Accelerator and Detector Control: Summary

Thomas Britton (JLab) and Benjamin Nachman (LBNL)

Day 3 morning

Conveners: Benjamin Nachman, Thomas Britton (JLab)

10:00	**Accelerator and Detector Control: Introduction** **Speakers:** Benjamin Nachman, Thomas Britton (JLab)
10:05	**Anomaly detection/Online data quality monitoring** **Speaker:** Kishansingh Rajput (JLab)
10:27	**Online low-level calibration and operational conditions** **Speaker:** Torri Jeske (JLab)
10:49	**Online high-level calibration and analysis** **Speaker:** Mike Williams (MIT)
11:15	**break**
11:30	**Trigger rate control and allocation** **Speaker:** David Miller (University of Chicago)
11:52	**Machine Learning in trigger deployment** **Speaker:** Nhan Tran (Fermilab)
12:14	**Accelerator control** **Speaker:** Aurelee Edelen (SLAC)
12:36	**Discussion**

Live notes:

https://docs.google.com/document/d/18-IHRK43APHDOUhpXW74VRvu4tD4zBhqyqQ7VmT13z0/edit

The following slides are borrowed from the talks to show some highlights - please see the individual talks for details!

CDC Results

At false positive rate of 0.005
True positive rate for Anomaly is 0.96

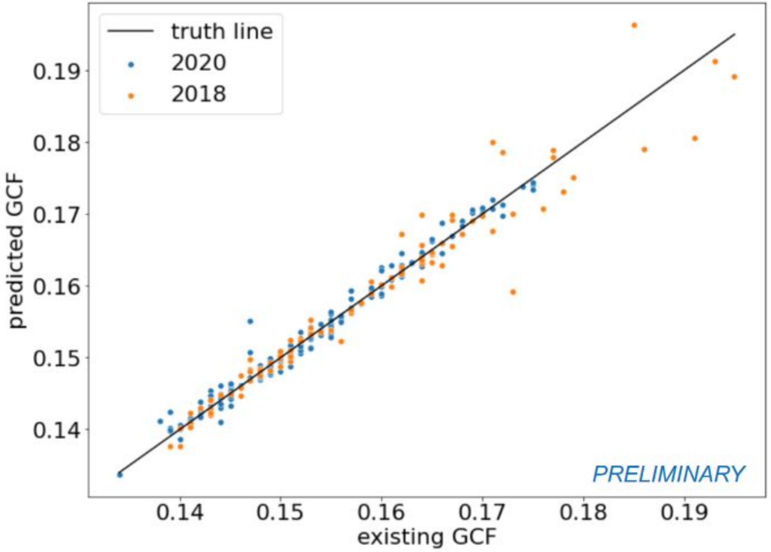
Hydra Fast Facts

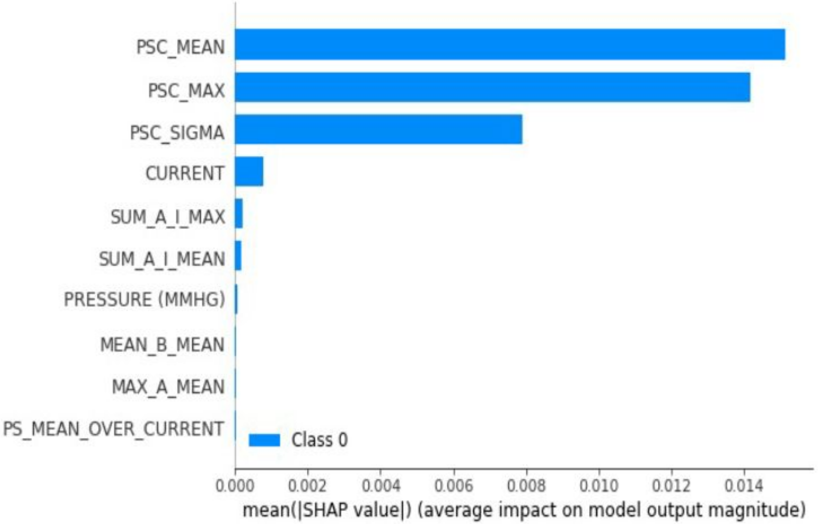
- Hydra looks at a finer time scale then any higher level monitoring the shift crew performs. **Approximately every minute**
  - Because who hits reset?
- **Operates** (conservatively) at about **3-4Hz**
  - From receiving an image to action **~300ms**. Most of the time spent on model inference
  - Inference accounts for **~71%** of the total processing time and is driven primarily by model size
- Currently focused on **go/no-go decisions**
  - Doctor classifying you as sick with no diagnosis as to what you are sick with. Refinement underway

HydraRun also saw the FDC problem, which I probably would have missed inspecting it by eye.

Koboldpress.com

Model predictions and feature evaluation





Recap + Ongoing work

Recap:

- Predict existing gain calibration constants with changing experimental conditions (2018 vs 2020 data)
- Established boundary for operating voltage of CDC based on previous run periods

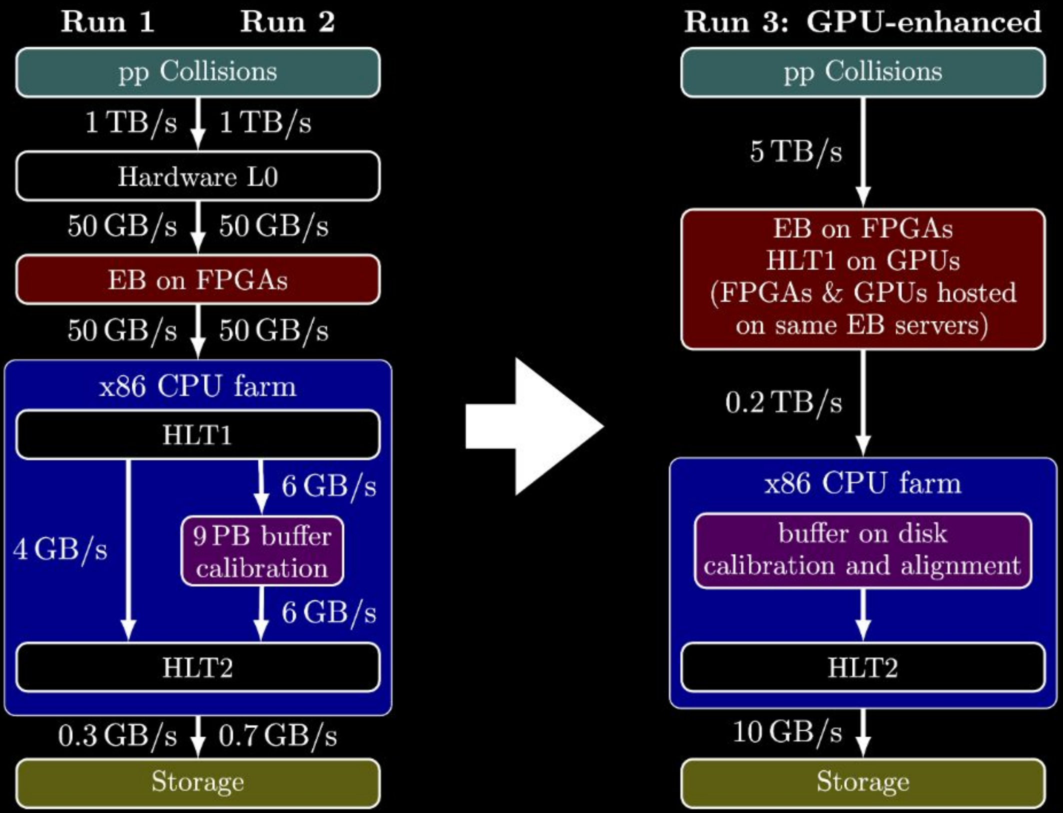
Ongoing work:

- Evaluating CDC resolution with predicted calibration constants
- Time to Distance model development in progress
- Incorporating physics information into model
- Application to CLAS12 Drift Chambers located in Hall B

Contact information:

– Torri Jeske | roark@jlab.org
– Diana McSpadden | dianam@jlab.org

Real-Time Analysis



GPU-enhanced option greatly increases our discovery potential in Run 3!

Summary

- LHCb successfully managed to calibrate — and fully reconstruct — all data in real time in Run 2.

- Since 2011, we have used ML-based selections in our primary trigger algorithms. Roughly 400 LHCb papers thus far are based on ML-selected data. These were based on a discretization method; a novel NN architecture has been developed for Run 3.

- Since 2015, fake tracks and clusters have been rejected in real time using NNs.

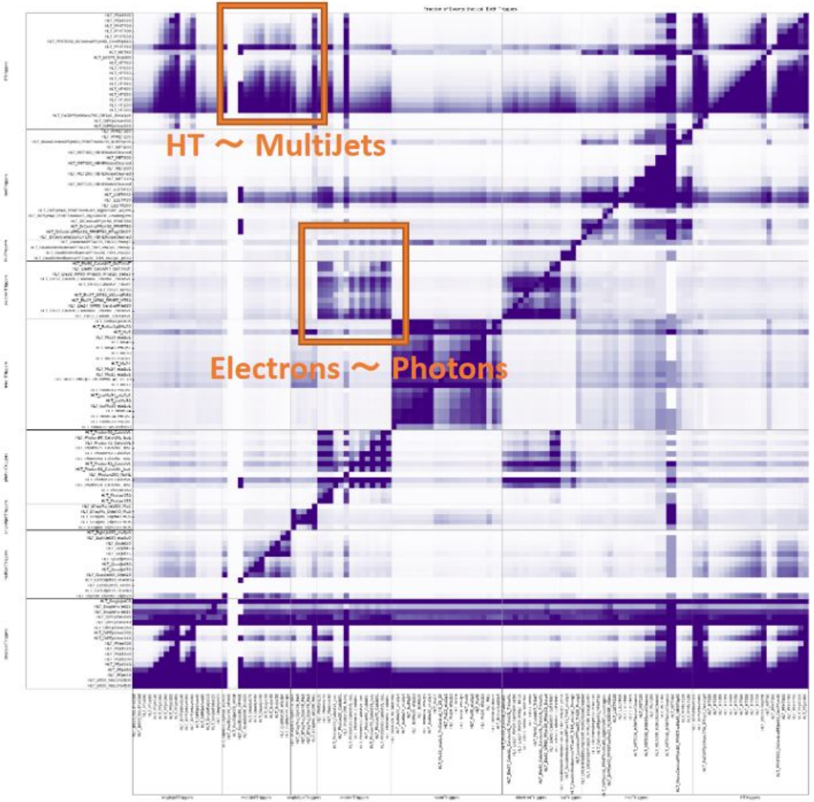
- Since 2016, NN-based particle ID selections have been used in the real-time selections. Several high-profile results published in Run 2 were only possible because of the performance increases provided by ML.

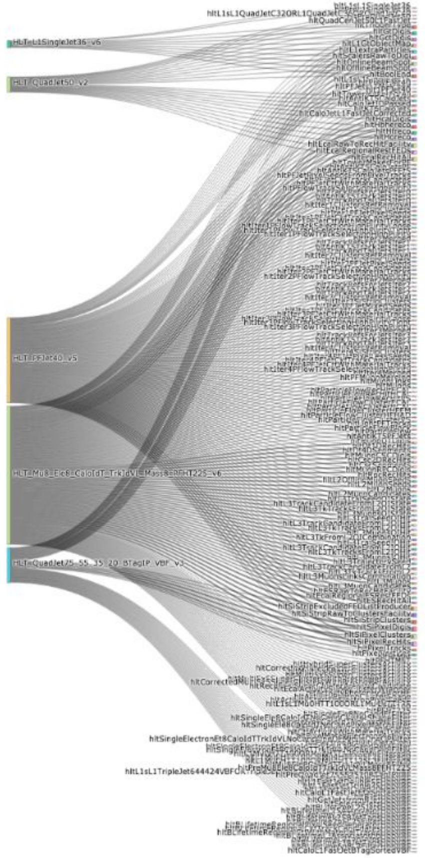
- In Run 3, we are removing our hardware trigger and will process every event in a GPU-based application that will track all particles with a very low pT threshold (and possibly do much more).

- Many other studies are underway to expand the use of real-time AI in Run 3 and beyond!

Modeling trigger “cost” using CMS open data

[NeurIPS 2020 WS] Self-Driving Trigger Paper





Summary and conclusions

- Our field is envisioning projects that span another 50 years, and so it is necessary that we allow ourselves to ask big questions!
- The concept of an autonomous data filtering and processing system for high-throughput physics facilities is well-aligned with physics goals
- We must ask what such a system could and should do for it to be useful, let alone feasible
- We have demonstrated some simple principles regarding interpretation of models and cost effectiveness using toy and open data
- Expect results soon demonstrating proof-of-principle using realistic dataset and cost models and functions

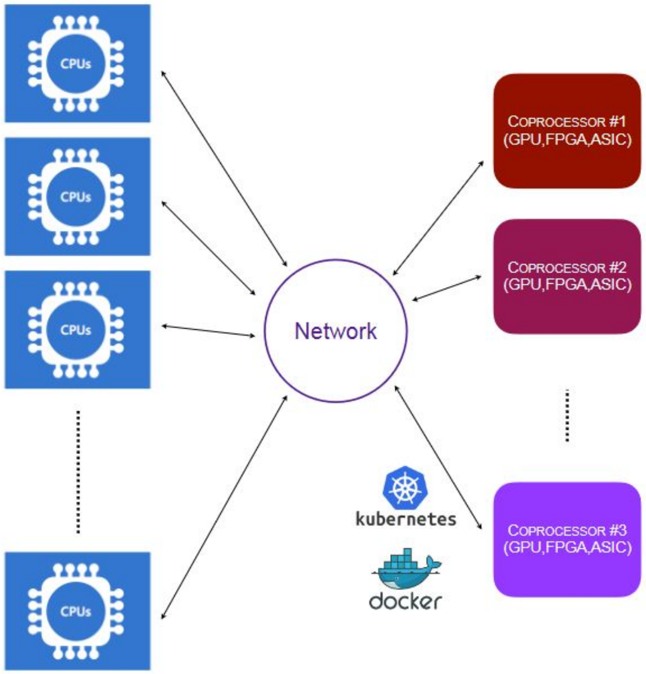
SONIC

Services for **O**ptimized **N**etwork **I**nference on **C**oprocessors

Flexible - optimize the hardware based on task; no need to support many ML frameworks in experiment software

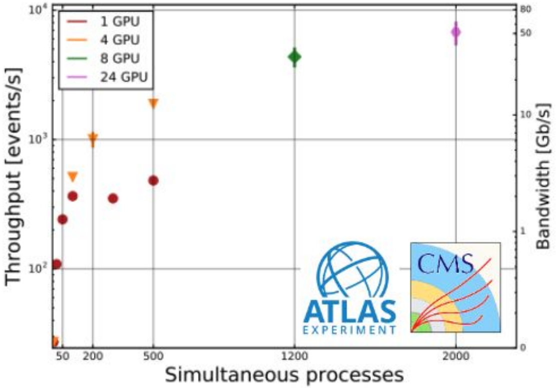
Adaptable - right-size the system to the task, you choose the number of coprocessors based on computing needs

Scalable - coprocessor need not be co-located next to existing CPU infrastructure; common software framework



$N_{CPU} \;!\!= N_{coprocessor}$

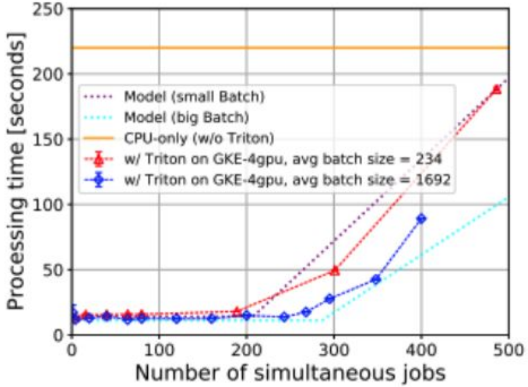
Results



Demonstrated significant and efficient acceleration of LHC/ProtoDUNE tasks

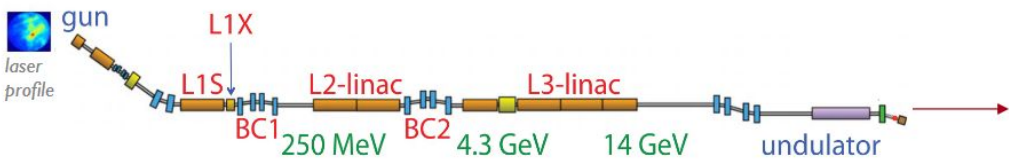
Broad range of tasks — cluster calibration, jet tagging, cosmics 1D, Graph NNs

	Wall time (s) ML module	non-ML modules	Total
CPU only	220	110	330
CPU + GPUaaS	13	110	123



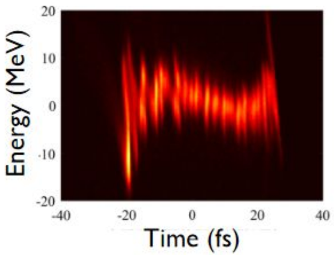
Deployed on-premises, in the cloud, and at HPC - exploring all types of new hardware (FPGA, GPU, TPU, ...)

References:
arXiv:1904.08986
arXiv:2007.10359
arXiv:2009.04509
arXiv:2010.08556

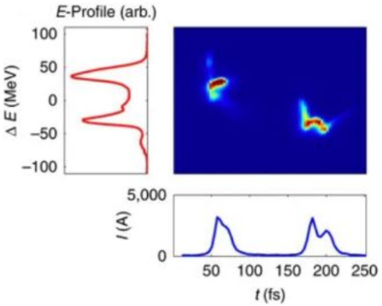


photon beam to 7 experiment stations

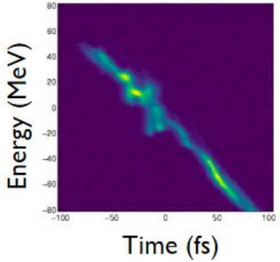
J. Qiang et al, PRAB (2017)



A. Marinelli, et al., Nat. Commun. 6, 6369 (2015)



A. Marinelli, IPAC'18



Approximate Annual Budget: $145 million

Approximate hours of experiment delivery per year: 5000

About $30k per experiment hour to run!

400 hours hand-tuning in a year → $12 million value
~10 additional experiments

ML Future Directions / Needs for Accelerator R&D

- Uncertainty quantification
  - Detect when model may not be accurate (e.g. outside training range)
  - Leverage for safe exploration of parameter space
- Active learning
  - Retraining to account for drift or adapt during search
  - Sampling strategies to efficiently explore large parameter space + generate training data (maximize information with the least samples)
- Efficient ways to handle high dimensional data:
  - Images, 6D phase space
  - More variables (full accelerator vs. small test cases)
- Physics-informed / constrained ML
  - Improve robustness / generalization to unseen regions of parameter space
  - Reduce need for additional data
  - Extract physics from measured data
- Differentiable Simulators
  - Wide range of types of simulation codes for accelerators (analytic matrix transport codes, particle-in-cell) → relatively unexplored area
- Interpretability
  - Important for ML-based tuning, identifying physic underpinning a prediction
- *Many shared challenges with other SciML domains → accelerators are unique test beds for these kinds of problems*

Summary

There are many interconnected topics related to real time calibration, analysis, and control for both detectors and accelerators.

The talks in our session covered specific examples, but also many of the speakers provided a broader context.

There are clear paths for synergy between EIC, LHC, and other particle/nuclear experiments and other accelerator complexes. We look forward to future discussions, innovation, and progress!