Muon integration in GPU Demonstrator

M. Bauce

Meeting GAP Roma - February 18, 2016

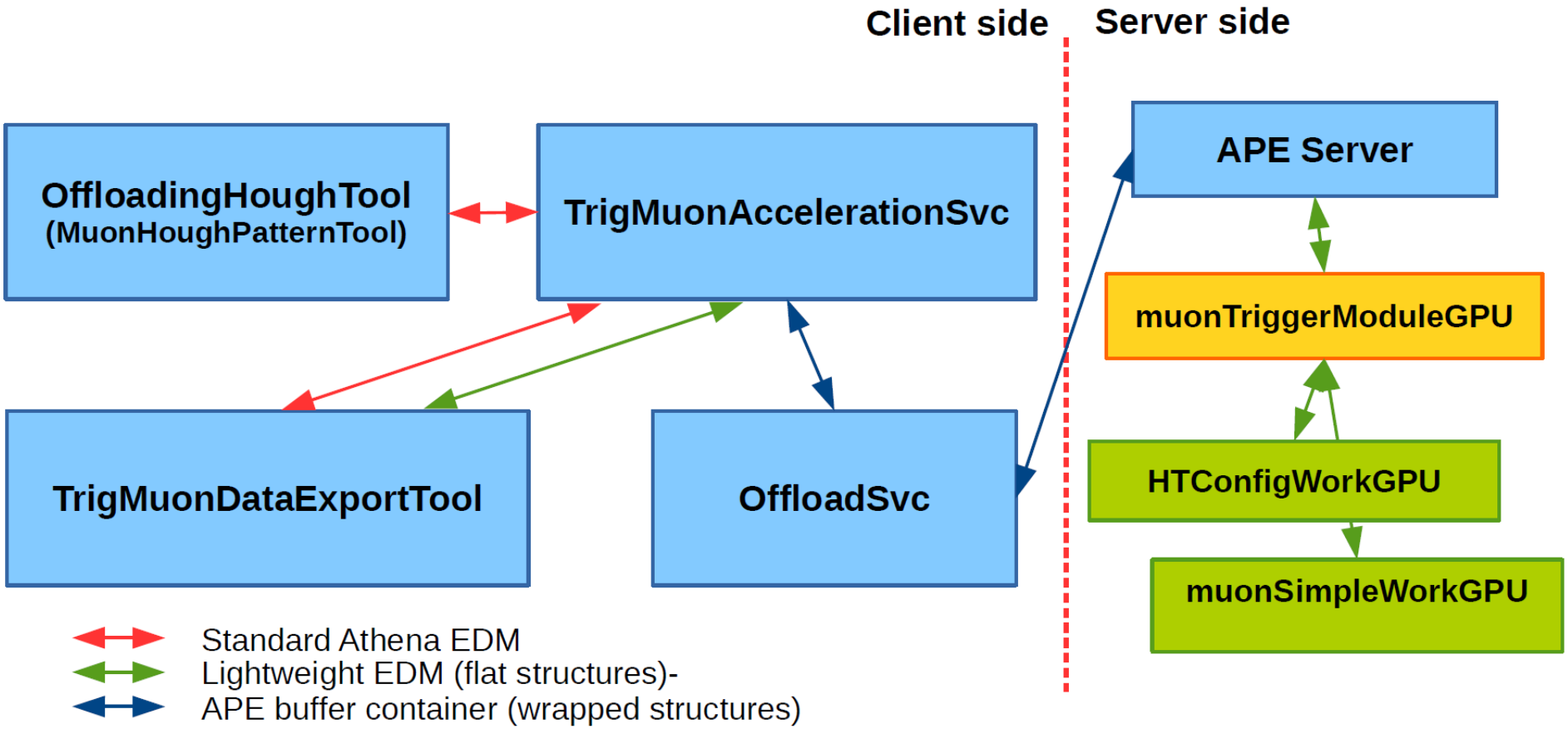
Introduction

Current involvement in an ATLAS-wide effort with contributions from UK, Lisbon, Roma, Bologna.

- GPU demonstrator: aiming at benchmark measurements of GPU deployment in ATLAS High Level Trigger
  - Algorithms studied as case study in different subdetectors: Inner Detector, Calorimeter, Muon
- Our contribution on a Muon algorithm
  - muon track segment reconstruction through an Hough transform

▶ Bulk of the code/framework in place: finalization and first measurements ongoing

Offloading architecture



- OffloadingHoughTool: replacement for athena hit-pattern finder tool
- TrigMuonAccelerationSvc: prepare input hits, algo configurations for offloading
- TrigMuonDataExportTool: convert algo configuration and MS hit data into lightweight EDM

Athena algorithm execution

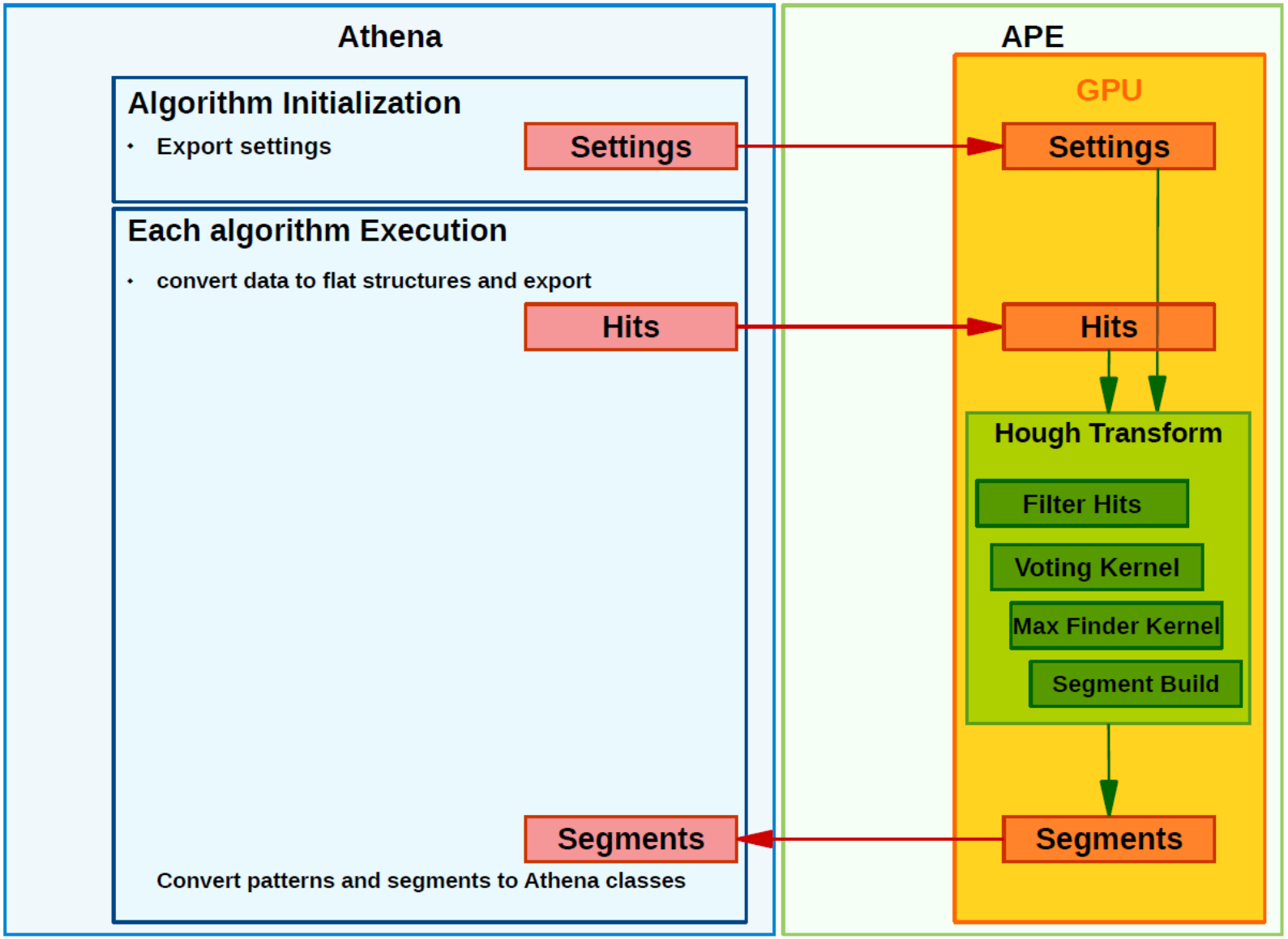
Segment reconstruction is the first step of the muon reconstruction chain

1. Collect hits from muon spectrometer in a Region of Interest (RoI) (or Full Detector)
2. Select hits in the forward region and reconstruct segments through an Hough Transform (straight) in the **xy** plane
3. Select hits in the central region and reconstruct segments through a **curved** Hough Transform in the **rz** plane
4. Pack everything together and send segment and patterns to the next step in the chain.

▶ So far only **xy** projection implemented on GPU, **rz** toward finalization (hopefully) by the end of this week.

- **rz** Hough Transform involves 10x more hits wrt **xy** and a similar factor for execution on CPU.

Data processing scheme



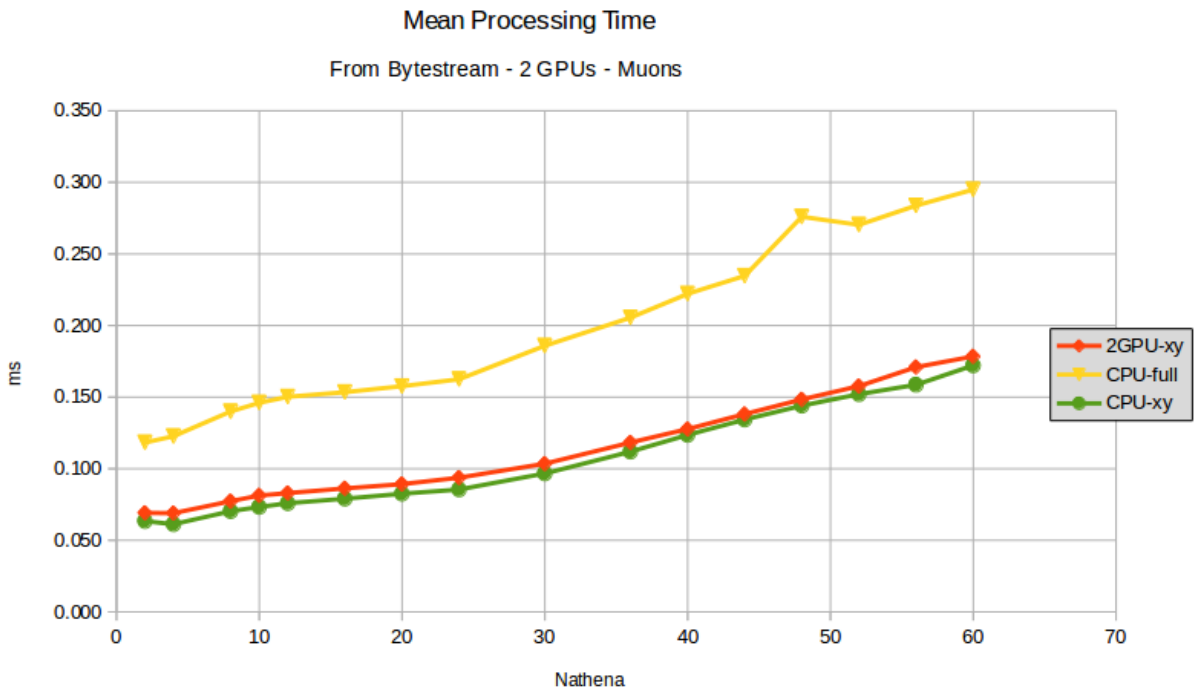
Kernel Design: Kernel structure

- **Sift Kernel:** Starting from input data representing hits, reduces the collection to hits fit for the voting part and the association to maxima part
- **VoteHough Kernel:** It fills the matrix representing Hough space using as input the fit-for-voting hits milked by Sift kernel
- **MaxFinder and Sorter Kernel:** For each sector the most voted bin is found and sector maxima are sorted according to the highest voting
- **ComputeOutput Kernel:** Using the maxima and the fit-for-association hits it computes the variables of the Hough Pattern and stores them to the final output struct representing the HP itself

Preliminary measurements

- Benchmark measurements using multiple istances of Athena ('Nathena') using AthenaMP, multiple
- Multiple (12) processes on the server available for round-robin algorithm execution
- 2 GPU exploited – K40

This configuration let us reach GPU resources saturation – optimal condition

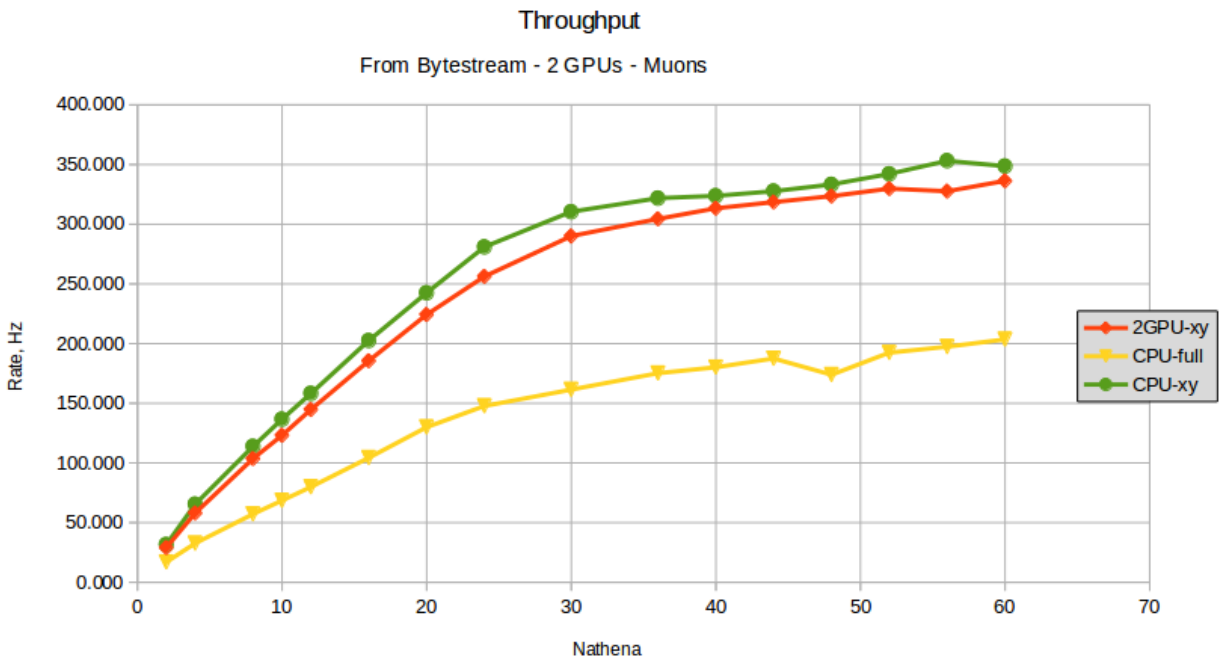


This time include overall athena execution (/event). GPU **xy** execution is 9-10% slower. **rz** projection is still the most time-consuming part.

Preliminary measurements

- Benchmark measurements using multiple istances of Athena ('Nathena') using AthenaMP, multiple
- Multiple (12) processes on the server available for round-robin algorithm execution
- 2 GPU exploited – K40

This configuration aim at reaching GPU resources saturation – optimal condition



This time include overall athena execution (/event). GPU **xy** execution is 9-10% slower. **rz** projection is still the most time-consuming part.

(Some) Technical issues

- Common effort(/decision) to switch from RDO to BS inputs:
  - BS reading is failing because a muon tgs hit retriever tool, fixed in a more recent framework release
  - going from RDO for the tests so far
- Common decision to swicth to a more recent framework release (Athena 20.1.4.1 $\rightarrow$ Athena 20.7.4.2)
  - Need to update all the code to comply to the updated algorithms
- Requested to switch to code compilation using cmake (instead of Makefile)
  - Easier to keep track of framework and package changing
  - Not so easy as expected

Conclusions

- No observed benefit from GPU deployment in the muon Hough Transform
  - haven't reached a computational intense scenario (yet) – CPU-GPU communication overhead still dominant
- Working ongoing (close to the end) for **rz** projection implementation on GPU
- Working to fix all the technical issues

Upcoming plans:

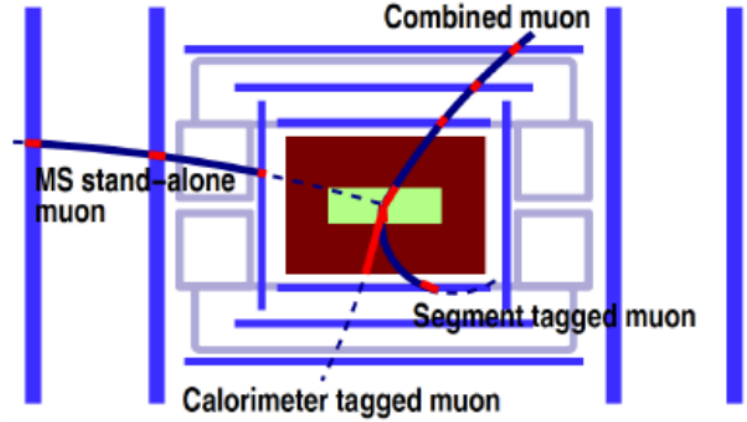
- Soon will have the complete code in place, easy-to-use: further measurements will follow
- Benchmark ATLAS measurements expected by end of May.

BACKUP

Muon reconstruction basic structure

Different classes of muons in ATLAS:

- equivalent in HLT and Offline
- Relying on Muon Spectrometer (MS) segments and Inner Detector (ID) tracks (and Calorimeter info)

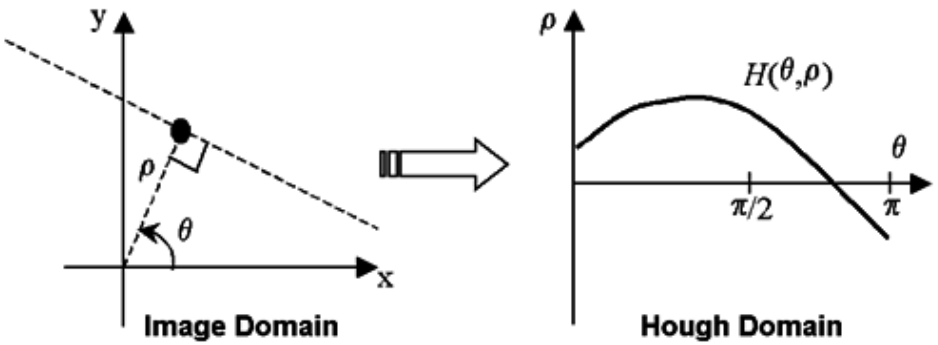


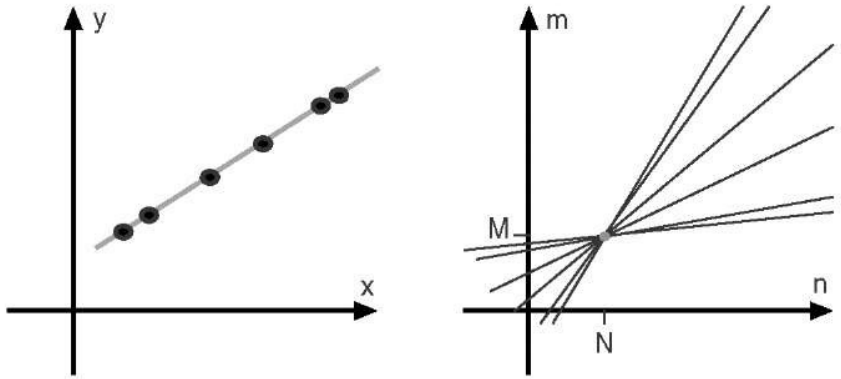
► Several steps of reconstruction relying on many many tools in Athena

1. Find segments in the MS ◄◄ Hough Trasform
2. Build Tracks in the MS
3. Extrapolate to primary vertex (PV) obtaining a Standalone Muon (SA)
4. Combine MS muon with ID tracks for a Combined muon (CB)

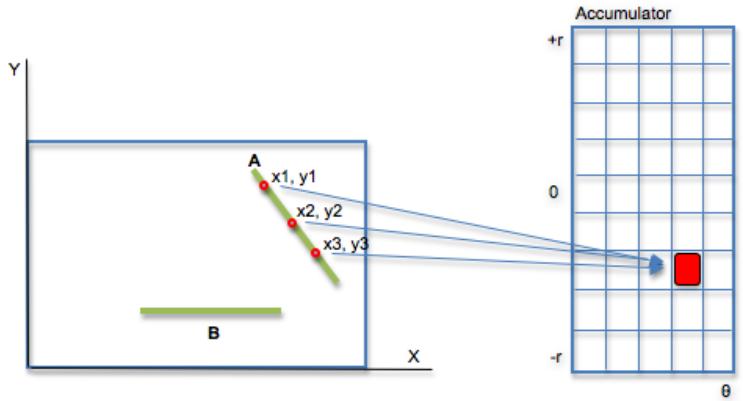
Hough Trasform in a nutshell

Replace track fitting with search for maxima in the Hough (dual) space





- Each point correspond to a *path* in the Hough space
- Points belonging to the same track produces *crossing paths*



- Tracks correspond to *maxima* in a Hough space *accumulator*

Kernel Design: Kernels' score

- Algorithm can be executed up to **5 times** provided hits and number of maxima found are enough
- **N.B.** Only XY version translated so far: CurvedAtCylinder in finalization
- **Average Times** for Muon Standalone execution (100 events per worker) for the full loop
  - Preparatory variable setting (host): 0.9 ms
  - CUDA part (Mem I/O + Kernels): 1.369 ms
  - Total Worker execution: 3.234 ms
- Residual overhead to be measured and hopefully its impact reduced (see memory usage considerations)
  - Algorithm still needing some validation-relevant tweaks from 2^{nd} iteration onward may reduce time by removing not-needed iterations