GPU deployment in Atlas High Level Trigger

M. Bauce

Meeting GAP - Ferrara - 23/03/16

SAPIENZA
UNIVERSITÀ DI ROMA

GAP RT

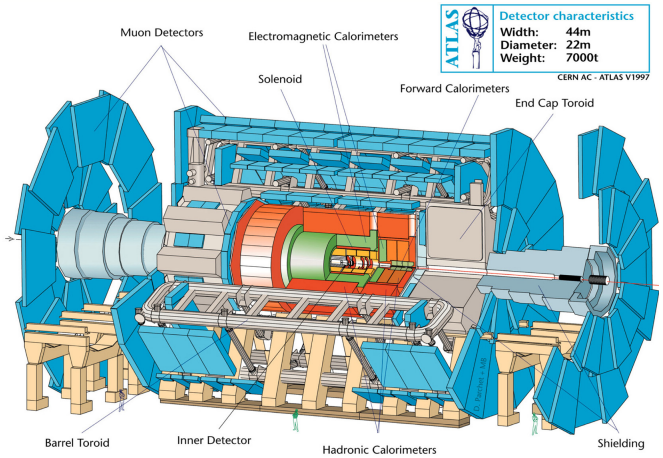
INFN

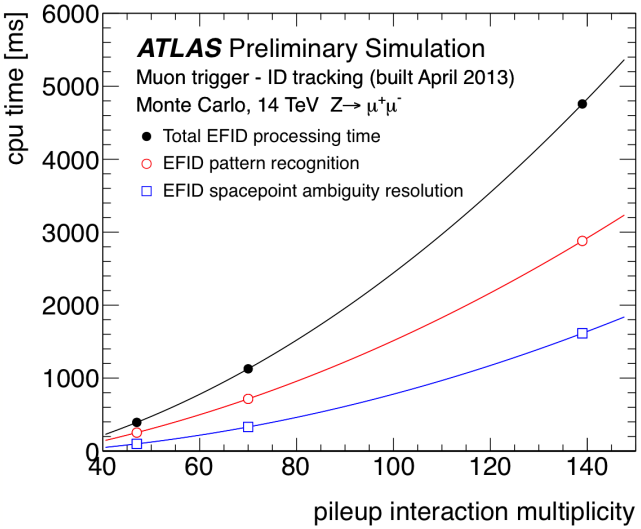
Introduction

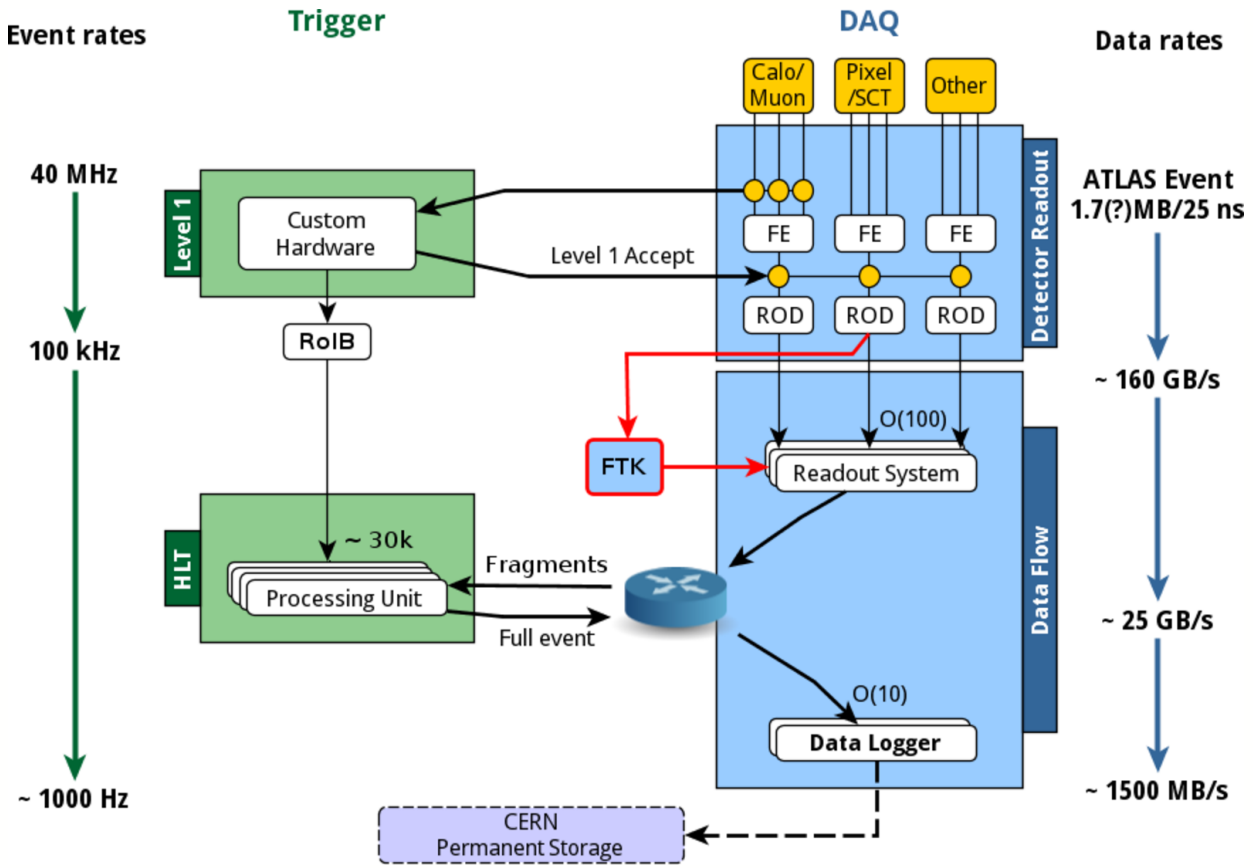
▶ HEP activity in Rome focusing on GPU deployment in ATLAS High Level Trigger (HLT) for muon reconstruction

- First *hand-made* implementation started ∼2 years ago
- Current involvement in an ATLAS-wide effort with contributions from UK, Lisbon, Roma, Bologna.
  - ▶ GPU demonstrator: aiming at benchmark measurements of GPU deployment in HLT upgrades
  - ▶ Algorithms studied as case study in different subdetectors: Inner Detector, Calorimeter, Muon
- Finalization part ongoing, preliminary measurements available

ATLAS and its Trigger system



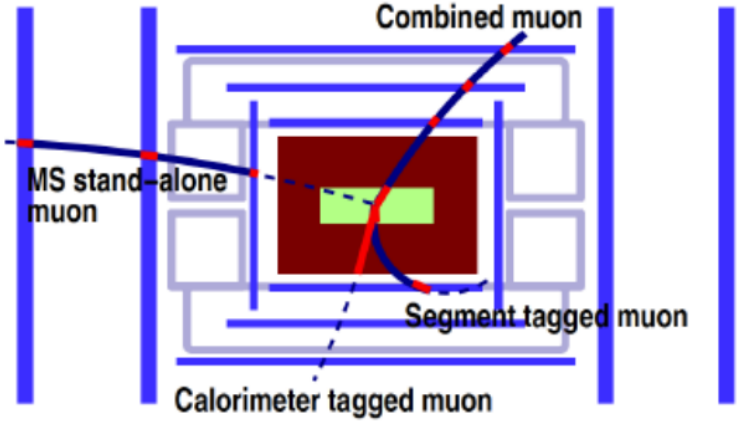




Muon reconstruction basic structure

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)

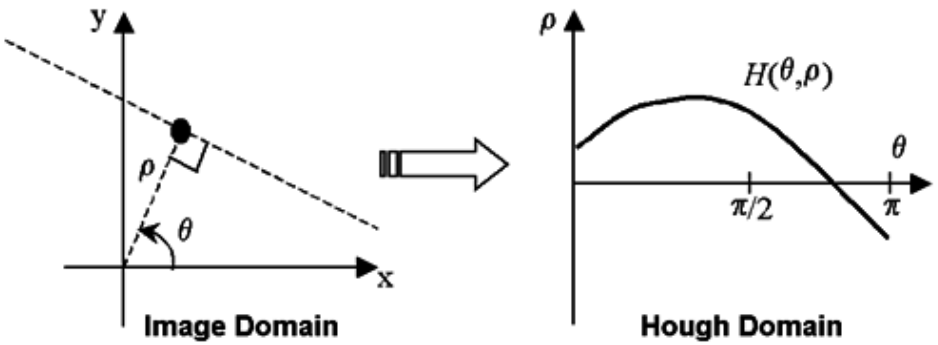


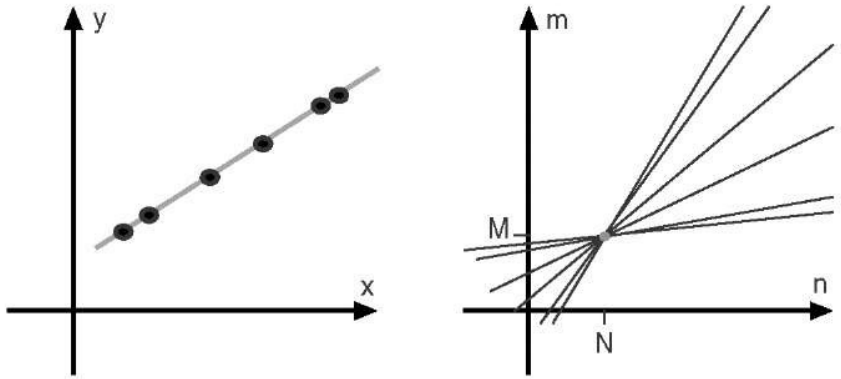
► Several steps of reconstruction relying on many tools passing informations to each-other

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)

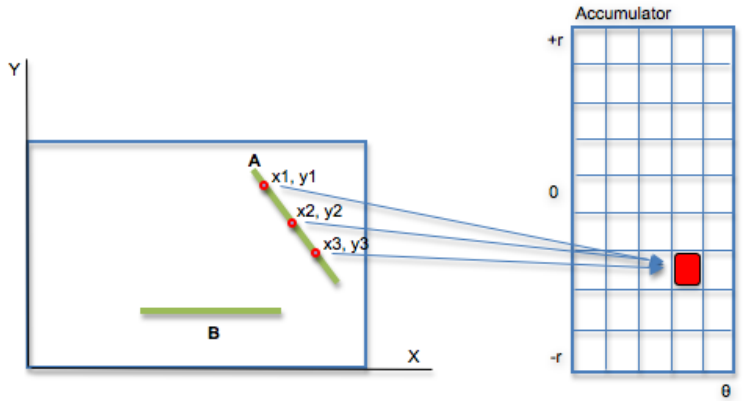
Hough Trasform in a nutshell

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





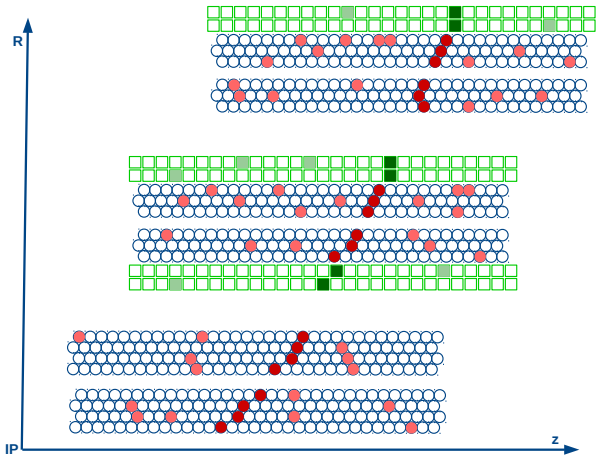
- Each point correspond to a *path* in the Hough space
- Points belonging to the same track produces *crossing paths*



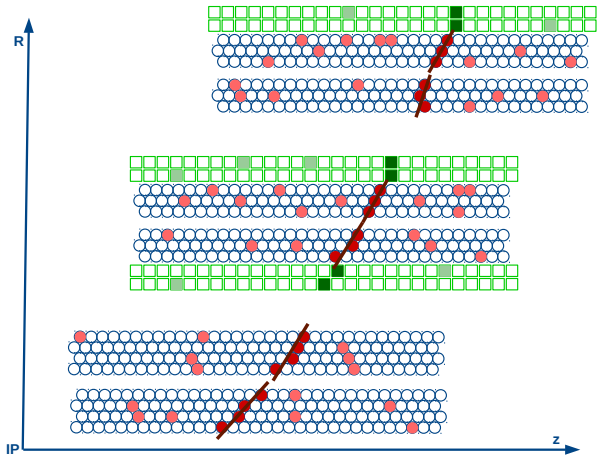
- Tracks correspond to *maxima* in a Hough space *accumulator*

Algorithm implementation

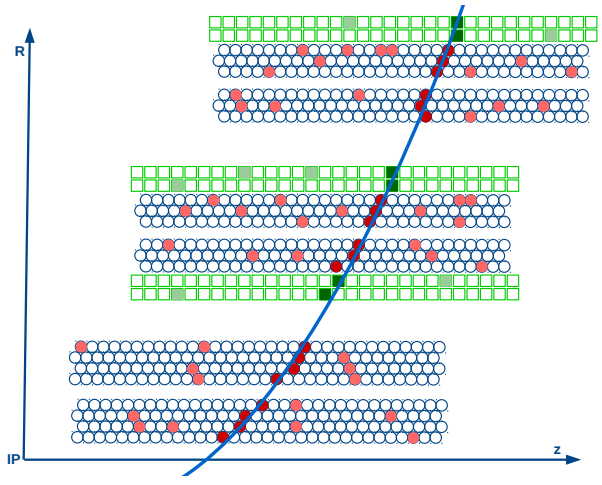
Muon Spectrometer hits



Segment reconstruction



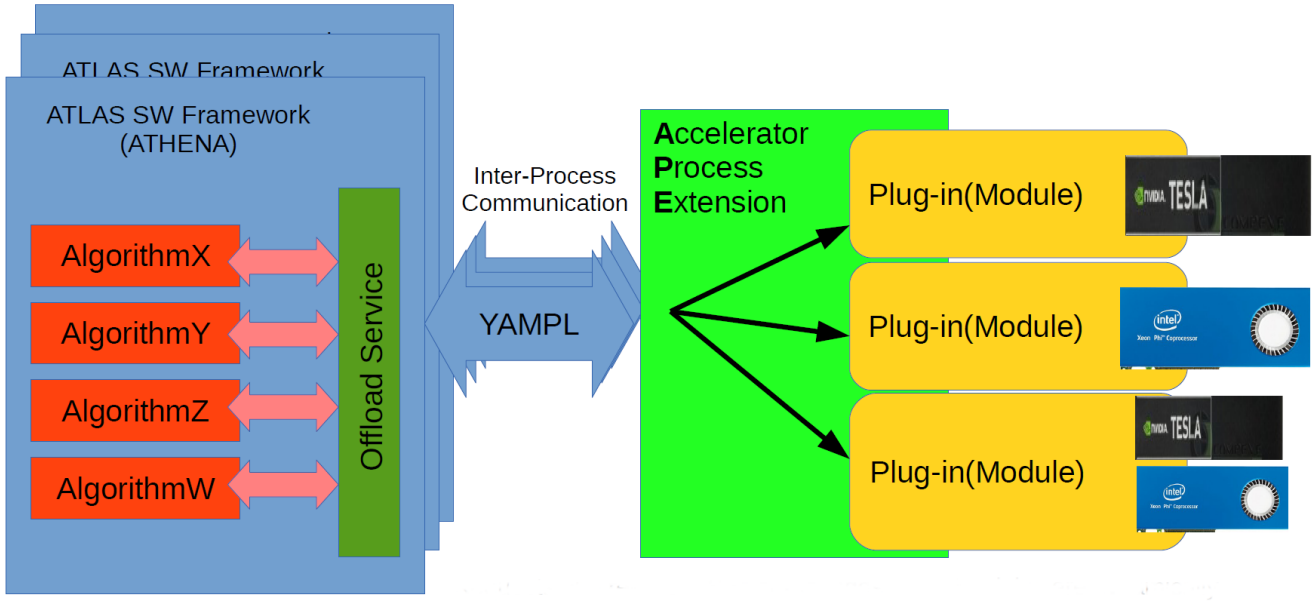
Track building



Muon segment reconstruction is the first step of the muon reconstruction chain

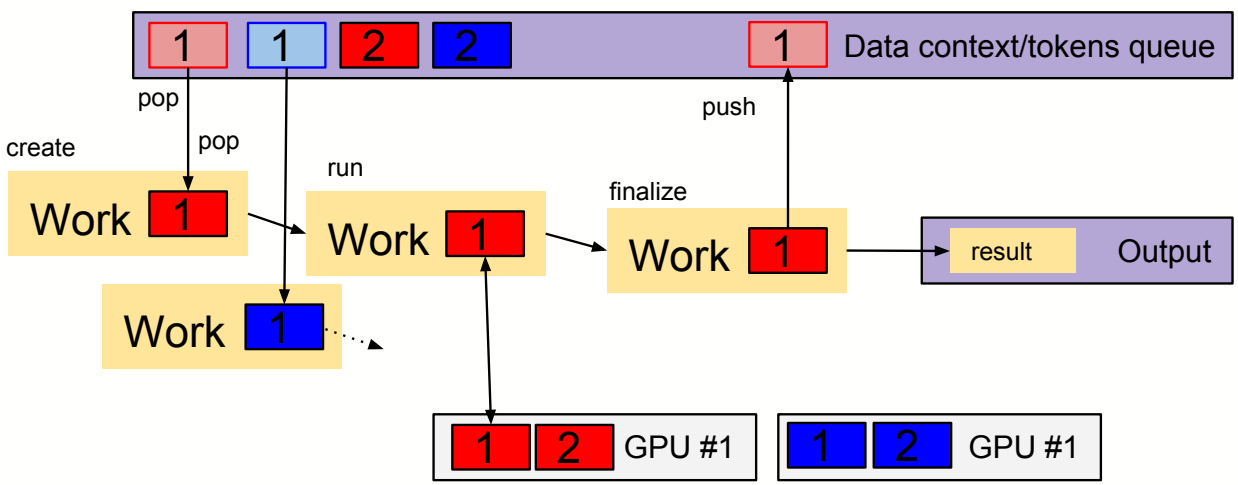
1. Hits collected from muon spectrometer in a Region of Interest (RoI) (or Full Detector)
2. Segments reconstructed in the transverse (x, y) plane first as straight segments
3. Segments reconstructed in the (r, z) plane of the detector as curved trajectories
4. Projected segments are combined by a different tool and returned to the chain

CPU-GPU communication framework

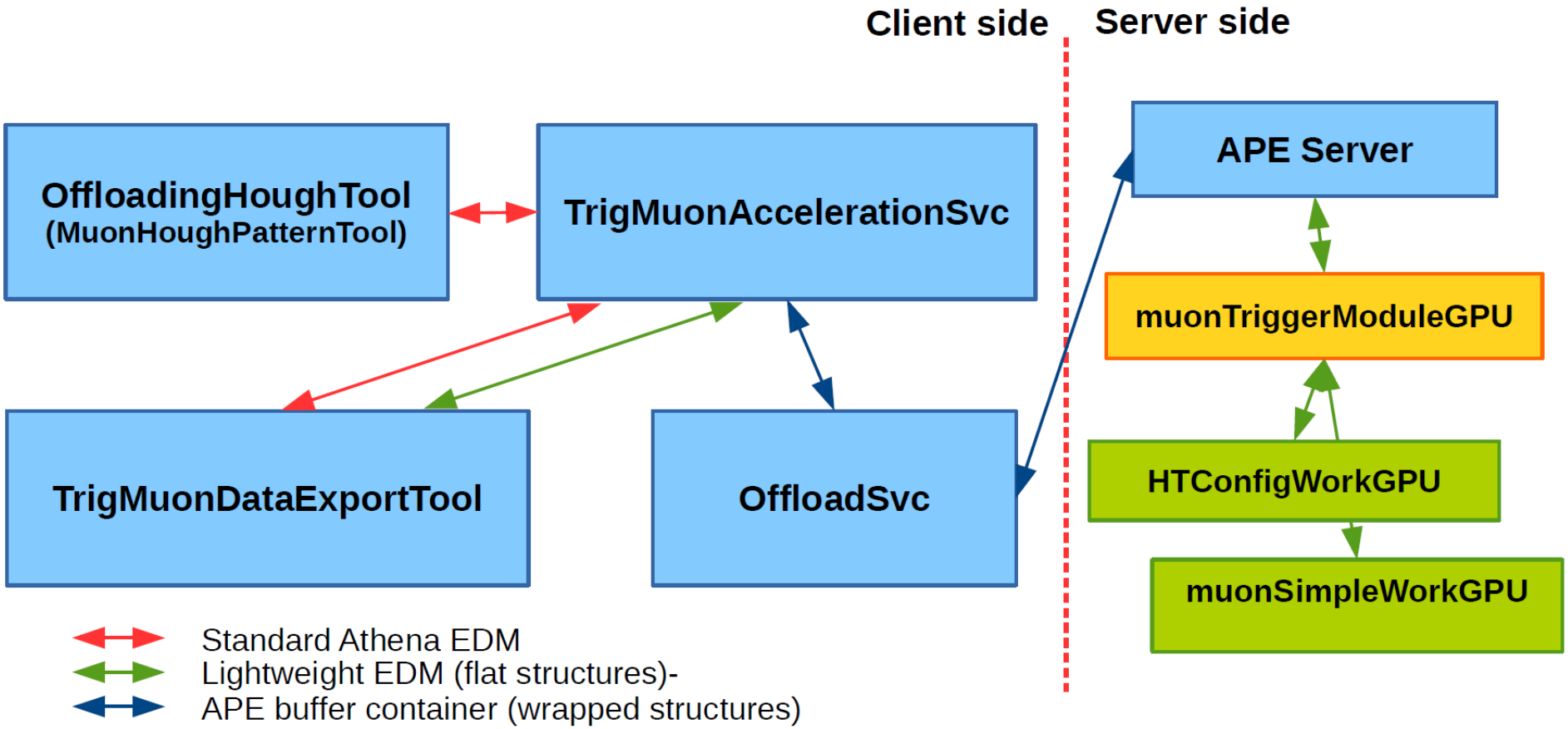


Server-Client
Flexible approach to integrate in pre-existing framework.

APE server
Algorithm execution queue aiming to saturate GPU computing capability

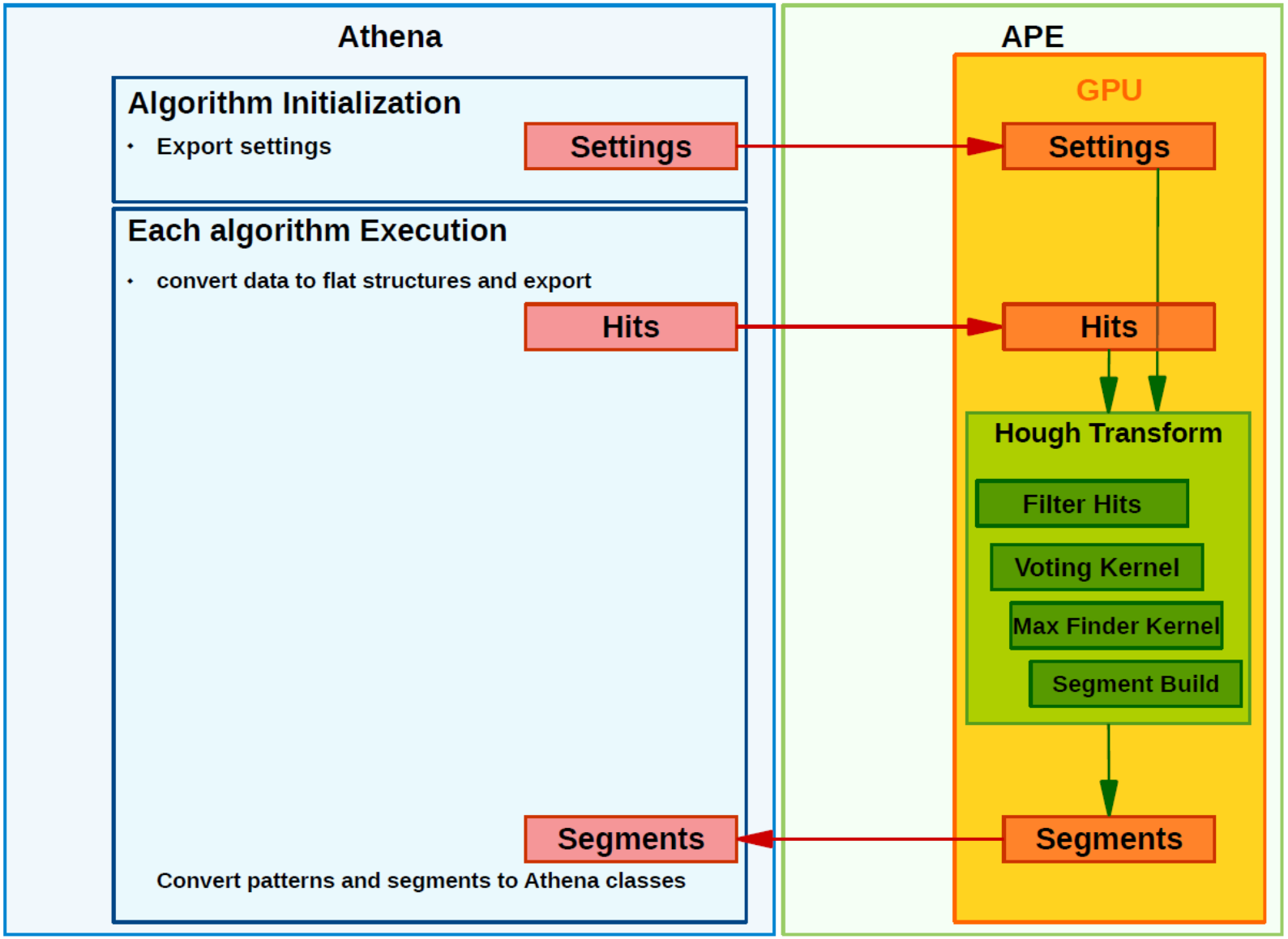


Offloading architecture



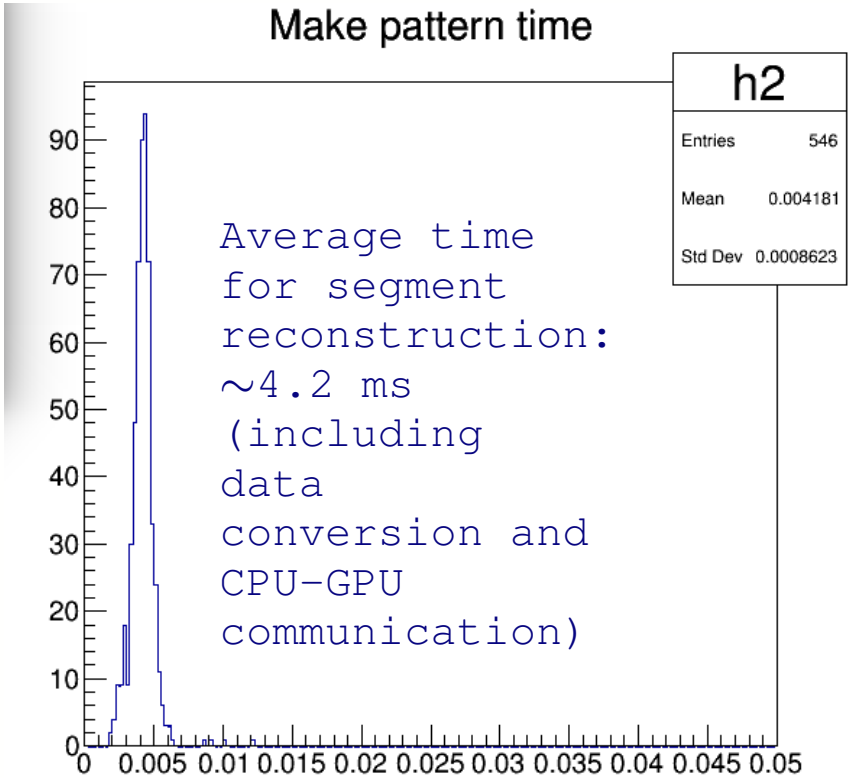
- OffloadingHoughTool: replacement for athena hit-pattern finder tool
- TrigMuonAccelerationSvc: prepare input hits, algo configurations for offloading
- TrigMuonDataExportTool: convert algo configuration and MS hit data into lightweight EDM

Data processing scheme



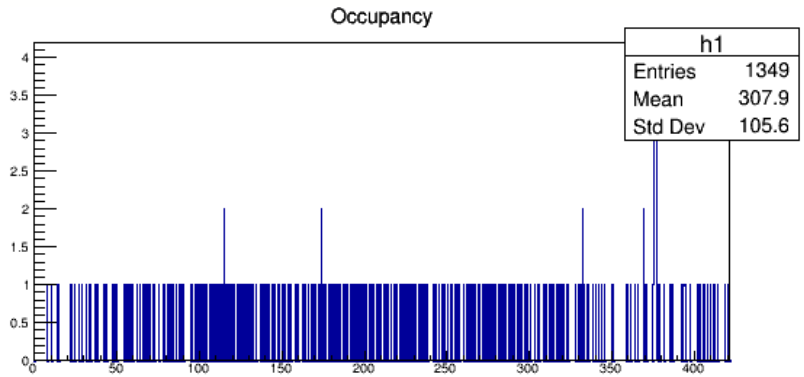
Preliminary measurements

- Benchmark measurements using multiple istances of Athena clients running in parallel (on CPU) to process the events
- Multiple (12) processes on the server available for round-robin algorithm execution
- 1 GPU exploited – K40 (4 available overall)



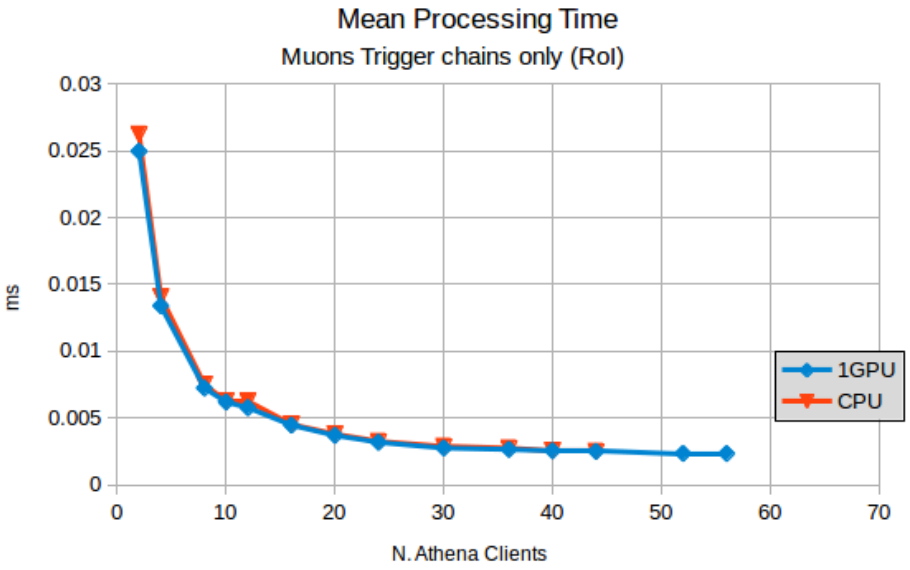
Average time for segment reconstruction: ~4.2 ms (including data conversion and CPU-GPU communication)

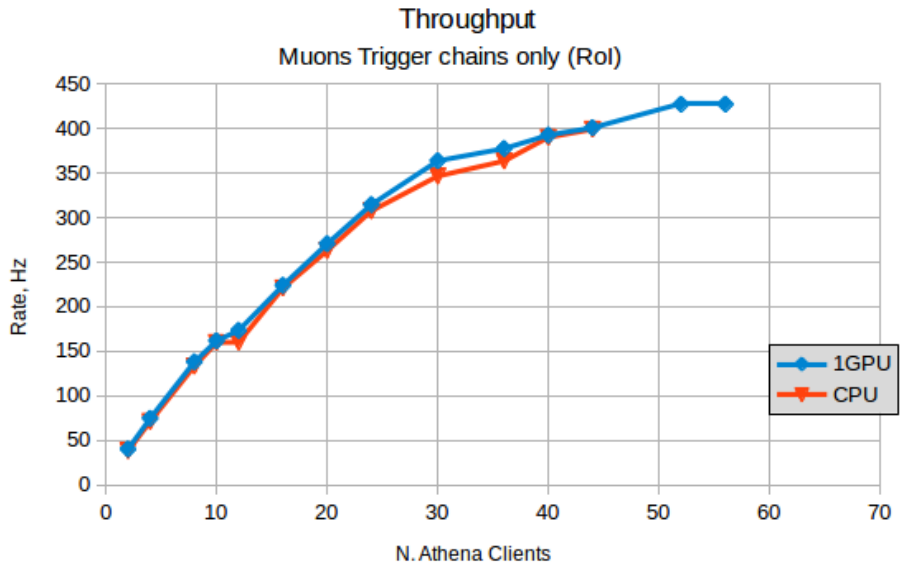
With 44 parallel client running the GPU offloaded part is not saturating the Server-GPU.

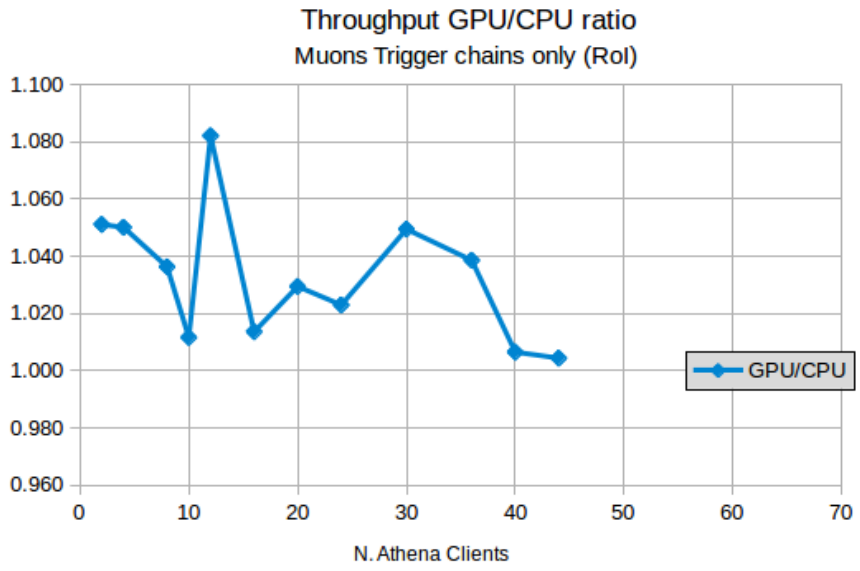


Active GPU processes through time

Preliminary measurements





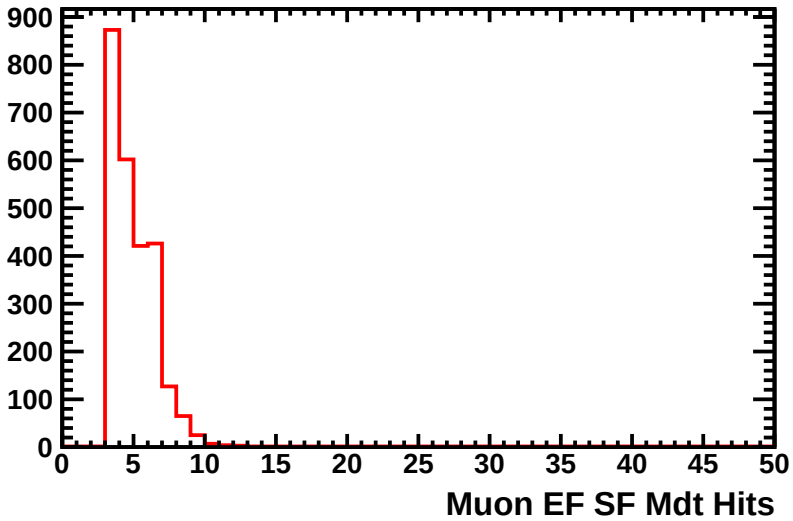


- No significant difference between CPU–GPU
- Rate/Time dependent on trigger chain considered
  - RoI muon only
- Need to consider further scenarios *(ongoing)*

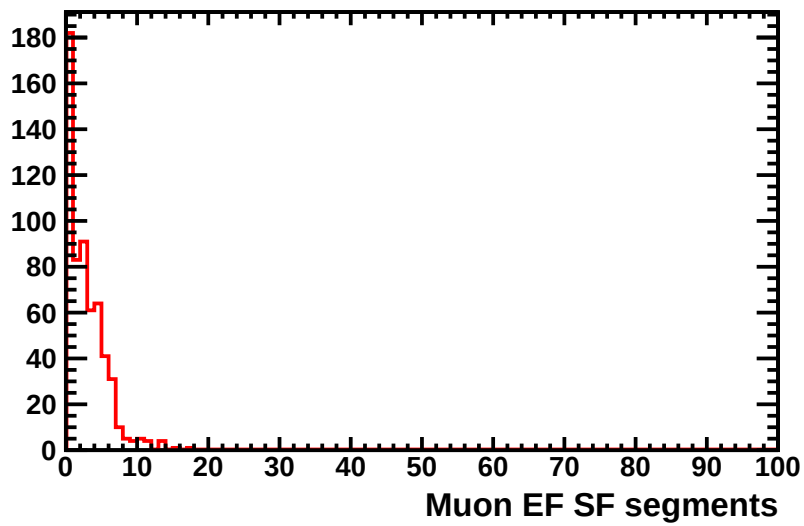
Preliminary measurements

Tipical multiplicities for the tests reported are not sufficiently large to get significant benefit from GPU deployment

Number of MS hits



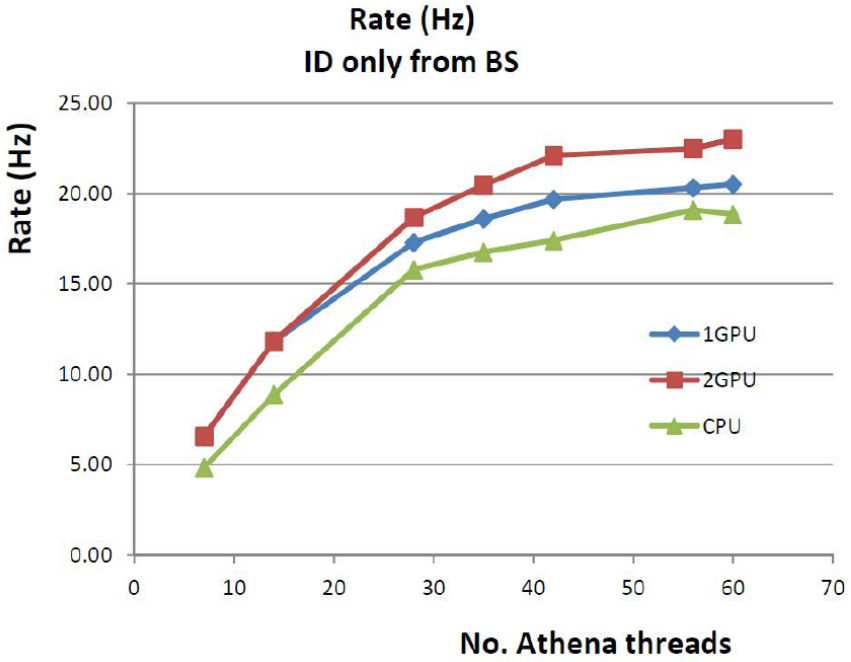
Reconstructed segments

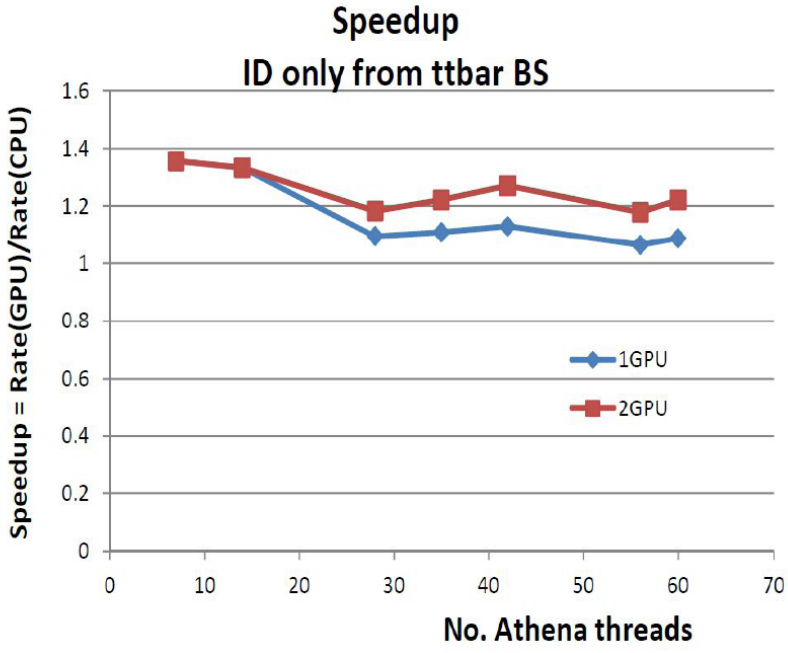


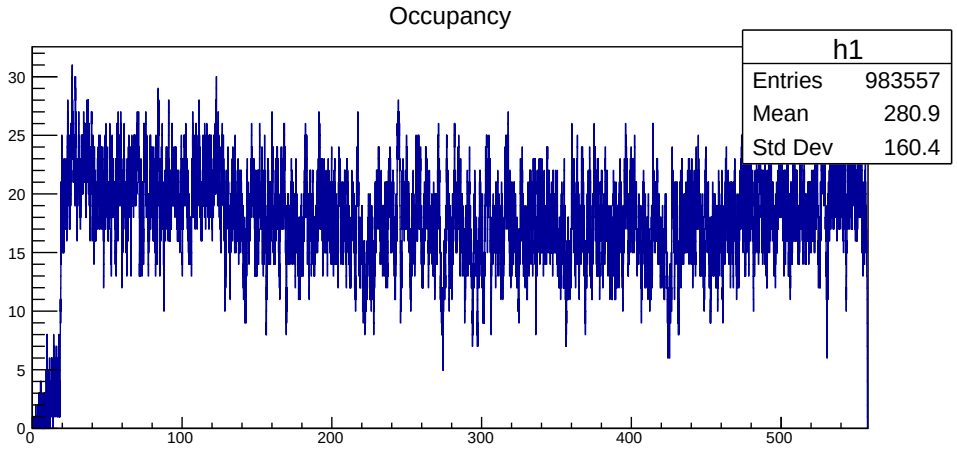
- RoI based reconstruction has reduced amount of data to process → considering algorithms with detector full scan
- Track multiplicities are dependent on the collision average luminosity (pileup) → will test more dense environment

Inned Detector reconstruction

(for comparison)







- Only ID reconstruction executed
- 29% of ID code is offloaded to the GPU
- ×5.5 speedup for accelerated ID part

Conclusions

- Work ongoing to evaluate the best gain achievable from muon GPU deployment
  - no significant benefit observed in the preliminary tests
- Diagnostic tools under finalization (GPU resource usage and timing)
- Standalone client developed for algorithm optimization studies
- Fully integrated in ATLAS Demonstrator activity
  - Benchmark measurements under definition: need to optimize a typical HLT scenario for upcoming LHC environment
  - Aiming at some *realistic* measurements by the end of May

Upcoming plans:

- Investigate possibility to offload on GPU different algorithm
  - So far manpower-limited – now infrastructure and model available
  - Room for *standalone* innovative algorithm development

BACKUP

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

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

Trigger Muon Reconstruction