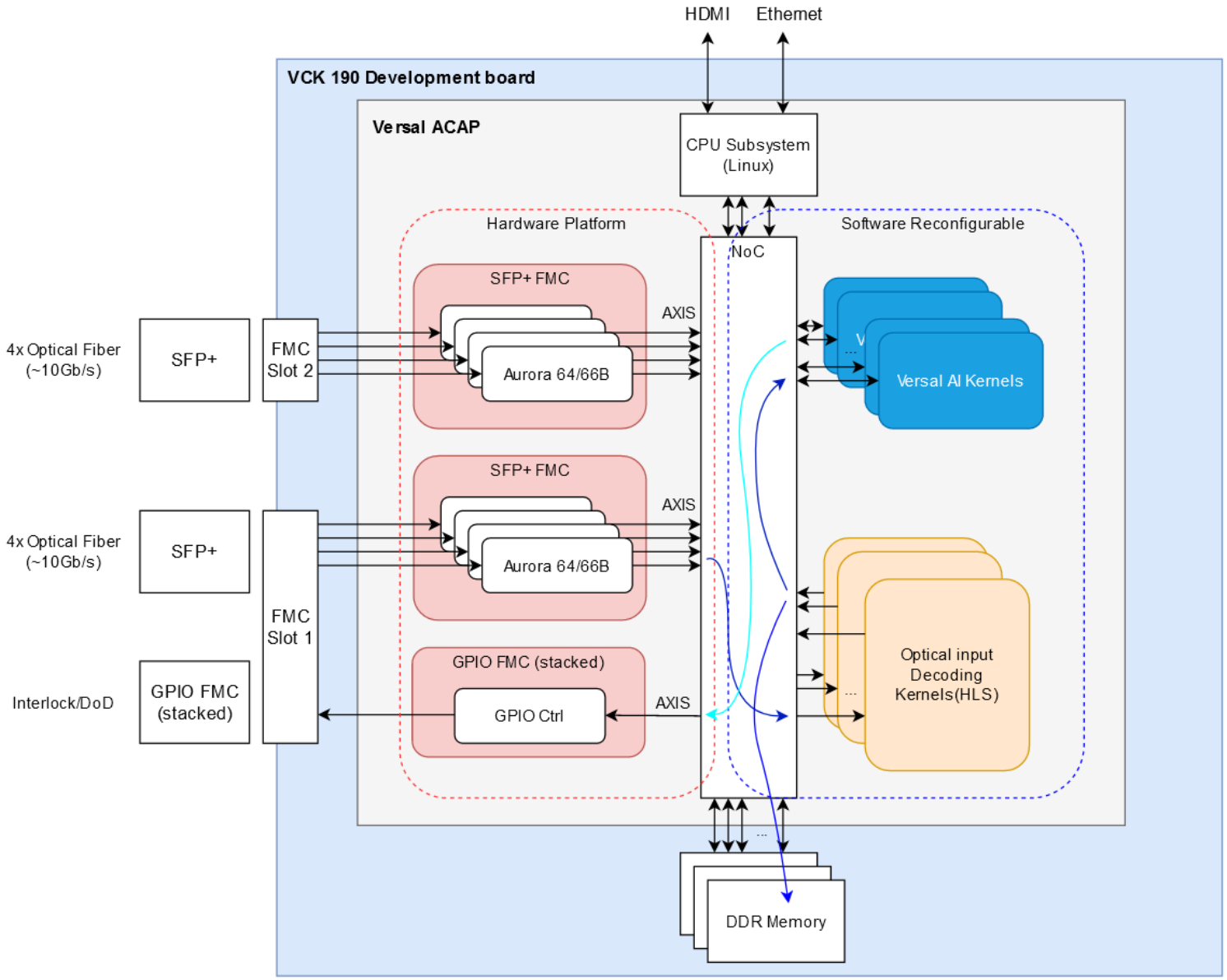
Summary of next steps

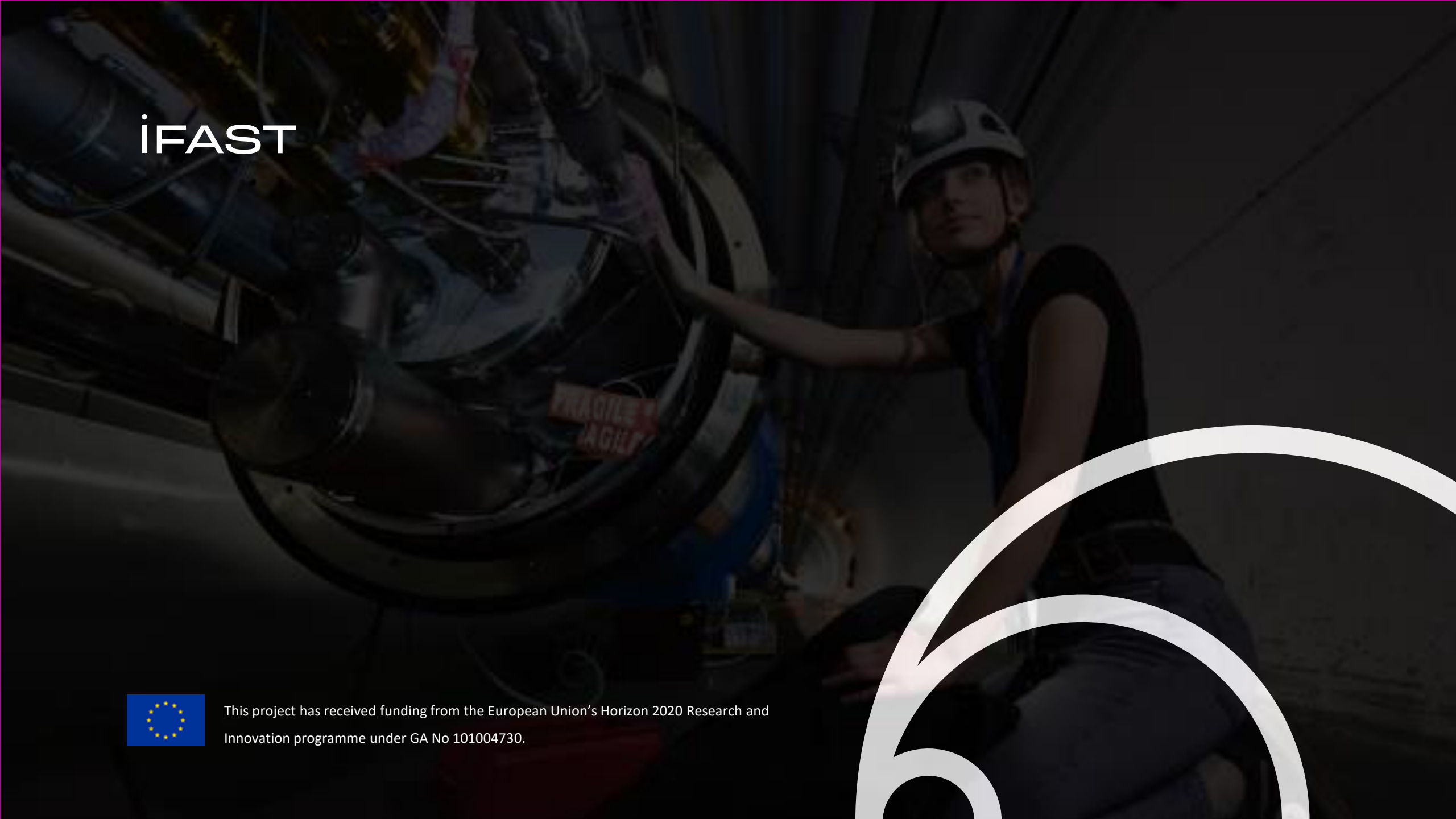
- **Investigate FPGA platform limitations.**
- Prepare simulated data to explore sensitivity to errant conditions.
- Explore compatibility with SLAC data standards.
- Acquire relevant data during commissioning phase with beam to ESS DTL1 (May – June 2022).
- Prepare for beam studies during commissioning phase with beam to ESS DTL4 (Q2 2023).

Extra

Platform

- Versal “Adaptable Compute Acceleration Platform, ACAP”
  - The NoC and Versal AI kernels in the diagram is made of dedicated hardware (not FPGA logic)
- The design can be divided into two parts (compare with tool diagram)
  - The “Hardware Platform” part is made by FPGA designer
  - The “Software Reconfigurable” part can be reconfigured in the SW compile flow with quick turnaround.
- Data can be streamed directly from input to the AI kernels for processing





This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under GA No 101004730.