# GPU Utilization in ATLAS

Sami Kama

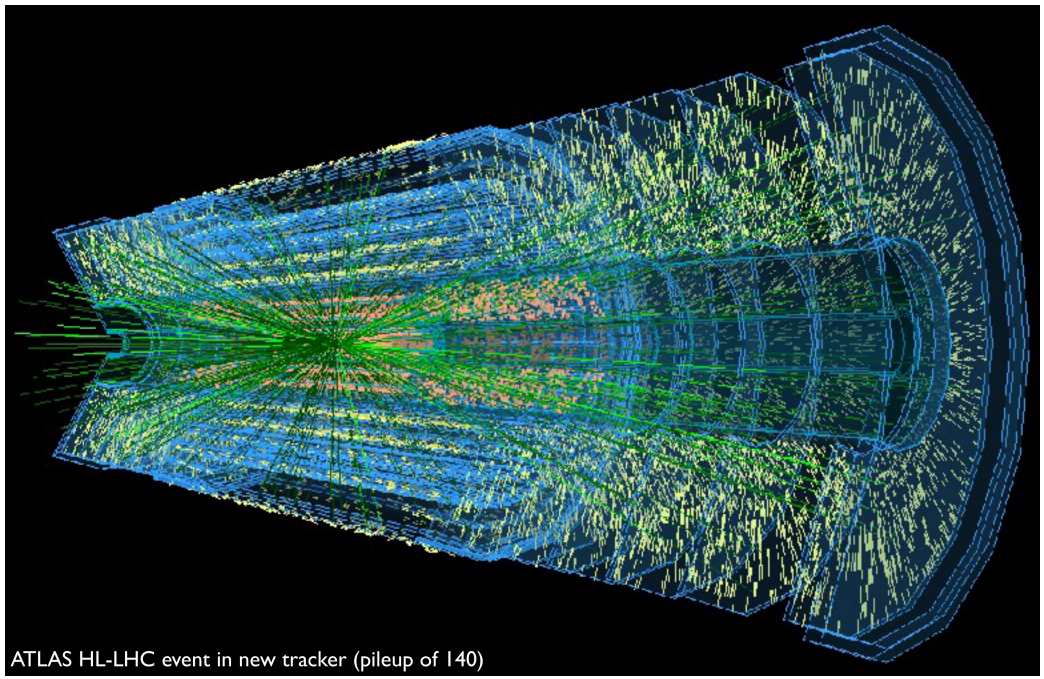
Southern Methodist University

2016-01-12

# Outline

# GPUs?

- Increasing LHC rate and luminosity leads to combinatorical increase in computational requirements, especially for tracking
- Many core machines already necessitates parallel processing
- Can we use GPGPUs to massively parallelize reconstruction?
  - In online reconstruction (Trigger)
  - In offline reconstruction (T0-Grid)
- Main criteria is the effective cost!


ATLAS HL-LHC event in new tracker (pileup of 140)

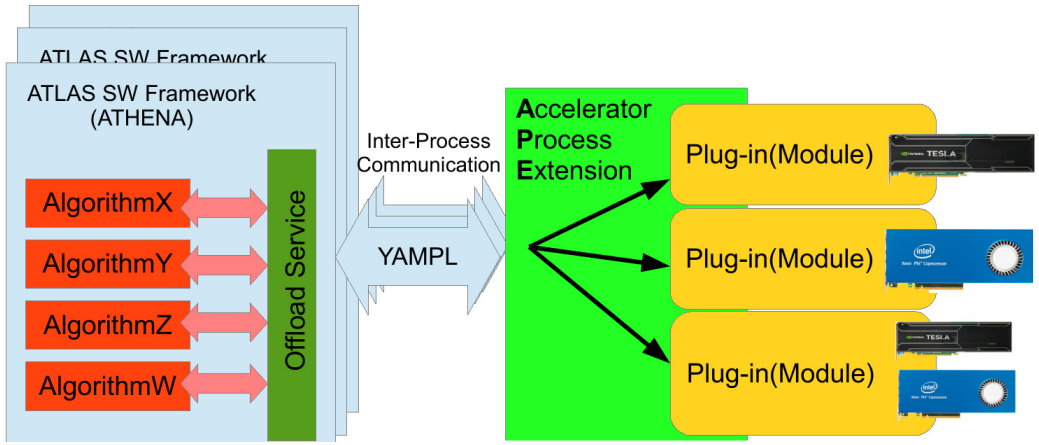
# Past Studies

- Prototyping in Trigger
  - Level 2 trigger inner detector tracking algorithm was ported to GPUs in order to study feasibility of the system
  - Upto 12x speedup achieved wrt single threaded CPU implementation
- Prototyping in Offline
  - Prototyping of Kalman filter based tracking achieved ~20x speedup wrt standard CPU implementation
  - Some ME calculations in Monte-Carlo programs are ported to GPUs

# Framework support

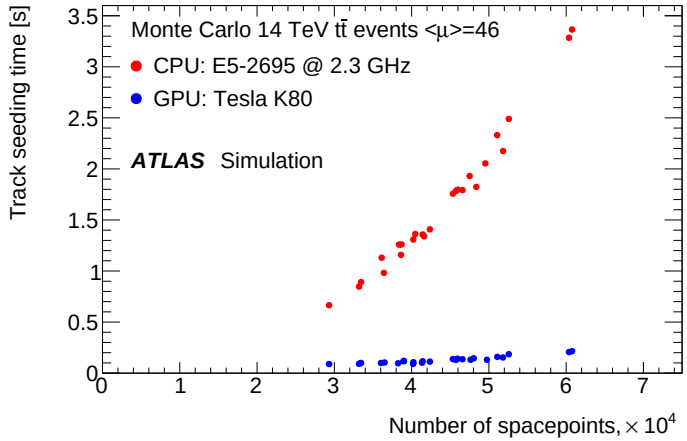
Introduced framework support for using offloading platforms through Accelerator Process Extension(APE) Framework.

- Hides details of implementation from framework
- Manages connection to multiple Athena processes
- Can run on the same host or a different host (Top-of-Rack machine)
- Can run on different platforms and architectures, as long as message format is consistent



# Current Status

- Trigger GPU Demonstrator group is formed to study feasibility of GPU usage in Trigger farms. Studies are extended to three domains
  - Inner Detector, tracking based on Cellular Automata(CA)
  - Calorimeter, jet finding and clusterization based on CA
  - Muon, tracking based of hough transforms
- All modules have been implemented, more code is being ported to GPU
- Good speedup so far,
  - optimization studies are on going
- Feasibility report at the end of February



# Issues

Adaptation in offline world is slow, mainly due to lack of manpower and availability of HW on Grid sites

- Few developers can write proper parallel code
- Even fewer can write GPU code
- HEP reconstruction code usually requires expert domain knowledge
- Due to limited computing budgets, existing experts dedicating most of their time to maintain current SW

- Grid sites usually choose their HW to serve a diverse scientific community
- Availibility of GPUs in the farms depend on other branches of science as well
- Different procurement policies result in very diverse HW
  - Code maintenance issues

# Issues

- Cost effectivenes requires up to 80% of the code to be run on GPU side
- However most of the algorithms are not parallelizable
  - Need to invent new algorithms or paradigms
- There is no standardization, technology is evolving too fast (or standardization is too slow)
- It takes about 5 years from R&D code to production code
  - About 3-5 new generations of chips
- Under controlled environments (such as trigger farms) can be managed
- Common development effort and algorithm R&D might help!

# Summary

- ATLAS has been investigating GPU utilization since Run 1
- Trigger GPU Demonstrator project is producing promising results
- Adaptation in offline is slow due to manpower, hardware diversity and maintenance issues
- Need to invent new paradigms and algorithms to exploit GPUs
- A common R&D effort could help in overcoming obstacles in GPU utilization

# THANK YOU