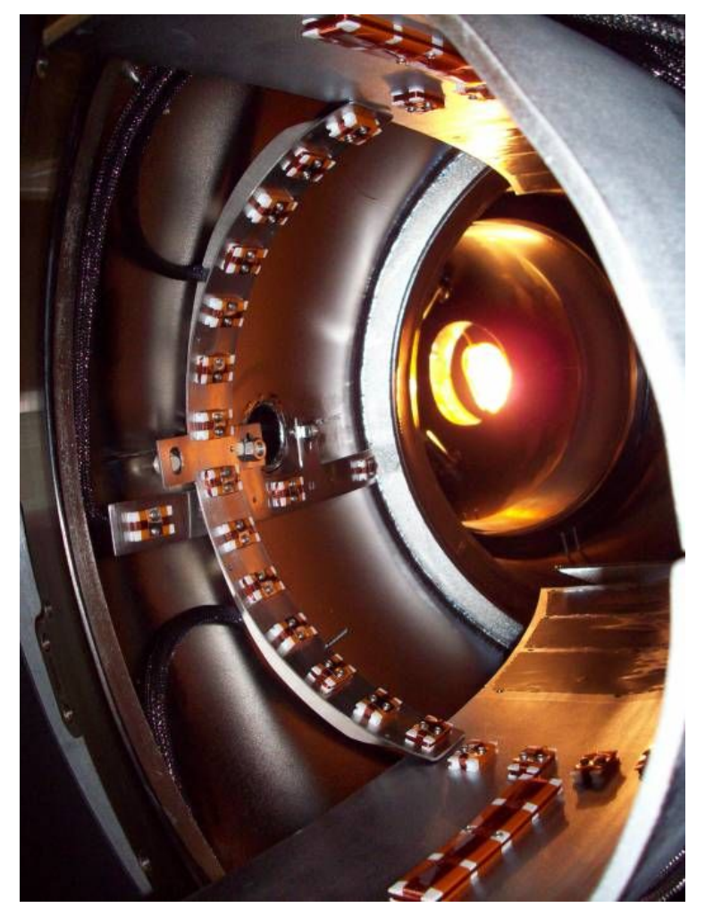


An open platform for in-situ high-speed computer vision with hls4ml

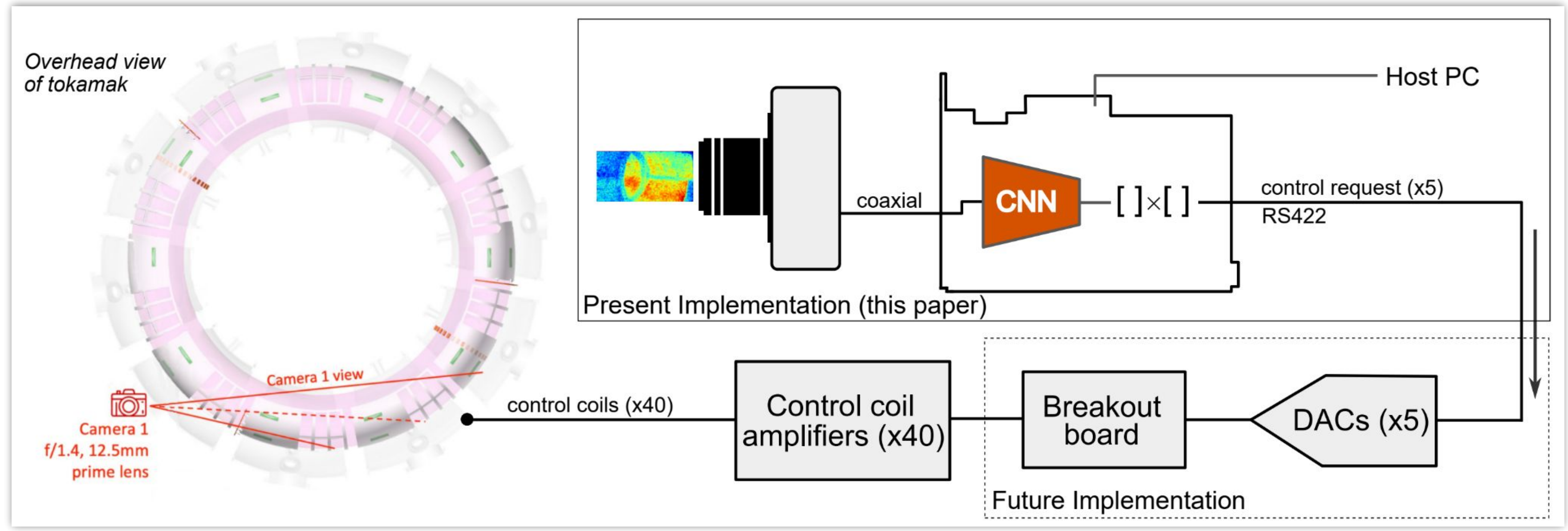
Fast Machine Learning for Science Conference 2024, Purdue University

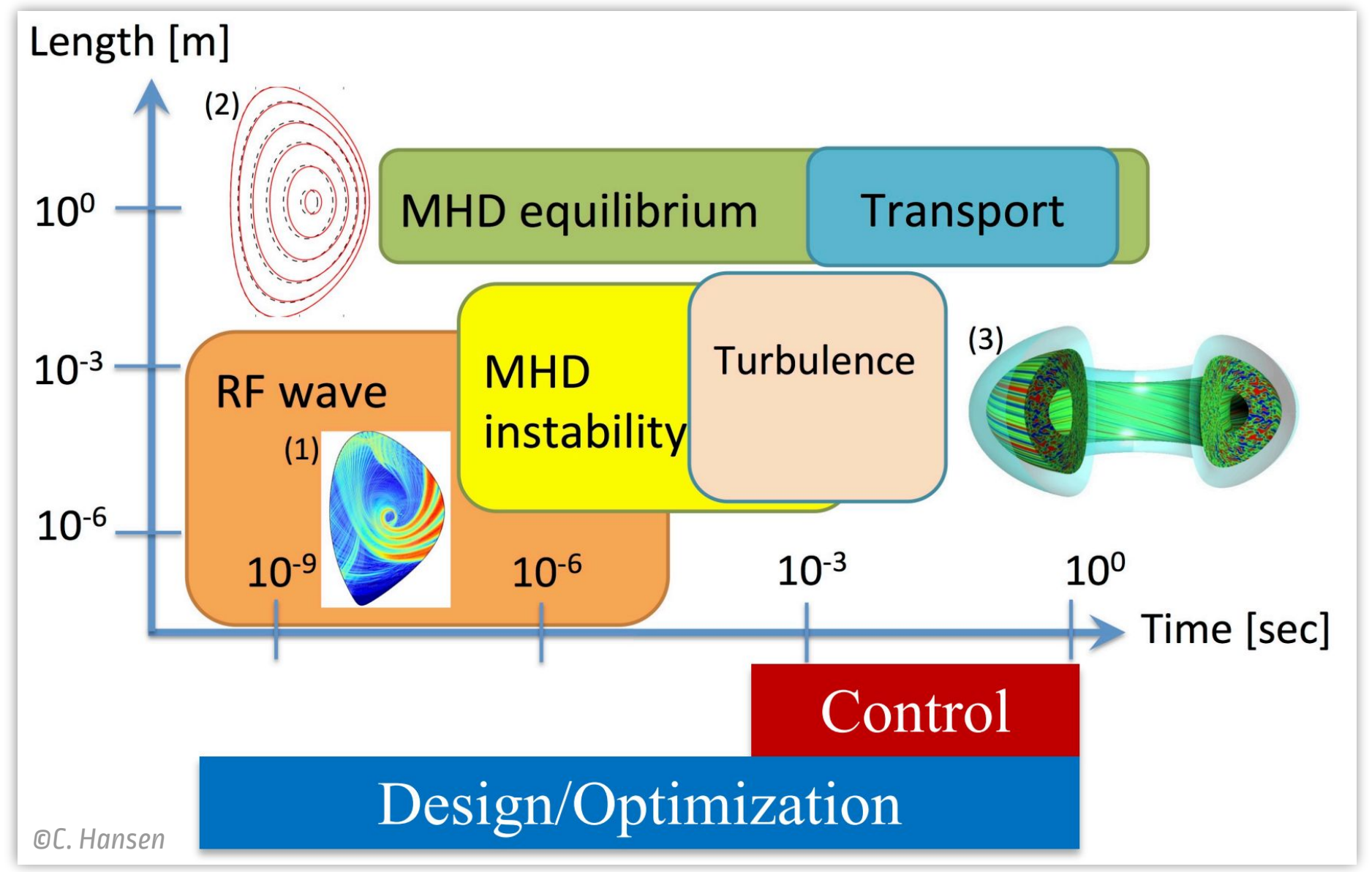
Ryan F. Forelli, Nhan Tran, Josh C. Agar, Yumou Wei, Chris Hansen, Jeffrey Levesque, Michael Cyros, Eric Janssen, Paulo Possa, Benoît Trémérie

Abstract

Low latency machine learning inference is vital for many high-speed imaging applications across various scientific domains. From analyzing fusion plasma to rapid cell-sorting, there is a need for in-situ fast inference in experiments operating in the kHz to MHz range. External PCIe accelerators are often unsuitable for these experiments due to the associated data transfer overhead, high inference latencies, and increased system complexity. Thus, we have developed a framework to streamline the process of deploying standard streaming hls4ml neural networks and integrating them into existing data readout paths and hardware in these applications. This will enable a wide range of high-speed intelligent imaging applications with off-the-shelf hardware. Typically, dedicated PCIe machine vision devices, so-called frame grabbers, are paired with high-speed cameras to handle high throughputs, and a protocol such as CoaXPress is used to transmit the raw camera data between the systems over fiber or copper. Many frame grabbers implement this protocol as well as additional pixel preprocessing stages on an FPGA device due to their flexibility and relatively low cost compared to ASICs. Some manufacturers, such as Euresys, have enabled easy access to their frame grabber FPGAs' firmware reference design. This reference design, aptly named CustomLogic, allows the user to implement custom image processing functions on the available portion of the FPGA. Moreover, open-source co-design workflows like hls4ml enable easy translation and deployment of neural networks to FPGA devices, and have demonstrated latencies on the order of nanoseconds to microseconds in domains ranging from particle physics to materials science. We provide the necessary wrappers, support files, and instruction to integrate an hls4ml model onto a frame grabber device with a few lines of code. We present two comprehensive tutorials in collaboration with Euresys to demonstrate the full quantization-aware training-to-deployment and benchmarking process, in addition to hls4ml's advanced feature set. We will also discuss and explore existing and potential applications. This work ultimately provides a convenient framework for performing in-situ inference on frame grabbers for high-speed imaging applications.

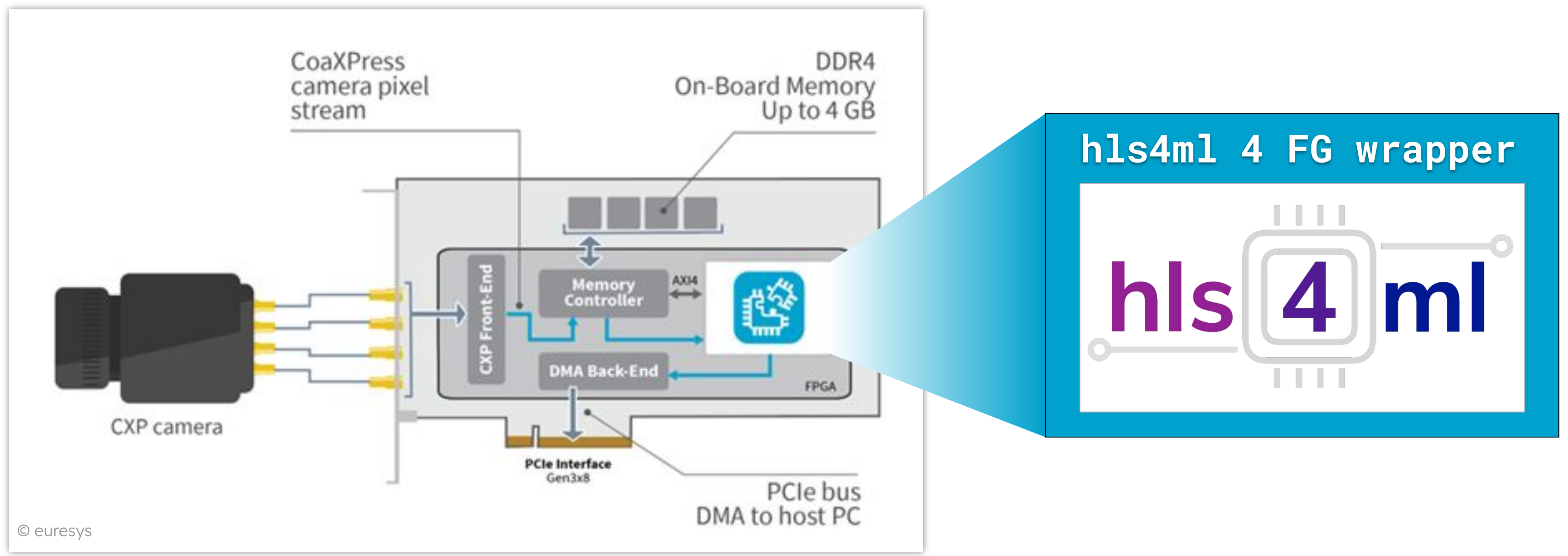
Background & Prerequisite Work





- **Countless applications in industry and science rely on high-speed imaging and machine vision analysis** for real-time feedback control, manufacturing quality control, compression/data reduction, etc.
- **Low latency machine learning for Tokamak instability control and suppression** (see arXiv:2312.00128)
  - Application timescale necessitates **<10us CNN inference latency** to effectively suppress magneto-hydrodynamic instabilities
  - Requires **in situ deployment** on frame grabber FPGA and optimization to meet latency constraints

Methods



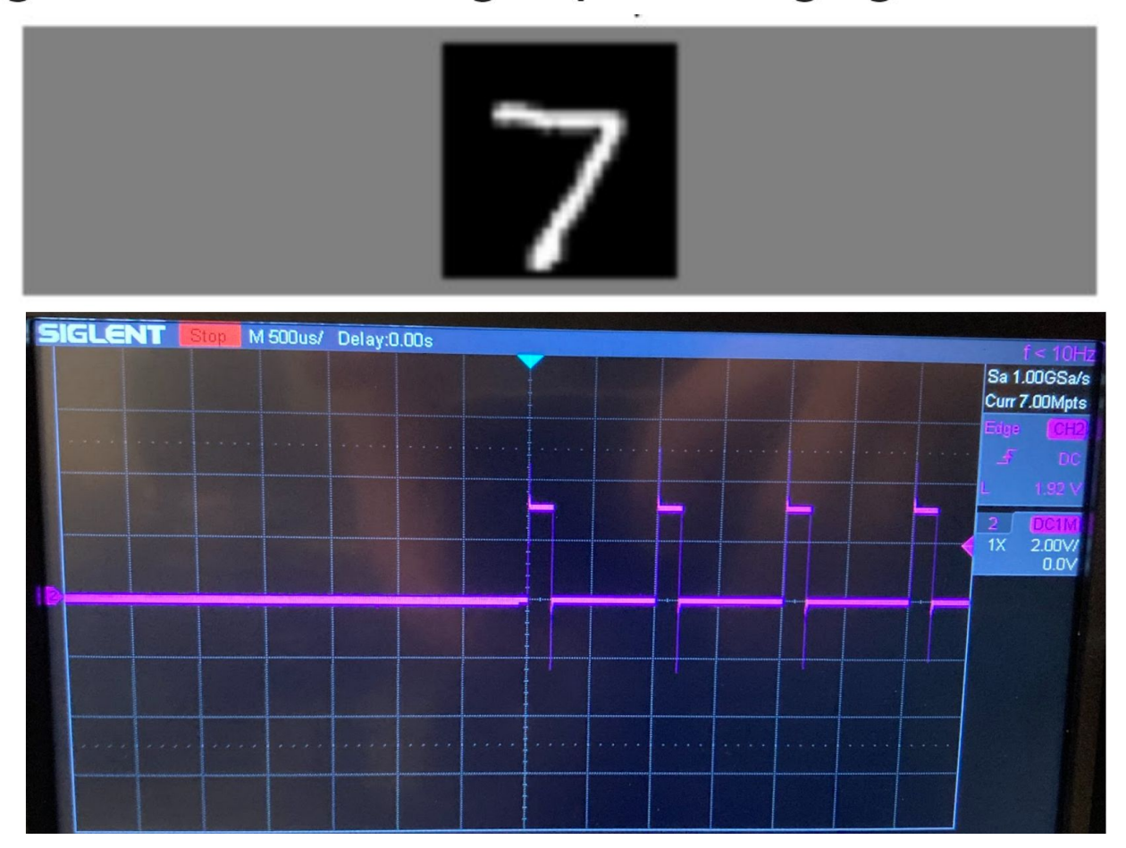
- **Enabled simple deployment of standard streaming hls4ml neural networks to off-the-shelf machine vision camera hardware** for high-speed imaging applications by leveraging open reference design from manufacturer and hls4ml.
- Developed HLS wrappers and necessary auxiliary files to:
  - ➢ automatically integrate hls4ml IP into frame grabber reference design
  - ➢ unpack and crop user-defined region of interest from CoaXPress camera stream
  - ➢ parse axi side channel for spatial information
  - ➢ reorder and duplicate camera stream
  - ➢ embed predictions with output
  - ➢ serially output result via TTL IO
  - ➢ enable easy hw benchmarking via frame grabber TTL IO
  - ➢ build Vivado project & synthesize design to meet 250MHz timing constraints.

Results

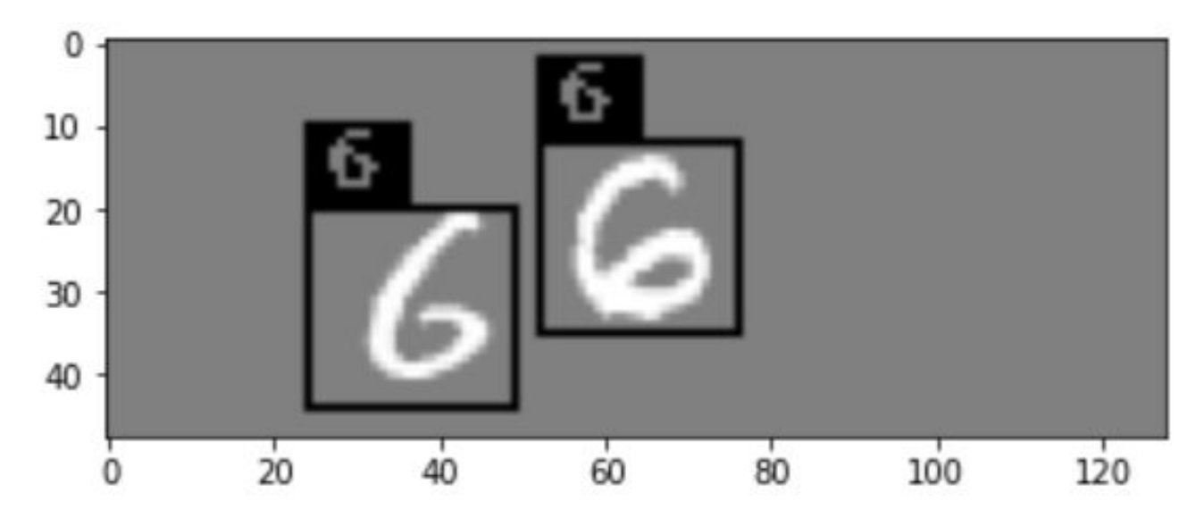
Two tutorials to get you started:

1. **Basic deployment and benchmarking** tutorial (Jupyter & Medium format) covers: *i)* SW/HW requirements, *ii)* pretrained quantized MNIST model compilation, *iii)* the basics of hls4ml and on-chip deployment
2. **Advanced YOLO hls4ml** tutorial covers: *i)* Fake YOLO (FOLO) model, *ii)* dataset generation, *iii)* quantization-aware training, *iv)* hls4ml hardware optimization API for pattern pruning, *v)* extensions API for custom data reduction layer, *vi)* post-training quantization, vii) profiling tool, *viii)* FIFO optimization, *iX)* frame grabber firmware integration, *X)* benchmarking

Tutorial on deploying neural networks for in-situ inference on frame grabber FPGAs in high-speed imaging.



Part 9: Frame Grabbers and Advanced hls4ml Features



FERMILAB–VISION IN THE LOOP

Implementing (cutting-)edge image processing to ensure image based control on ultra-fast processes.

FIGURE 1 - *Peeking inside the HBT-EP Fermilab Tokamak- ©……*

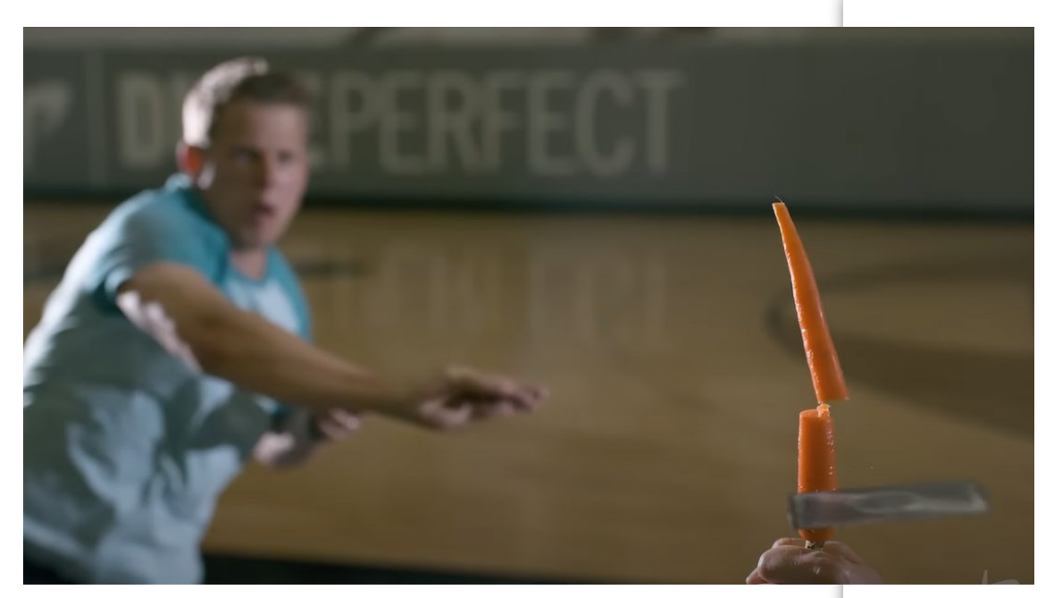
Euresys frame grabbers and their CustomLogic capability have enabled a team of researchers from Columbia University, Drexel University, Fermilab, and Lehigh University to develop new capabilities for fusion energy by greatly reducing the time between signal acquisition and control command in a camera-based high-speed tokamak plasma control loop.

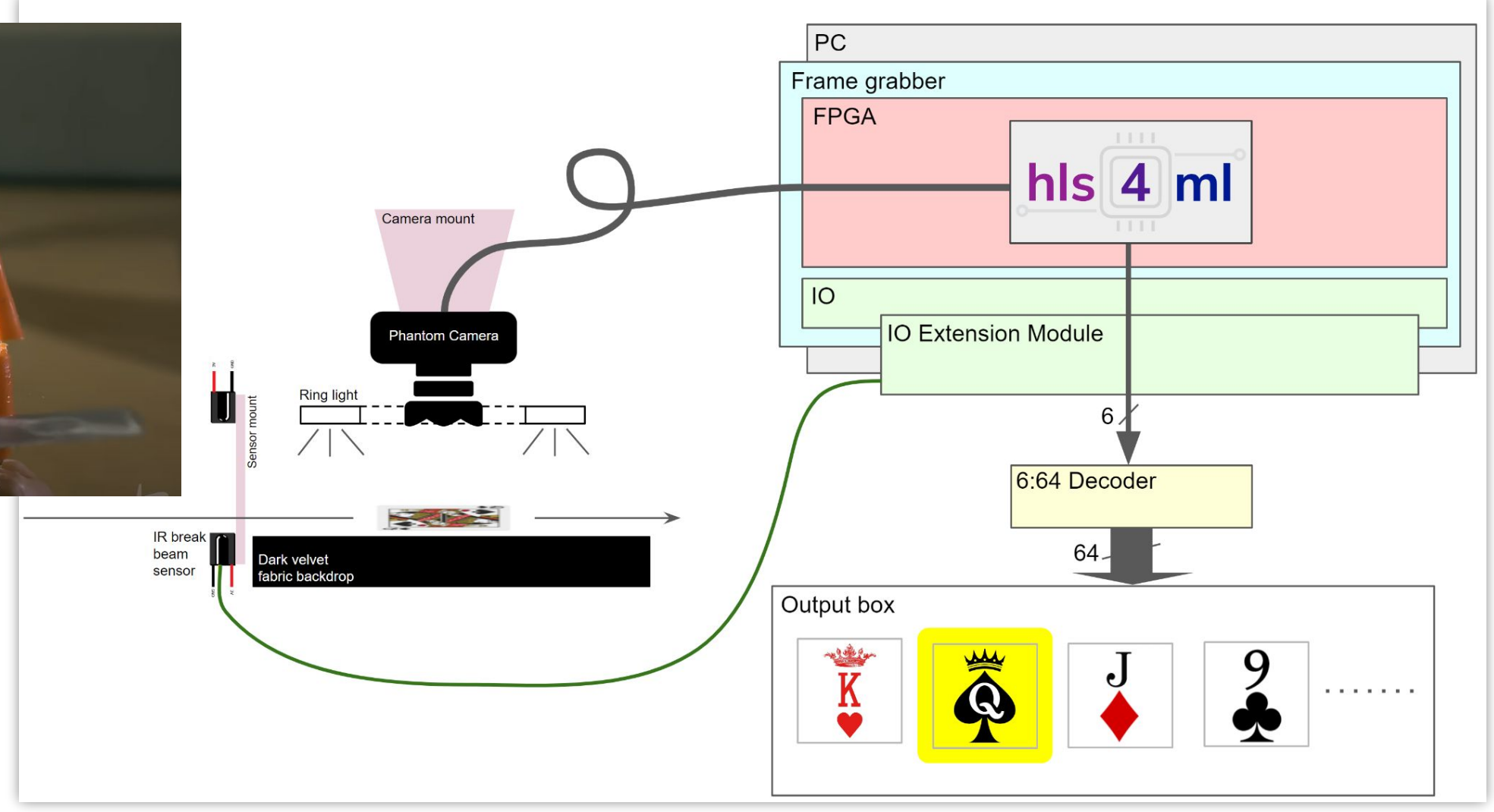
In another example of "Vision-in-the-Loop", plasma control requirements reach unprecedented speed for such complex control logic.

euresys Fermilab–Vision in the Loop - June 24 1

Future Work

- Add support for CoaXPress-over-fiber boards
- hls4ml card throwing demo
  - Classify cards thrown >90mph and display on hardwired output board





Northwestern University · Fermilab · euresys · LEHIGH UNIVERSITY · Drexel UNIVERSITY · COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK · U.S. DEPARTMENT OF ENERGY Office of Science

- Supported by US DOE Grant DE-FG02-86ER53222, DE-SC0022234, and US DOE ASCR under the "Real-time Data Reduction Codesign at the Extreme Edge for Science" Project (DE-FOA-0002501).
- RF gratefully acknowledges support from the Ryan Fellowship and the International Institute for Nanotechnology at Northwestern University.

Advised by: Professor Seda Ogrenci