Fast Tracking on Heterogeneous Hardware for the ATLAS Event Filter

Abraham Kahn, on behalf of the ATLAS TDAQ Collaboration

Overview

The High Luminosity upgrade of the Large Hadron Collider (HL-LHC) will deliver $\mathcal{O}(10)$ times the total integrated luminosity of LHC Runs 1-3 combined [1]. However, this increased rate of proton-proton collisions per bunch crossing (pile-up) poses significant challenges to the ATLAS Trigger and Data Acquisition System (TDAQ).

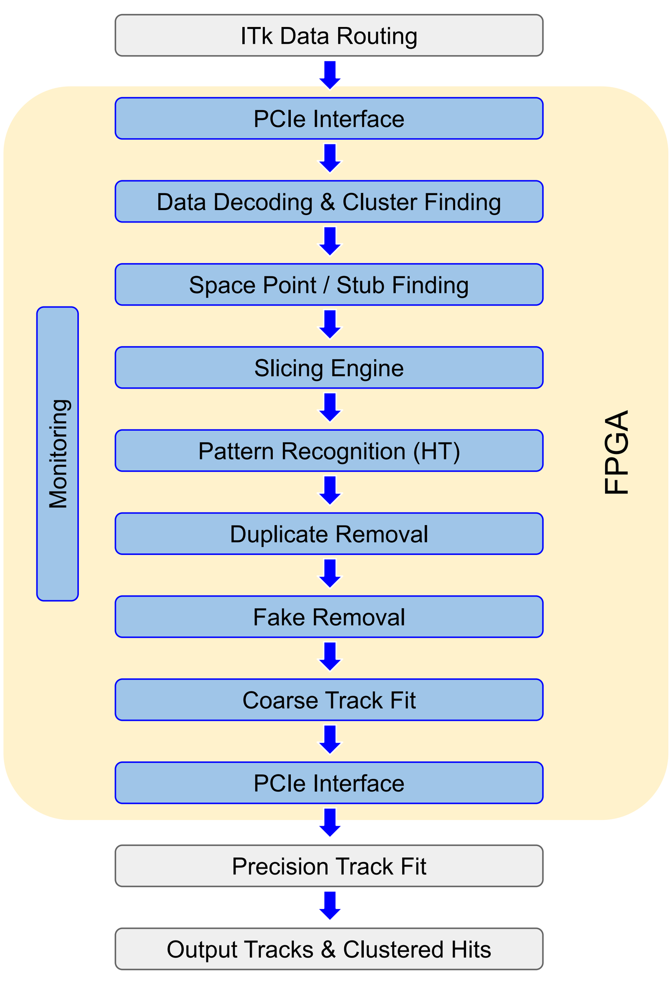
In order to meet the physics goals of the HL-LHC, an upgrade of the TDAQ system is underway, and a completely new Inner Tracker (ITk) will be installed. The Event Filter (EF) must be able to read in data from the ITk and quickly execute track reconstruction algorithms in this high pile-up environment.

Heterogeneous computing systems (systems which integrate multiple types of computational units: GPUs, FPGAs, ASICS, etc.) prove to be a compelling option for a new EF design, with the potential for higher performance, lower power consumption, and lower latency than CPU-only systems.

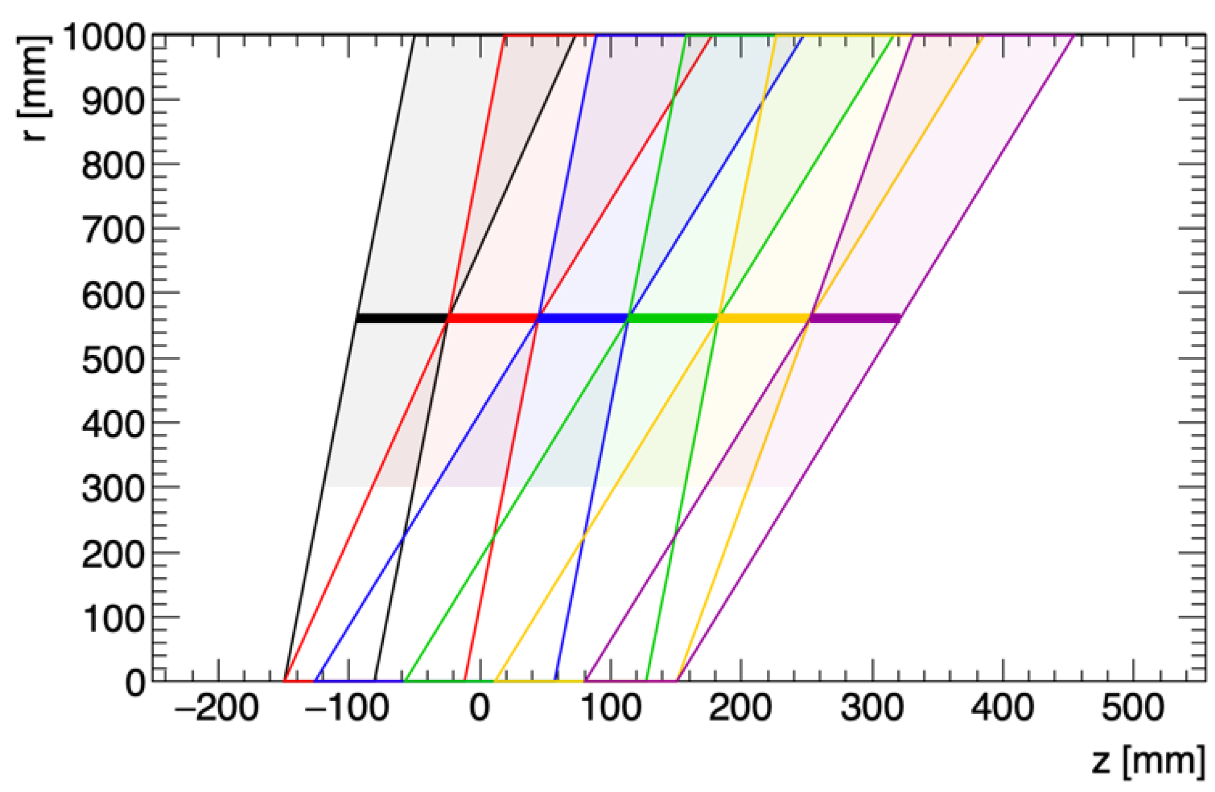
FPGA-based Tracking

- One such heterogeneous computing system would use lightweight CPU software to load an event onto an available FPGA, which then performs the track reconstruction and outputs a set of track candidates.
- Each FPGA would have the following implemented in firmware:

  - **Data Decoding and Clustering:** Provides cluster position and size information from front-end ITk data.
  - **Space Point/Stub-Finding:** Pre-filters hit clusters ("stub-finding") and combines strip hit clusters on opposite sides of a stave ("space points") to place 3-D constraints on a track location and reduce input to the pattern recognition stage.
  - **Slicing Engine:** To combat large occupancy, data packaged in $\eta - \phi$ regions is divided in the z-direction into "slices". (see below)
  - **Pattern Recognition:** The most resource intensive step - identifying track candidates from hit clusters. Two versions of the Hough Transform algorithm have been explored.
  - **Duplicate and Fake Removal:** Reject fake tracks using a Neural Network (NN), whose architecture was developed with an FPGA firmware implementation in mind.
  - **Coarse Track Fit:** The track candidates from the pattern recognition stage are further cut down using a linearized χ^2 test, and their track parameters are evaluated.



- An important strategy for reducing the occupancy in the accumulator is to divide each $\eta - \phi$ region into "slices" in the z-direction, referred to as z-slicing.



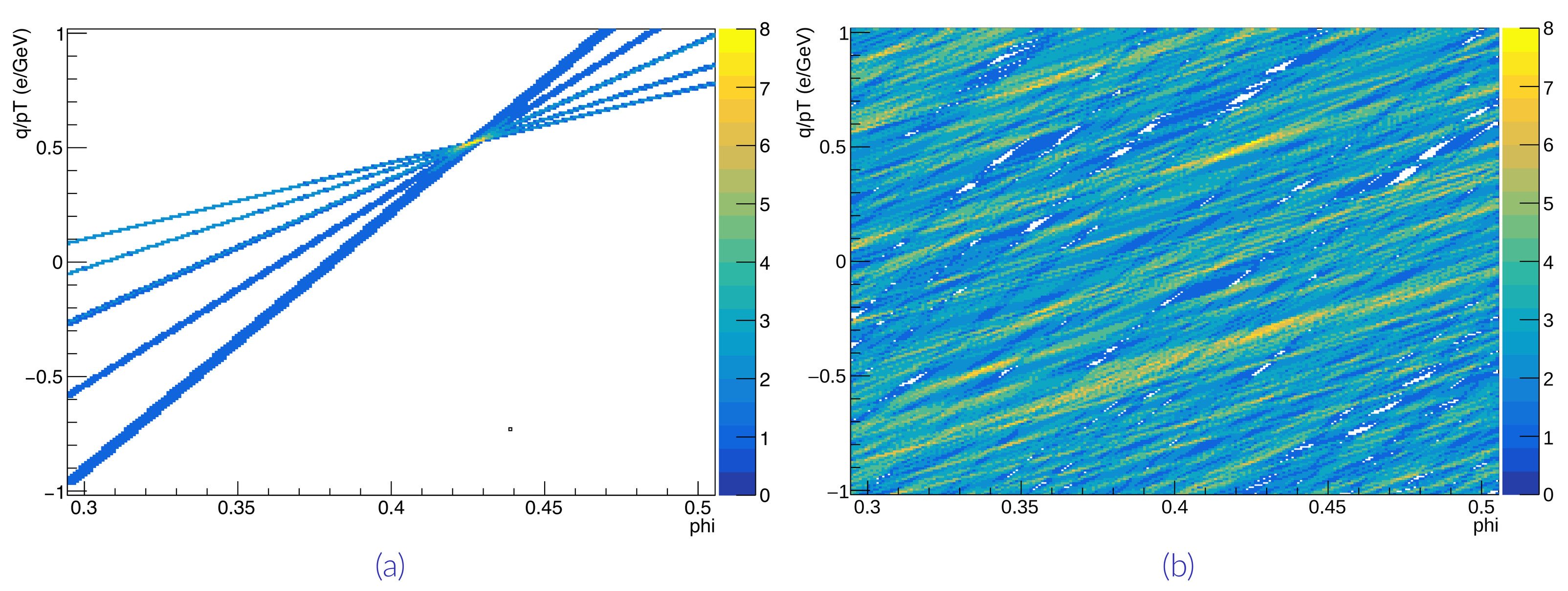
An Illustration of the z-slicing strategy. Each colored region represents a slice.

2D Hough Transform for Pattern Recognition

- the Hough Transform converts each hit on the $x - y$ plane into a line on the $qA/p_T - \phi_t$ plane, which is the set of possible track p_T and track ϕ consistent with that hit position.

$$\frac{qA}{p_T} = \frac{\sin(\phi_t - \phi_h)}{r_h} \approx \frac{\phi_t - \phi_h}{r_h}$$

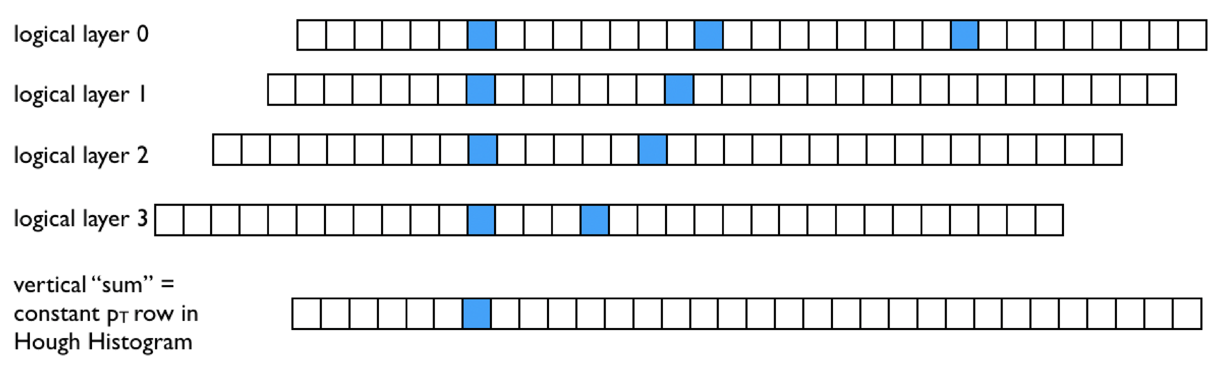
- Where q is the charge of the particle, A is the curvature constant for a 2 T Magnetic field, p_T is the transverse momentum of the particle, and ϕ_t is the azimuthal angle of the particle at the origin.
- If five lines intersect the same $\frac{qA}{p_T} - \phi_t$ bin, it suggests that those five hits came from the same charged particle track, with p_T and ϕ_t equal to the intersection point.



Hough transform accumulator for a single muon track without (a) and with (b) pile-up

1D Hough Transform for Pattern Recognition

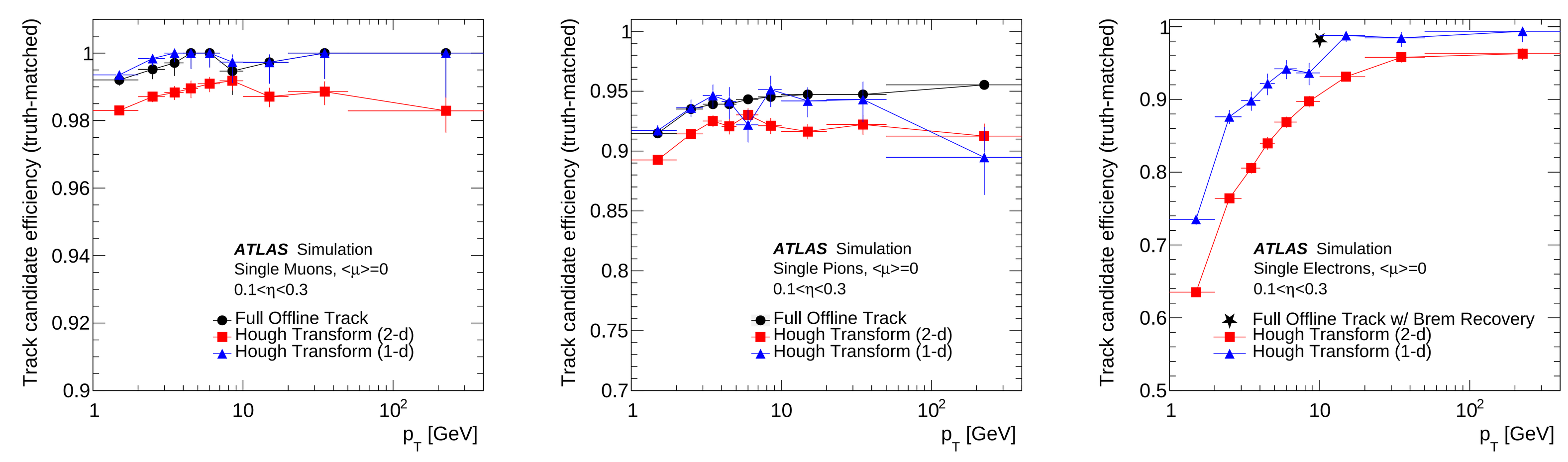
- An alternative method to track finding, the "1D Hough Transform" approximates that all hits in a layer have the same radius, so that the entire set of hits can be represented as one bit string per layer with each bit indicating whether a ϕ bin contains a hit.
- It follows that different p_T hypotheses simply correspond to shifting the bit strings relative to each other, and a track candidate is found if a ϕ bin has a hit in enough layers.



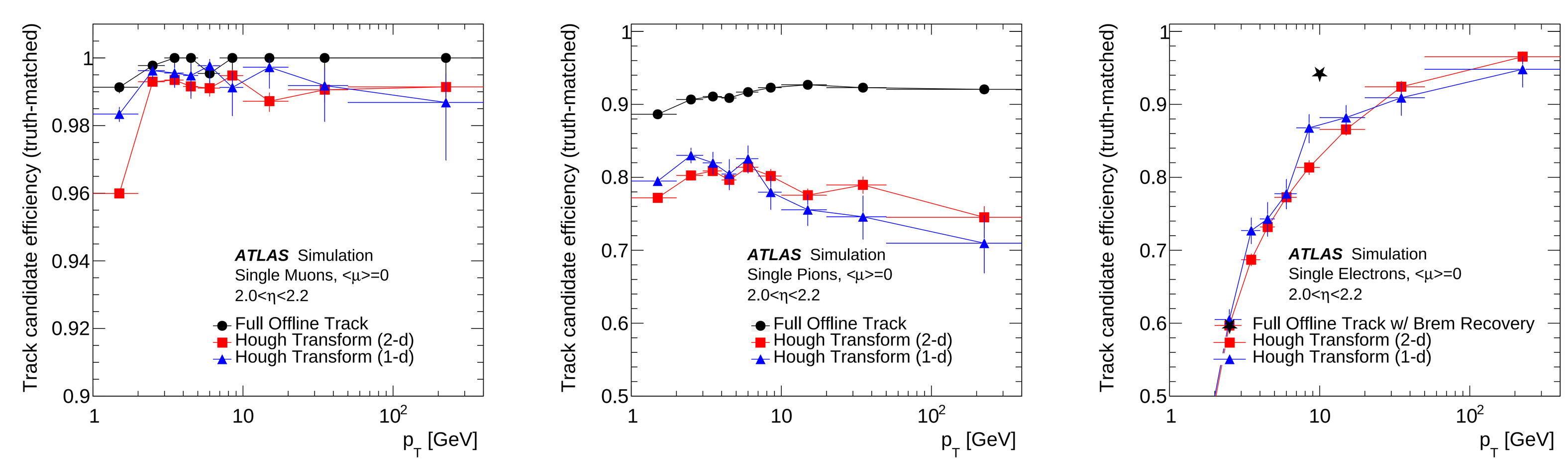
1D Hough approach. After applying a p_T shift perform a vertical sum, if it's above a given threshold a track candidate is found.

Pattern Recognition Performance

- The truth-matched track candidate efficiency was calculated as a function of p_T for muons, pions, and electrons in two regions of η [2].
- The performance of each Hough implementation is comparable to the Offline fast tracking software in the central part of the ATLAS detector.
- In the less studied forward region of the detector, the performance is not as good as Offline. This may be addressed with further optimization of each Hough transform configuration, or potentially by using an alternative pattern recognition algorithm.



Muons in $0.1 < \eta < 0.3$ — Pions in $0.1 < \eta < 0.3$ — Electrons in $0.1 < \eta < 0.3$



Muons in $2.0 < \eta < 2.2$ — Pions in $2.0 < \eta < 2.2$ — Electrons in $2.0 < \eta < 2.2$

Resource Usage

- The estimated size and total power of an FPGA-based heterogeneous system is shown in the table below [2].

	Run 4	Run 5
# of Accelerator Cards	270-680	
CPU resource requirement [MHS06]	1.1-1.7	1.6-2.4
Accelerator Power [MW]	0.08-0.18	
CPU Power [MW]	0.28-0.42	0.32-0.48
Total Power [MW]	0.4-0.6	0.4-0.7

- This can be compared to an estimate of the required resources for a CPU-only fast track reconstruction [2]:

	Pile-up 140		Pile-up 200	
	full-scan	regional	full-scan	regional
Rate [MHz]	0.15	1.0	0.15	1.0
CPU Resource requirement [MHS06]	3.41	2.49	5.36	3.81
Tracking resource requirement [MHS06]	5.90		9.17	
Tracking power requirement [MW]	1.47		1.83	

Conclusion

A heterogeneous commodity system has been demonstrated to be a viable option for EF tracking in ATLAS. This has been shown with the Hough Transform, but a broad range of pattern recognition strategies are still being explored, including using Graph Neural Networks and GPUs.

An FPGA-based heterogeneous system also offers notable power savings in comparison to CPU-only systems, while still maintaining high efficiency across $\eta - d_0 - z_0 - p_T$ phase space. In addition to being low risk and commercially available, this system is incredibly flexible and can scale or adapt in response to the future physics needs of the ATLAS experiment.

References

[1] ATLAS Collaboration, "Technical Design Report for the Phase-II Upgrade of the ATLAS TDAQ System," Tech. Rep. CERN-LHCC-2017-020. ATLAS-TDR-029, CERN, Geneva, Sept., 2017. https://cds.cern.ch/record/2285584

[2] ATLAS Collaboration, "Technical Design Report for the Phase-II Upgrade of the ATLAS Trigger and Data Acquisition System - Event Filter Tracking Amendment," Tech. Rep. CERN-LHCC-2022-004. ATLAS-TDR-029-ADD-1, CERN, Geneva, Mar. 2022, https://cds.cern.ch/record/2802799

abkahn@sas.upenn.edu