Online Track-finding and Event Selection in Hardware at 40 MHz

David Monk

Northwestern University

This material is based upon work supported by the National Science Foundation under Grant No. 2111556

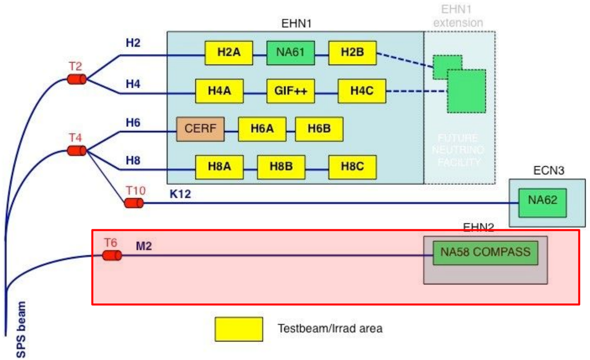
Outline

- Introduction
- MUonE DAQ System
    - FE Hardware
    - BE Hardware
    - Online monitoring
- Event Selection
    - Example signal - elastic scatter of muon and atomic electron
    - Occupancy
    - Track Finding
    - Extensions to the algorithm
- Conclusions

Introduction

For a detailed introduction to the MUonE experiment, please see the talk from [Riccardo Pilato]

- The MUonE experiment in collaboration with CMS has conducted several beam tests in the last two years to develop and test its DAQ system - including one as I speak
- Current system has been shown to perform continuous readout at 40 MHz with little truncation
  - Over 100TB of data recorded
  - Over 500 billion tracker hits
- As system scales in size, so much the approach to recording data that is relevant to the physics goals of the experiment

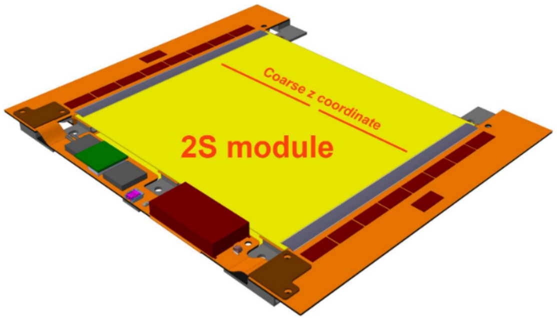


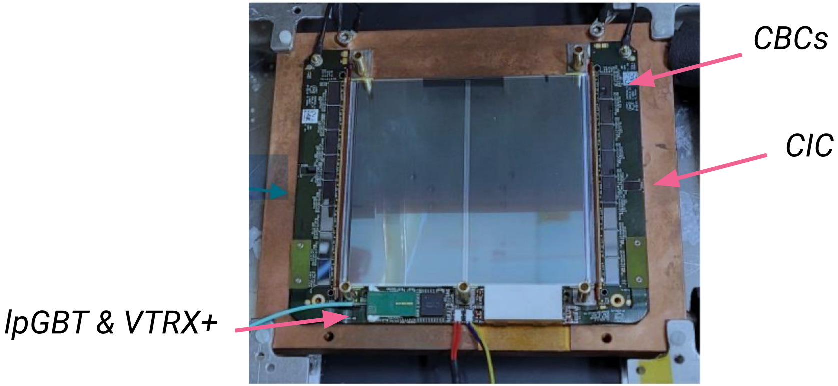
- CERN M2 beamline
  - Secondary beam generated from SPS (T6 target)
  - Up to $2x10^8$ muons per spill, 50 MHz asynchronous rate
  - 160 GeV muons or 40 GeV electrons (lower intensity)

MUonE DAQ System

MUonE DAQ System I - FE Hardware

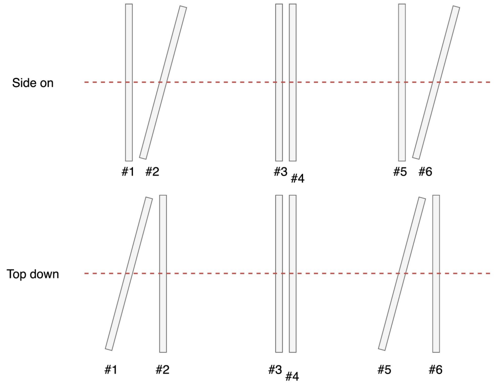
- 2S modules have been developed for the CMS Phase-II Tracker upgrade, composed of 2 layers of silicon strip sensors, whereby hits in the two layers are correlated to form a "stub"
- 10cm x 10cm active area, composed of 2 columns of 1016, 90 μm pitch, strips per layer
- Makes use of CERN-developed lpGBT+VTRx for optical readout at 5 Gbps
- Operates at "LHC" clock rate of 40 MHz
  - Asynchronous to M2 beam
  - Intended for CMS L1 trigger



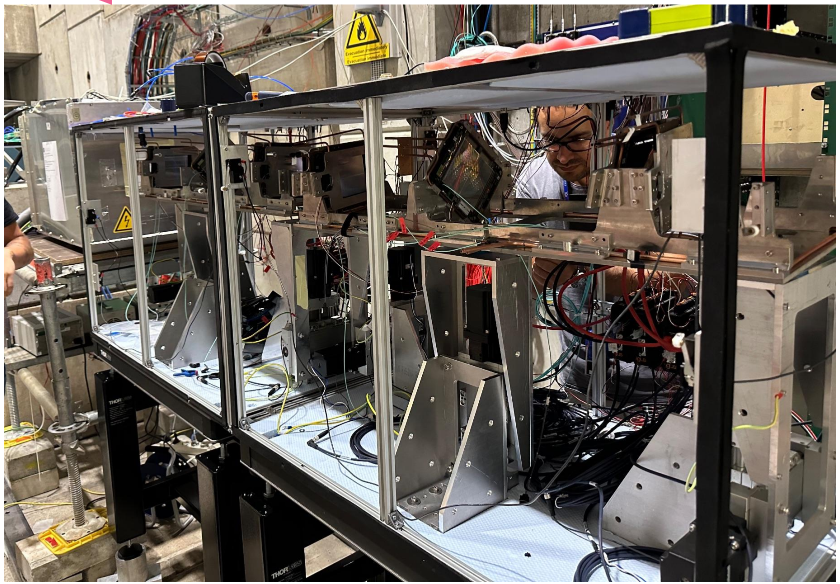


MUonE DAQ System II - Experimental Setup

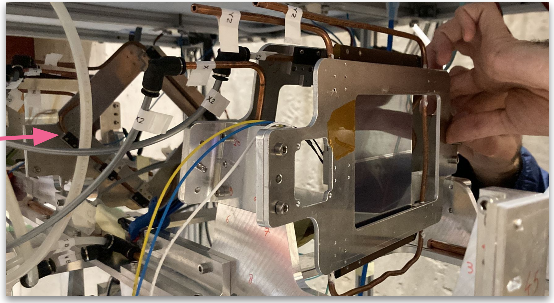
- Six modules placed within a “station”
  - Manages power, cooling, alignment and optics
  - Two stations installed on the beamline
- Modules arranged in pairs: x,y | u,v | x,y
- 2 cm carbon target placed in front of second station



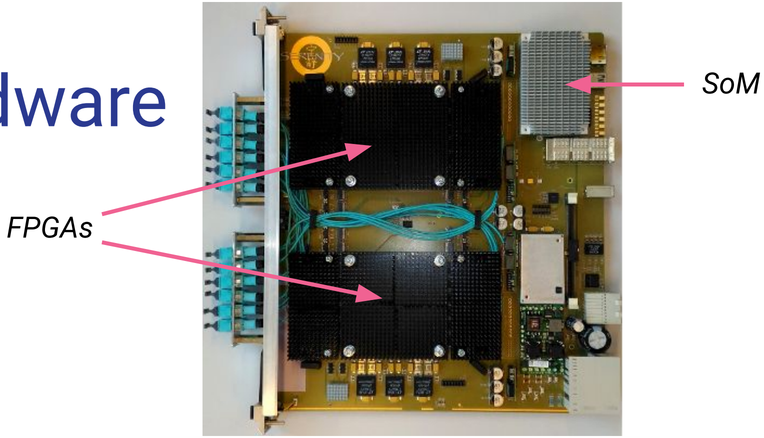
Beam direction



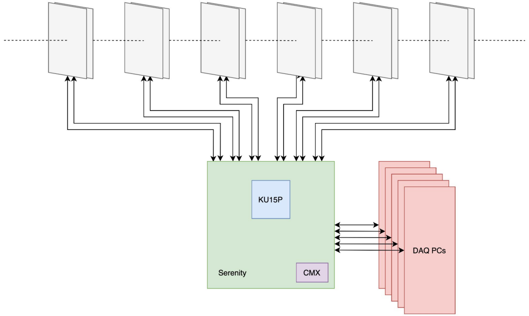
Housing for u,v modules

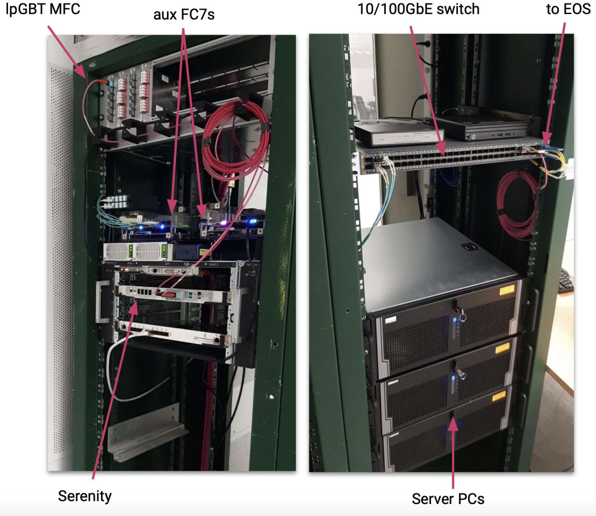


MUonE DAQ System II - BE Hardware



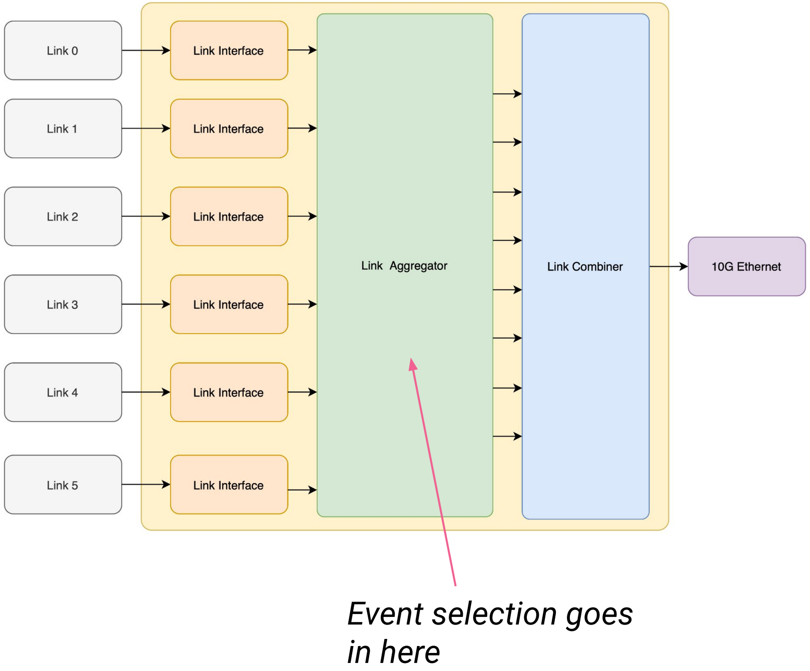
- Ingestion of data and configuration of modules is handled by the Serenity card
  - Prototype ATCA-class processing card developed for CMS Phase-II upgrade
  - Generic, composed of up to 2 AMD-Xilinx Ultrascale+ FPGAs and 144 optical transceivers for I/O
  - Also includes a System on Module (SoM) for management (Intel i5-based CoM-Express)
- Data transferred onward via 10 Gbps ethernet links to commercial PCs
  - PCs consolidate and chunk the data, before transfer to EOS for long-term storage and analysis
  - Direct link to EOS from experimental hall at 2 x 100 Gbps
  - No local buffering, data streamed live





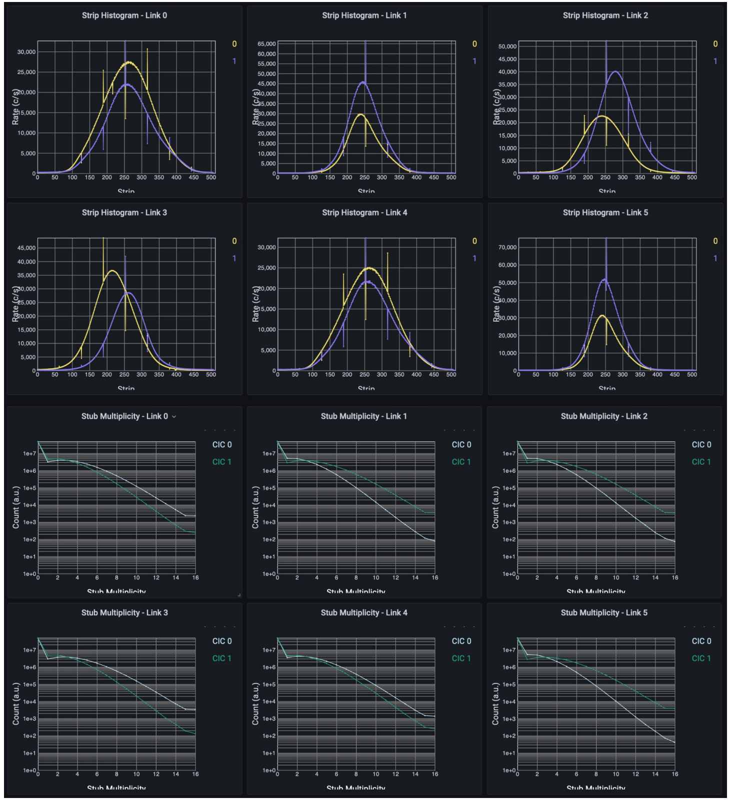
MUonE DAQ System III - BE Processing Firmware

- Makes use of EMP framework developed for CMS Phase-II upgrade
    - Abstracts infrastructure (links, clocks) away from algorithm
- Link interface firmware is common to CMS Phase-II tracker upgrade, rest of firmware custom to MUonE
- Stubs are collected by their clock ID across all modules, each collection sent sequentially to ethernet link
- Each station is handled independently and synchronously



MUonE DAQ System IV - Online Monitoring

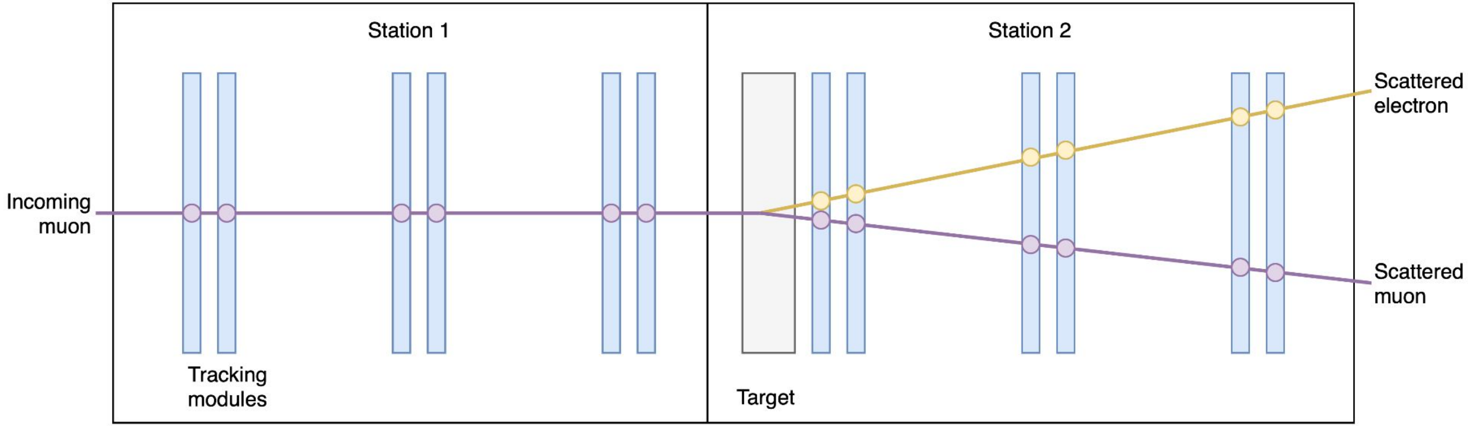
- Possibility to make use of both FPGA and SoM on BE processing card for monitoring of DAQ in real-time
- Two histogramming firmware blocks integrated into design
  - **Stub Address:** Provides real-time beam profile, generated from every stub sent from FE modules
  - **Packet size**: number of stubs histogrammed for every packet received. Useful for estimating truncation in FE modules
- Histograms are readout to the SoM via IPBus, then exposed as a web page to be scraped by Prometheus instance and plotted in Grafana
- Temperature, humidity sensors also connected as well as CAEN power supply



Event Selection

Example signal - elastic scatter of muon and atomic electron

- Physics motivation for MUonE is to measure angular distribution of elastic muon scatters against atomic electrons in a fixed target
- Signal is two tracks originating from a common vertex within the target, matched to single incoming track
- Tracks can be generated by combining multiple tracker hits; no magnetic field means tracks are straight lines
- PID of electron vs muon to be achieved with downstream ECAL (talk by Adrian Gutierrez)
- Primary backgrounds are non-interacting muons - one or more tracks without a common vertex close to station

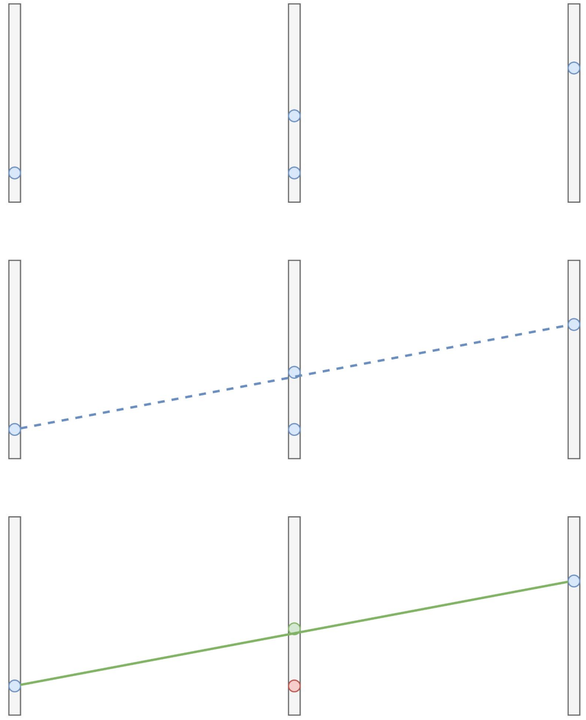


Occupancy cuts

- Most simple method for selecting candidate events is a cut on module occupancy
- For a two-track scattering event, each upstream module must have one hit and each downstream module must record at least two stubs in the same clock cycle
- Allow more than two stubs per module in downstream station to account for noise and other event topologies that may be of interest
- Cut in firmware trivial, per module occupancy available from buffer FIFOs in current DAQ system
- For data recorded in November 2022, occupancy cuts reduced the rate from 40 MHz to **5 MHz**

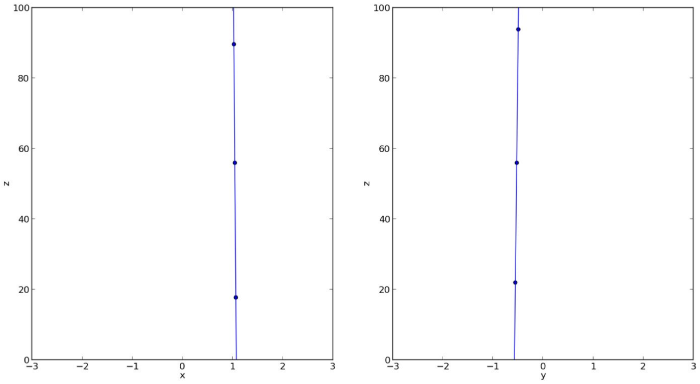
Track-finding I - Candidate Events

- Tracking in hardware a complex task, requires both resources and time
- Combinations of hits that could form a track increases exponentially, necessary to form candidate sets of stubs within event
- x and y axes can be considered independently for initial selection
    - Tracks can be formed from 3 hits: 1 hit at start of station, 1 hit at end, 1 “virtual” hit generated from combination of u,v planes
    - Candidate sets of hits created by propagating straight line made from outer hits to the u,v plane, then iteratively searched for compatible hits
    - Acceptance window can be programmatically tuned to maximise efficiency at a given occupancy
- Further 10% reduction in rate (**4.5 MHz**)



Track-finding II - Fitting

- Candidate sets of events sent to fitting stage
- Least Squares fit implemented using HLS
    - Tool capable of translating C++ code into VHDL, highly effective for rapid prototyping and complex operations (e.g. matrix inversion)
- Provides track parameters and associated errors
- Fitting performed independently in each axes, then 2D tracks are combined to form a 3D track
    - 2D tracks which share u,v hits are merged
- High latency at ~2us per candidate set, necessary to buffer event data for this time, intention is to have multiple fitters in the FPGA to pipeline stage.

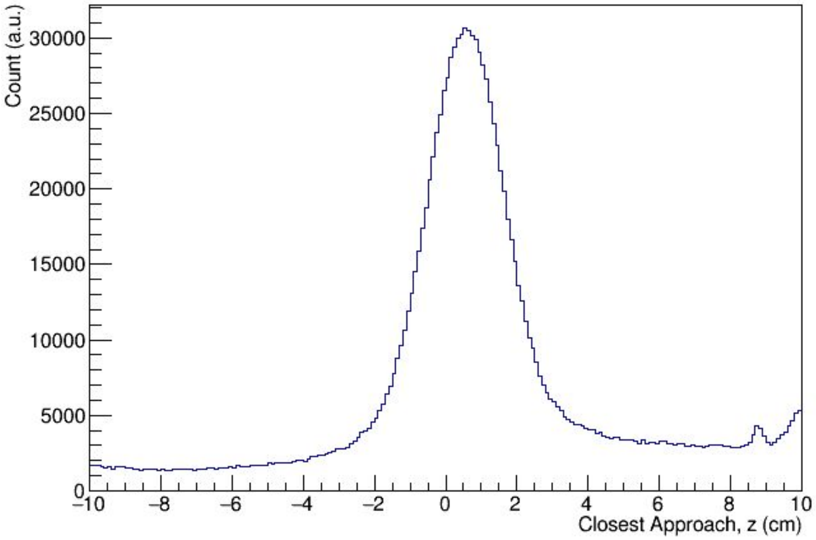



```
 readback X Covariance
-------------------------------
COVx[0] : 0x3201a5c4 (0.000000)
COVx[1] : 0xb4dcae0f (-0.000000)
COVx[2] : 0xb4dcae0d (-0.000000)
COVx[3] : 0x37f27176 (0.000029)

 readback Y Covariance
-------------------------------
COVy[0] : 0x3201c60a (0.000000)
COVy[1] : 0xb4e8594d (-0.000000)
COVy[2] : 0xb4e8594e (-0.000000)
COVy[3] : 0x3803506f (0.000031)

 readback x fit parameters
-------------------------------
Px[0] : 0xba0d3680 (-0.000539)
Px[1] : 0x3f895515 (1.072909)

 readback y fit parameters
-------------------------------
Py[0] : 0x3a58fda0 (0.000828)
Py[1] : 0xbf130e67 (-0.574439)
```

Vivado™ HLS

Extensions to the Algorithm

- Once tracking information is available online, further steps can be developed
- **Vertexing**: search for two tracks with intersection
  - Should offer ~6x reduction in data rate (**800 kHz**)
- **PID**: opportunity to use ML, in particular online
  - hls4ml project provides framework to translate trained networks into VHDL for use on an FPGA for inference

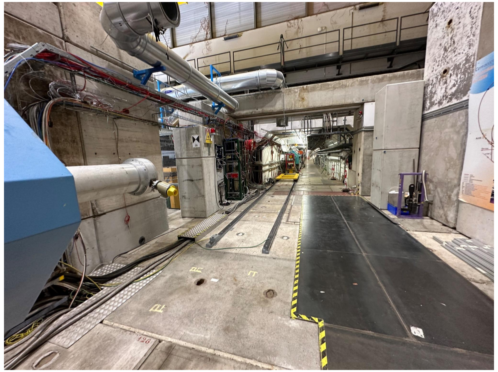


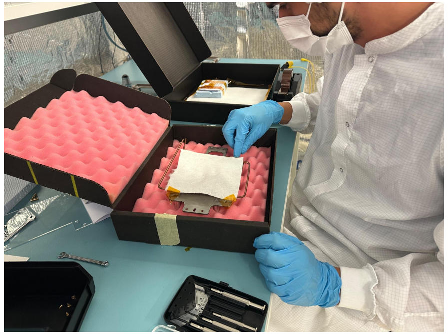
hls4ml

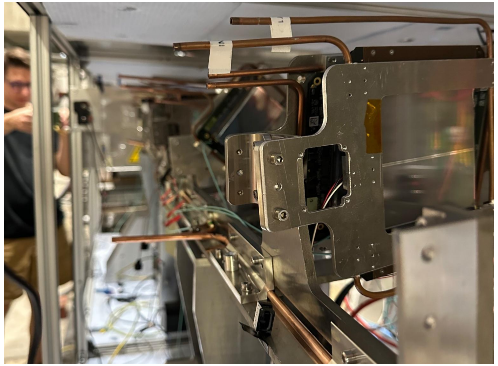
Current Tests

2023 Test Run

- Multi-week test beam, 12+ modules installed
    - First test beam at this scale for CMS prototype hardware
- Mainline DAQ system will only use occupancy cuts to manage readout bandwidth, will be sufficient with 4x10 Gbps ethernet links as output
- Track-finding will also be implemented on the FPGA, using data duplicated from the mainline DAQ, for comparison with offline reconstruction







Conclusion

- MUonE and CMS have sustained the 40 MHz readout of Phase-II Tracker modules in several joint test beams
  - Many TB of stub data live-streamed to EOS
- With higher beam intensity and larger scale detectors, readout bandwidth rapidly becomes constrained
- This challenge can be addressed through the use of modern FPGA technology, which provides the platform for real-time even selection based on complex topologies without external triggers
- The DAQ framework presented makes widespread use of common technologies, allowing for flexibility and use beyond the MUonE experiment

Backup

Applications to other Beamlines - UA9

- Many challenges and solutions presented are designed to be generic and can applied to other projects
  - Use of common hardware, firmware and software ensure that effort is shared amongst large collaboration across multiple experiments and larger commercial ventures (Docker, Kubernetes, Prometheus, Grafana)
- UA9 DAQ now extremely outdated, opportunity to update with modern hardware and software
- Once data is in the FPGA, many blocks can be reused