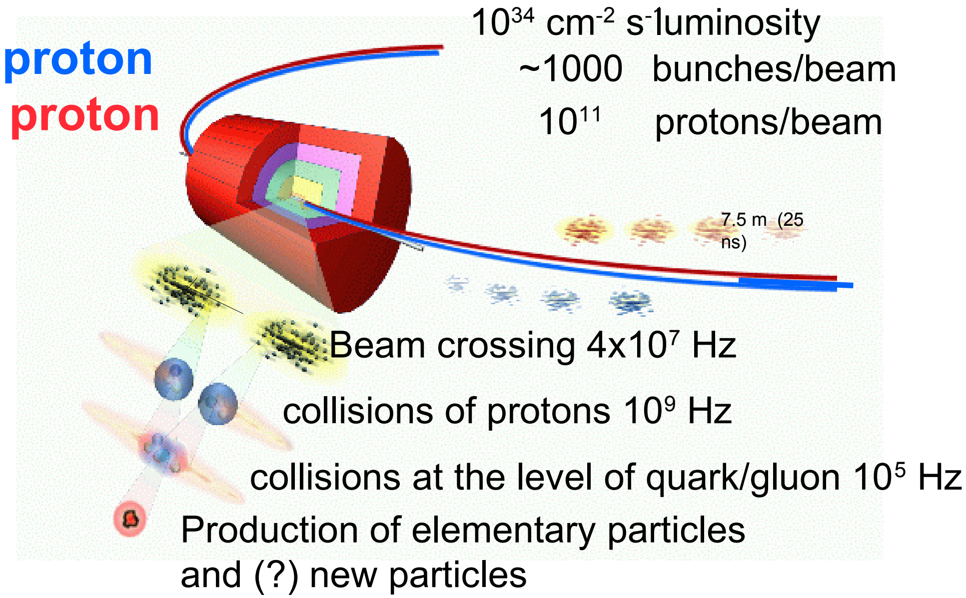
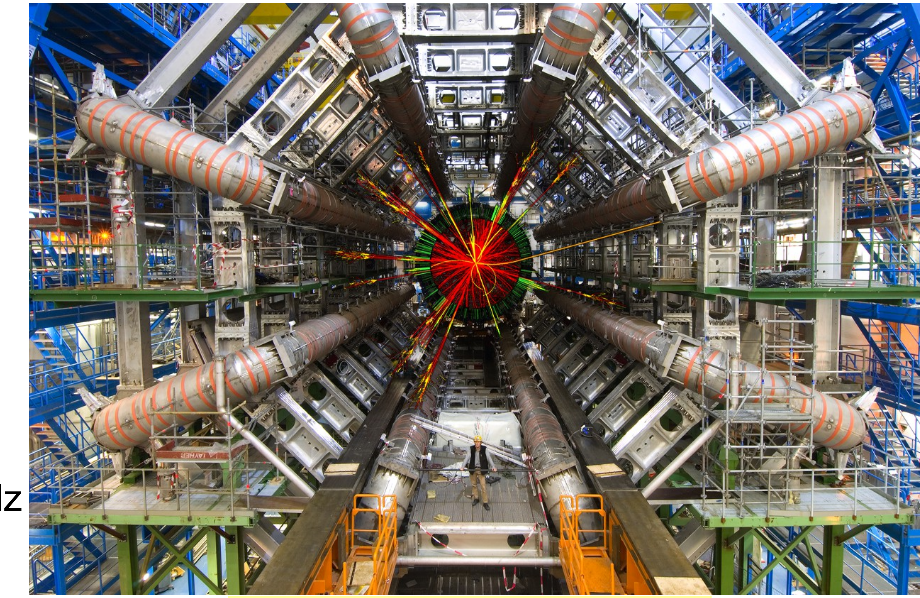
ATLAS usage of GPU
(within French groups)

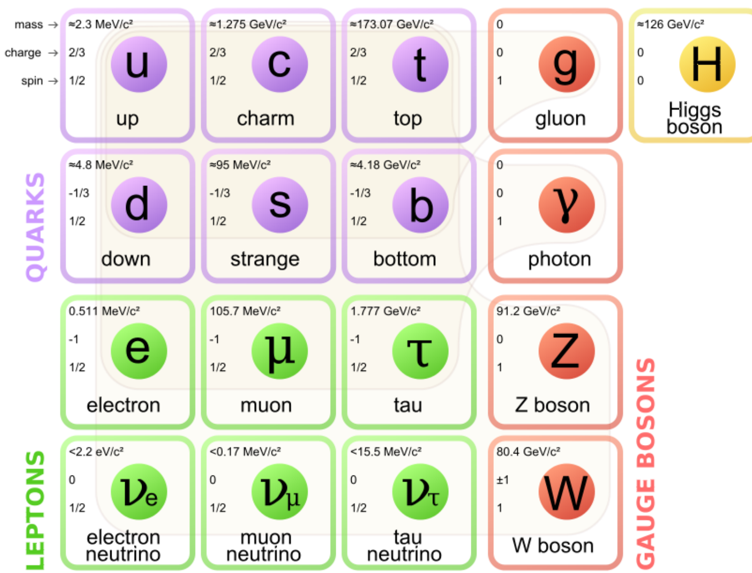
Frédéric Derue, LPNHE Paris
(on behalf of Computing ATLAS France group)

Workshop on GPU
CC-IN2P3 Lyon, 4th April 2019

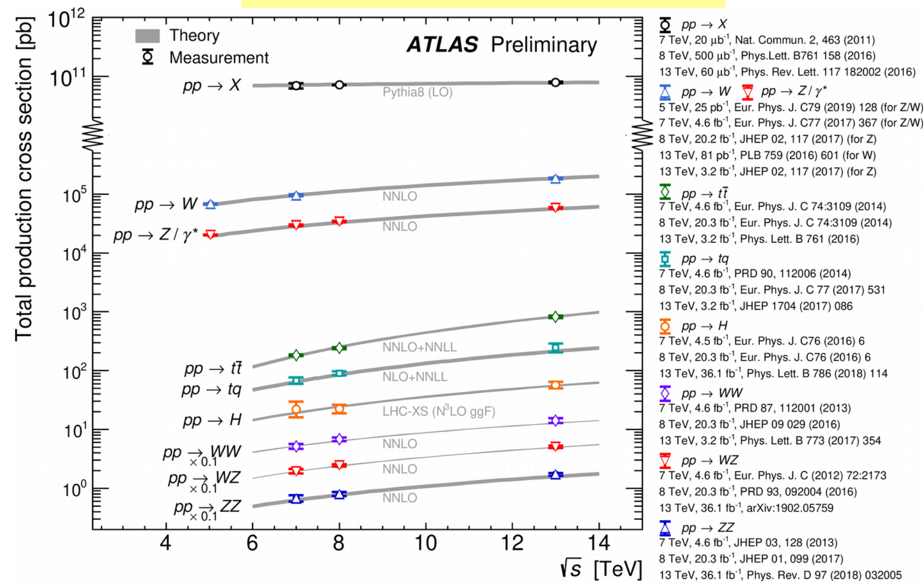
ATLAS and the Large Hadron Collider at CERN







ATL-PHYS-PUB-2019-010



Why ATLAS is interested in HPC/GPUs ?

- **Long term activities : High-Luminosity LHC (HL-LHC) Upgrade**

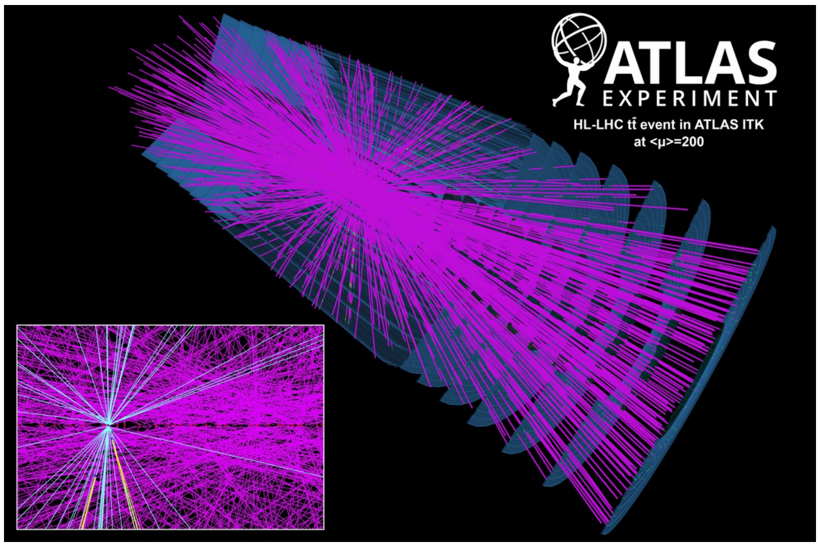
  - the HL-LHC represents the ultimate evolution of LHC machine performance
  - operation at up to L=7.5×10^{34}Hz/cm² (LHC run-2: 2×10^{34}) to collect up to 3000 fb^{-1} of integrated luminosity
    - ➢ vast increase of statistical reach, but challenging experimental conditions
    - ➢ up to 200 p-p collisions per bunch crossing
      - mitigated by extensive upgrades of experiments during LS3

- **Today and middle term activities**
  - Machine Learning : classification, regression, scan parameters

Event processing at HL-LHC

- **Higher luminosity**
  - more interactions per crossing
- **Increased event rate**
- **Bigger and more complex events**
  - ITk with>& 5B channels

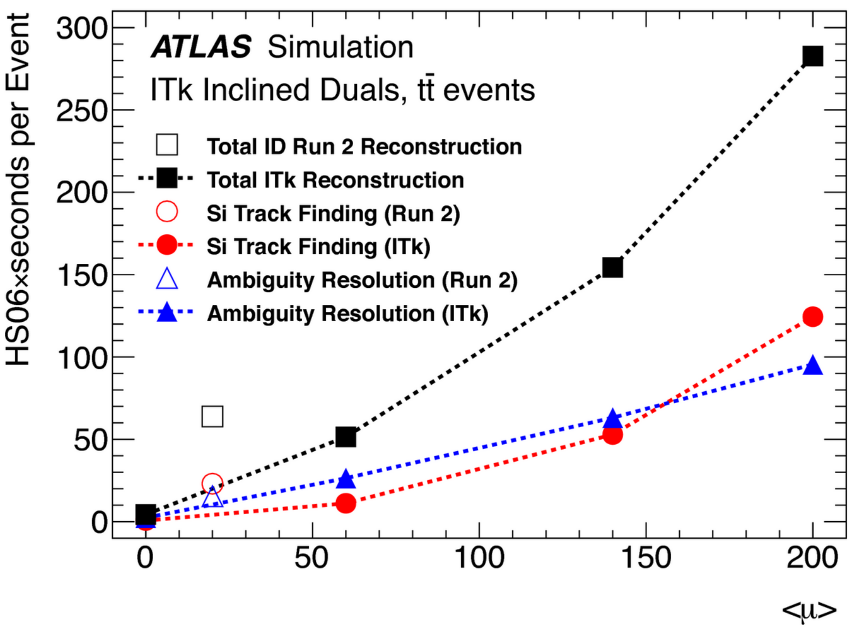
a simulated $t\bar{t}$ event at average pile-up
of 200 collisions per bunch crossing
[*Upgraded Event displays*]



- **Better physics performance**
  - improved algorithms
- **Better CPU efficiency**
  - better software engineering

- **Reconstruction**
  - environment is challenging in terms of CPU time for reconstruction
- **Multithreaded running at the event level**
  - exploit parallelism
  - technology watch
- **use diverse hardware architectures**

CPU required in HS06×seconds to reconstruct an event
in the ITk as a function of the average pile-up
[*Itk pixel TDR plots*]

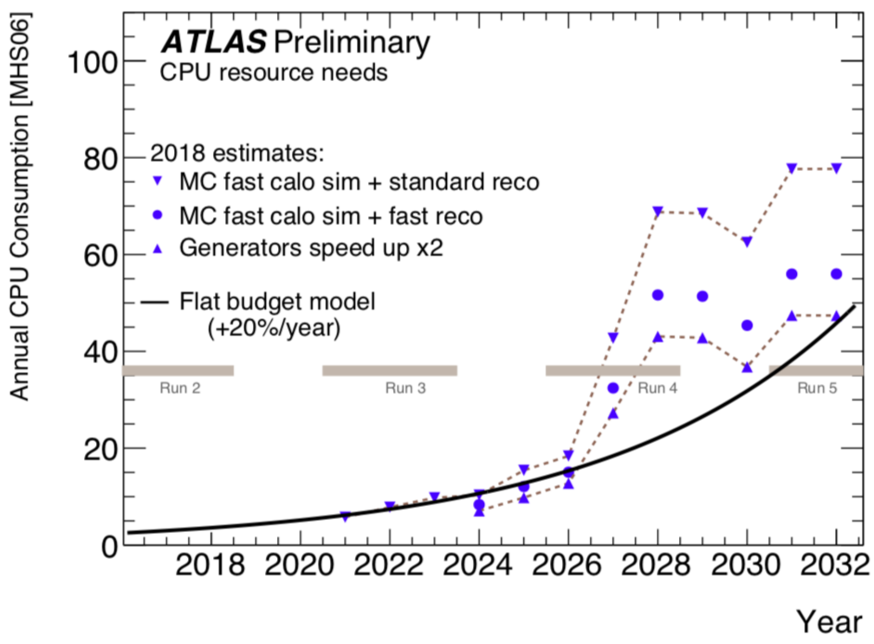


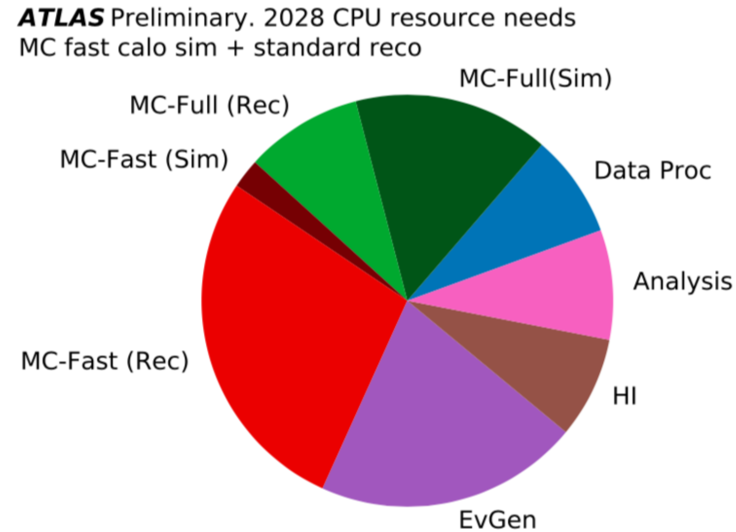
CPU projection for HL-LHC

ComputingandSoftwarePublicResults

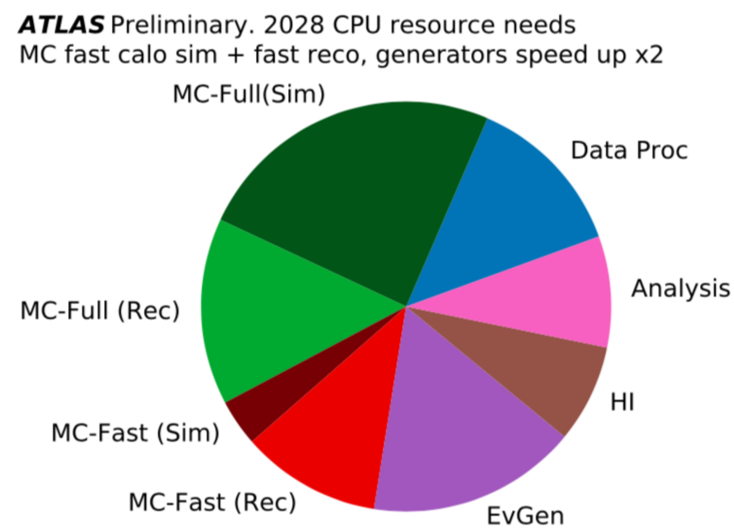
- **Fast vs Full simulation**
  - Run 3: 50% of simulation with fast sim
  - Run 4: 75% of simulation with fast sim

Estimated CPU resources (in MHS06) needed for the years 2018 to 2032 for both data and simulation processing





2/3 of global CPU time for simulation



Using accelerators and HPC

- **Supercomputers are evolving away from the « usual » hardware we have on WLCG resources**
  - e.g Summit Power9 + Nvidia V100
  - other architectures becoming popular – ARM
  - challenge of portability

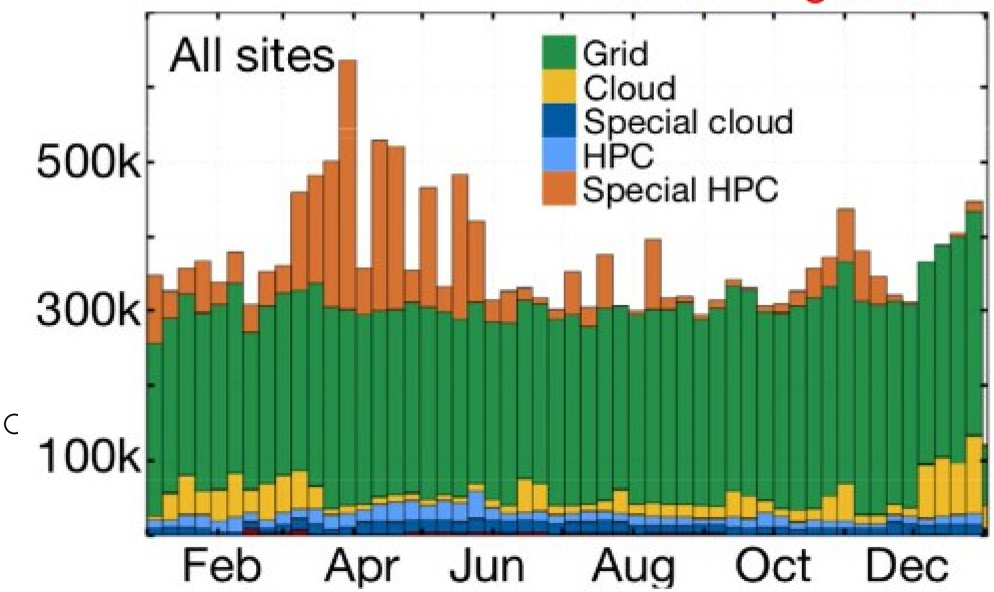
- **ATLAS is efficicient at using various resources**
  - grid
  - cloud
  - volunteer computing
  - HPC

- **Use of GPU on WLCG grid**
  - Manchester WLCG site has currently 2 grid queues setup for GPUs

ATL-SOFT-SLIDE-2019-068

D. Constanzo, WLCG2019, Mar 2019

CPU usage 2018



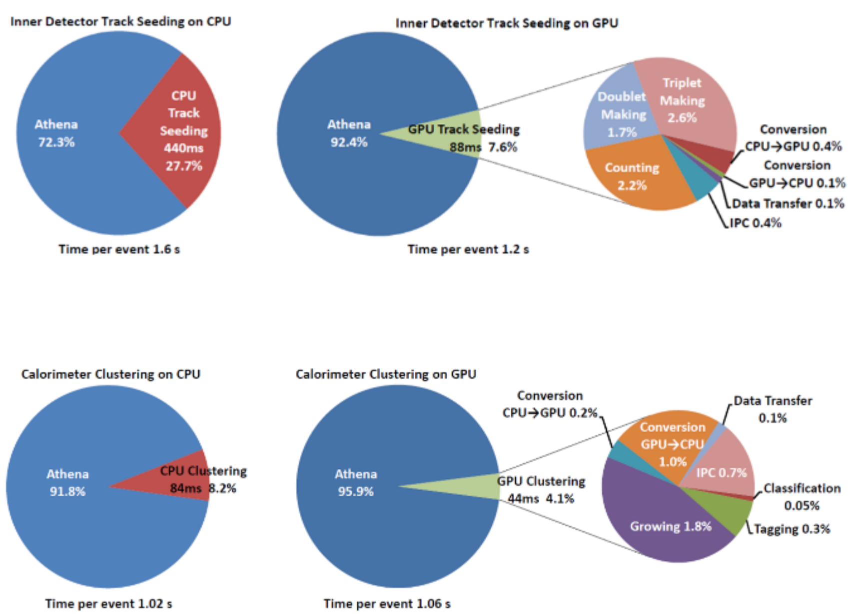
Using accelerators at triggering level

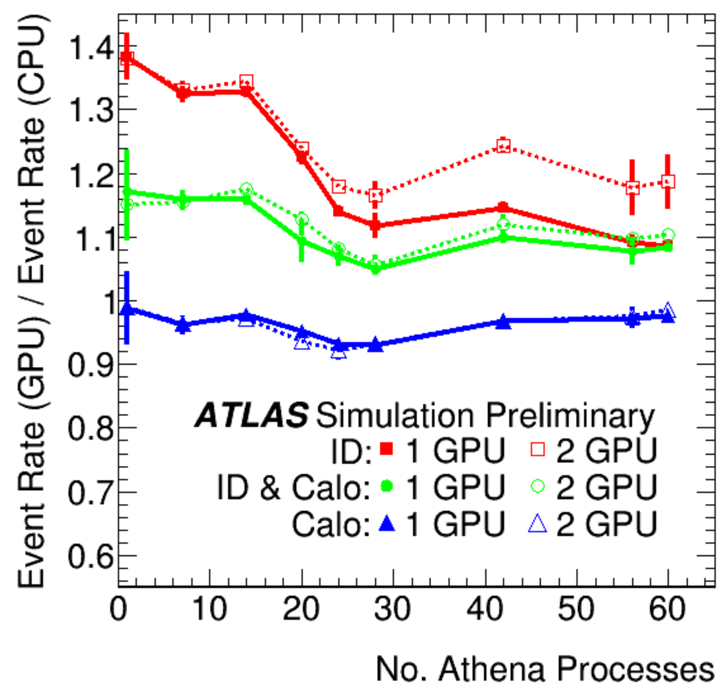
- **Conversion of significant part of ATLAS HLT code to GPU**
  - ported code can run significantly faster than on CPU ×5 for single E5-2695 vs Tesla K80
  - overall speed-up limited to ×1.4
    - data transfer/conversion costs
    - acceleration only applied to part of the workload
  - NB GPU resource barely used (1 GPU per 60 CPUs)

Multi-Threaded Algorithms for GPGPU in the ATLAS High Level Trigger

ATL-DAQ-PROC-2016-045

TriggerSoftwareUpgradePublicResults





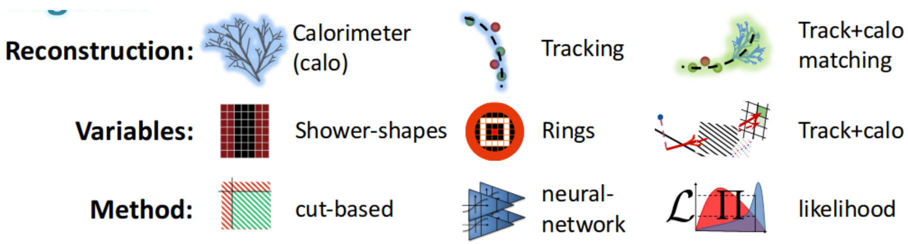
Neural Networks for Online Electron Filtering

Studies done at UFRJ Rio / LPNHE (Werner Spolidoro Freund)

→ see *this talk* by Werner at ILP ML workshop in November 2018

→ see *this poster* by Werner at Saas Fee March 2019

- **Application for High Level Trigger / Fast Calo (electron selection)**
  - Neural Ringer applies ML to reduce CPU demand



  - replace computation of shower shapes
  - concentric rings are build for all calo layers
  - compact cell information used to describe the event throughout of the calorimeter

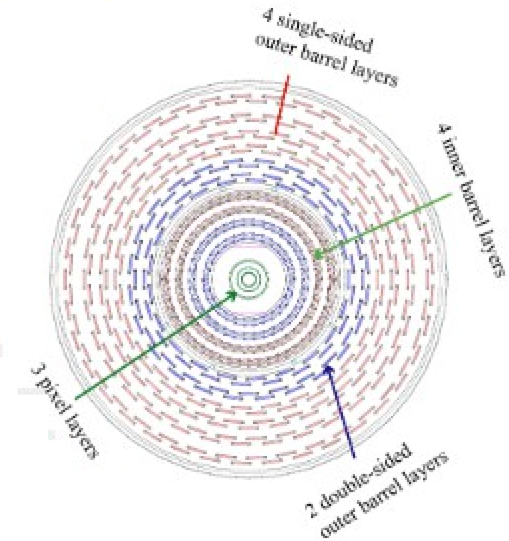
- **MLP training**
  - with simulated (2017 collision) data in 2017 (2018)
  - computing resources from WLCG, Techlab (CERN) and from Advanced High-Performance Computational Center (NACAD) at COPPE/UFRJ

- **Results**
  - kept HLT signal efficiency unchanged after the switch in early 2017:
  - estimated primary chain latency reduction: ~200 ms to ~100 ms;
  - higher rejection power (~2-3X);
  - estimated electron + photon slice: ~1/4 latency reduction;

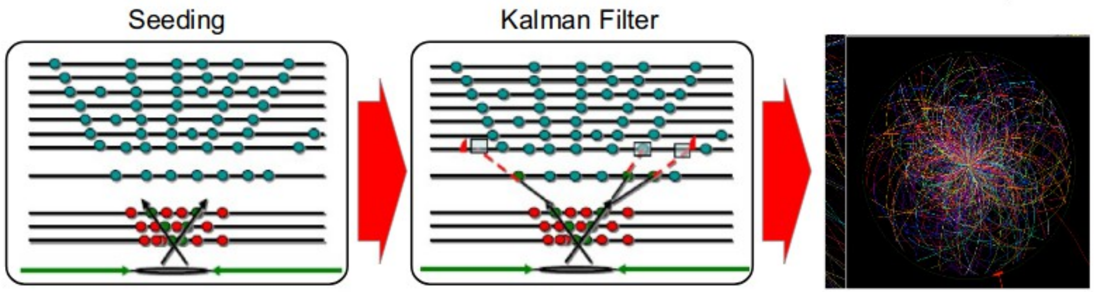
  - 20' to train 1 simple model on GTX 1080ti
  - 100 (initializations)*10 (cross-validation sorts)*36 (phase spaces)=36k tunings

===> 720k' ~1.5 year

Track reconstruction



- **Tracking in a nutshell**
  - particle trajectory bended in a solenoidal magnetic field
  - curvature is a proxy to momentum
  - particle ionizes silicon pixel and strip
  - thousands of sparse hits ; lots of hit pollution from low momentum, secondary particles



- Explosion in hit combinatorics in both seeding and stepping pattern recognition
- Highly computing consuming task in extracting physics content from LHC data

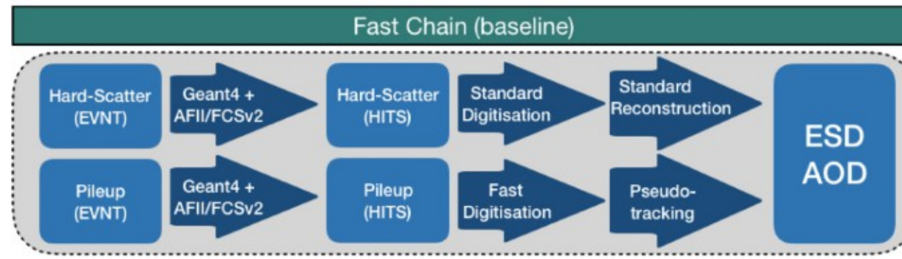
- **Standard solutions**
  - track trigger implementation for trigger upgrades development on-going
  - dedicated hardware is the key to fast computation.
  - not applicable for offline processing unless by adopting heterogeneous hardware.

- **Machine learning**

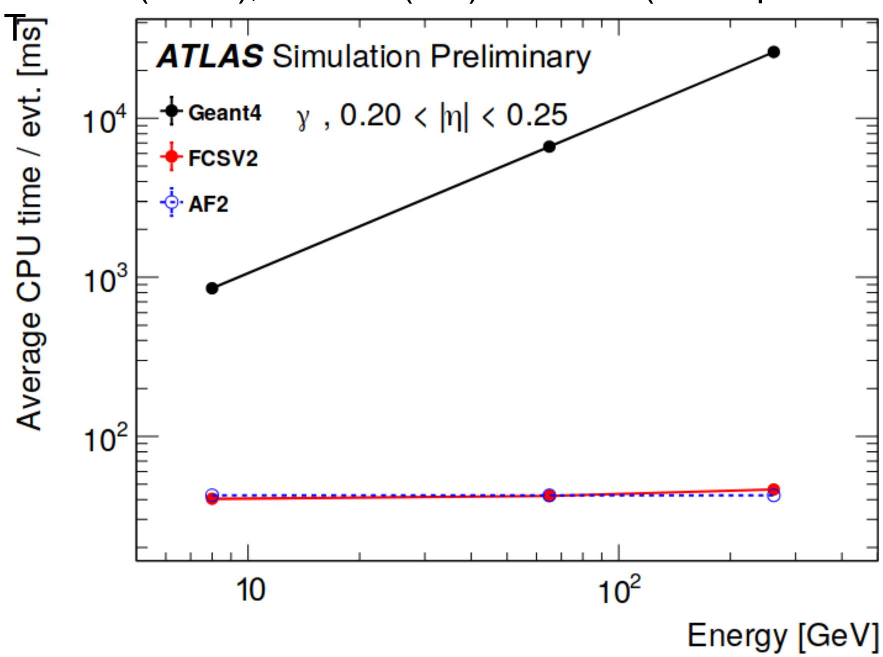
on going *TrackML challenge*
(D. Rousseau et al.)

Simulation evolution : from full sim to fast chain

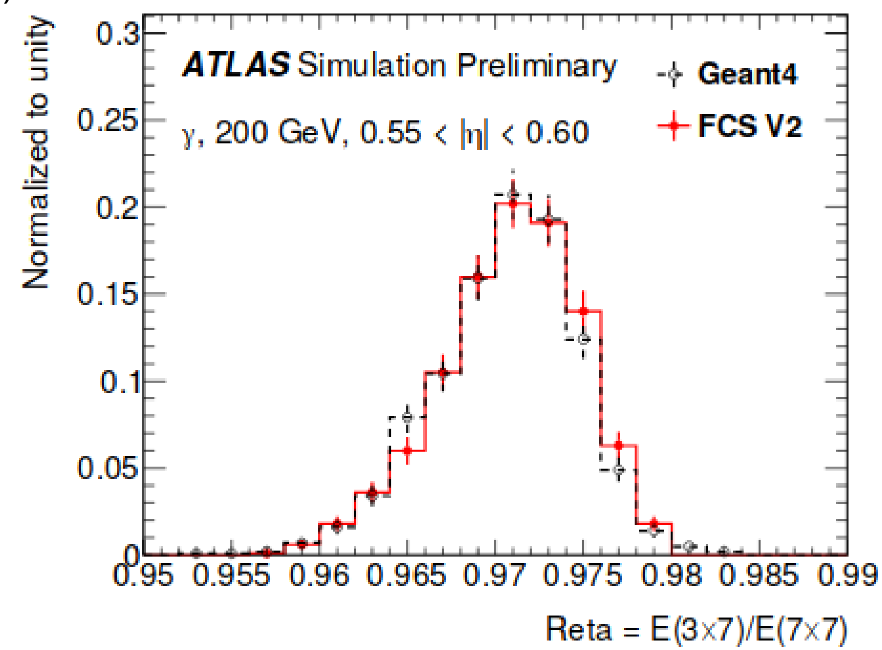
- **At least 1/4 of CPU to be used for Full simulation**
  - tuning and improvement of simulation very important
- **Fast chain as a key ingredient**
  - e.g Fast Calorimeter Simulation
  - validation as « good for physics » is a major challenge



CPU time to simulate photons of 8 GeV, 65 GeV and 256 GeV in the range 0.20<|η|<0.25 using Geant4 (black), FCSV2 (red) and AF2 (blue open circle).



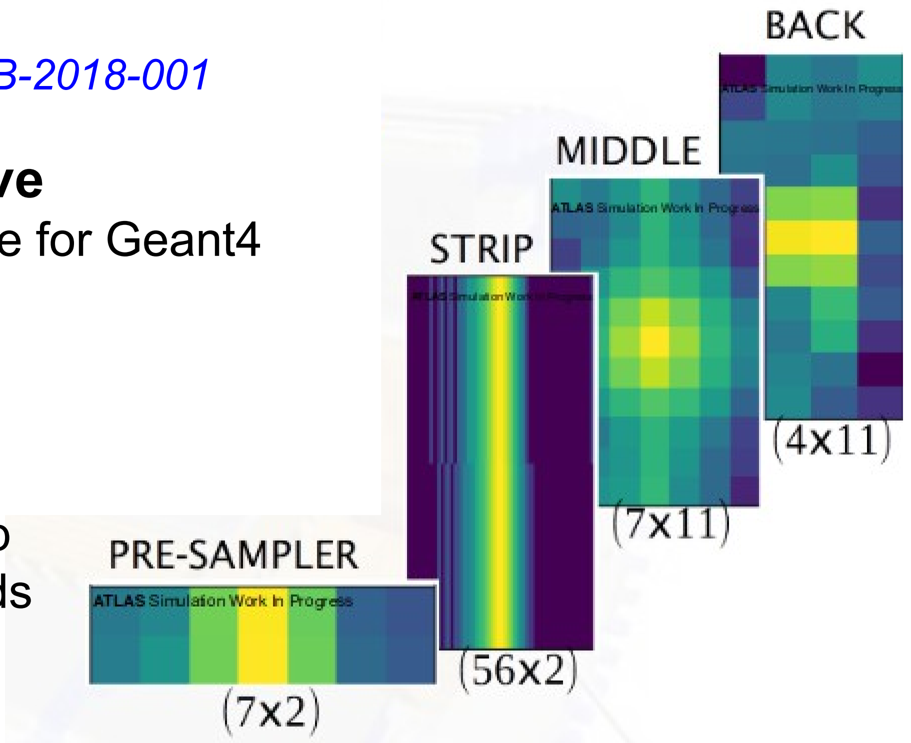
ATL-SOFT-PUB-2018-002



- **new ideas are needed, e.g GAN**

Deep generative models for fast shower simulation

Studies done at LAL (A. Ghosh, D. Rousseau)
→ see *this talk* by Aishik at IN2P3 ML workshop in March 2018
+ *poster* at Saas Fee March 2019
+ details in this note *ATL-SOFT-PUB-2018-001*

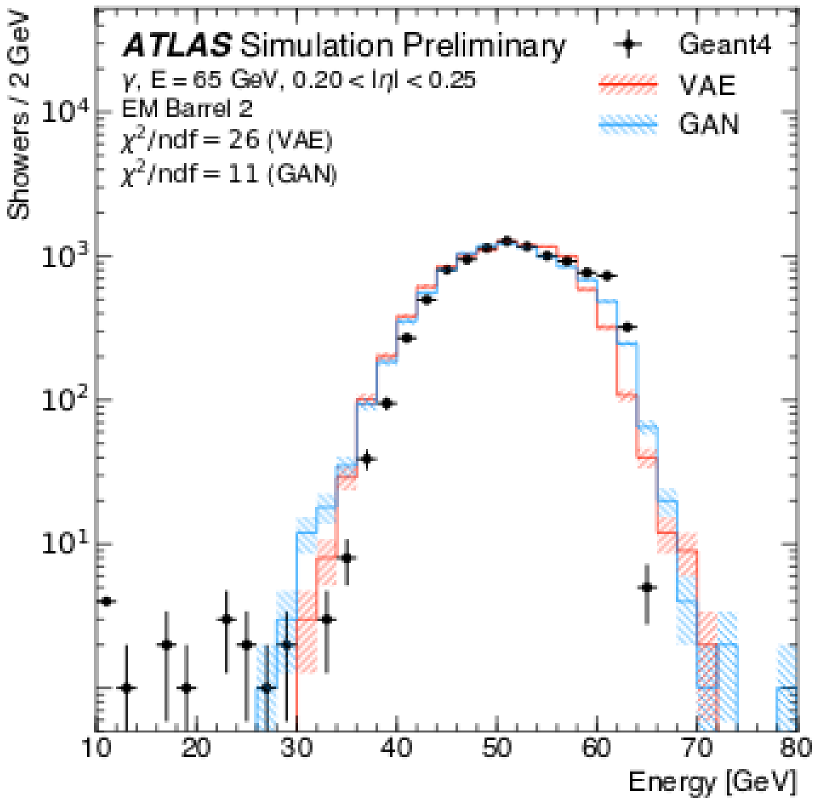


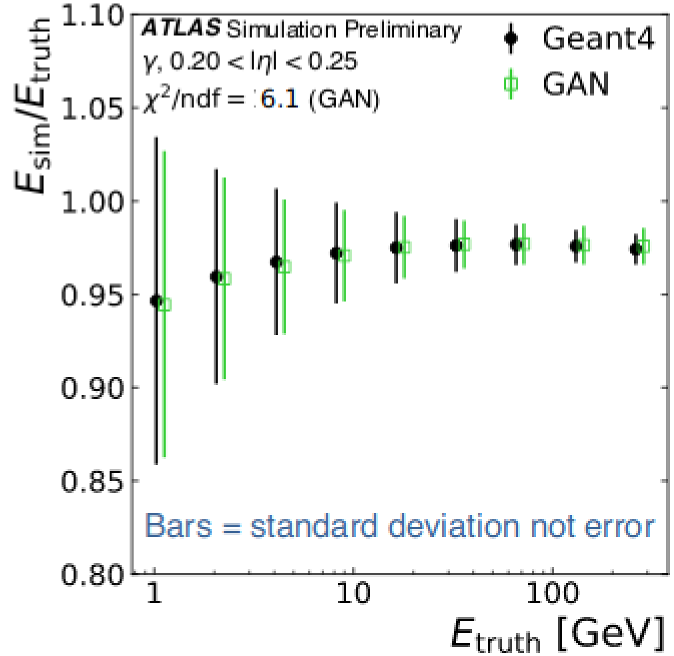
- **Showers computationnally expensive**
  - cascade quantum showers are expensive for Geant4
  - only final image is recorded
- **Compare two methods**
  - Generative Adversarial Networks (GAN)
  - Variational Auto-Encoder combining deep learning with variational Bayesian methods
- **Simulation of images**
  - Train on Geant4 Monte-Carlo simulated single photon shower data
  - Run on 3 GPU platforms (PRP-USA, Texas-Arlington, LLR-Palaiseau, CC-Lyon) for Lyon : 1 GPU per job with >50% GPU utilisation
  - GAN training time: 2 days per training for 15k epochs
  - GPU speed: 2x over CPU for Calo
  - GAN generation time: 0.7ms/shower - as FastCalo images on CPU with Keras+TF

700 jobs * 2 days per job = 16800 hours ~1 year

Deep generative models for fast shower simulation

Look at single photon showers at {1,2,4,8,16, 32, 65, 131, 262} GeV in barrel
Assume Geant4 is ideal. Compare VAE, GAN to Geant4





GAN reproduces the detector resolution mean and σ(E) ~ 10% √E

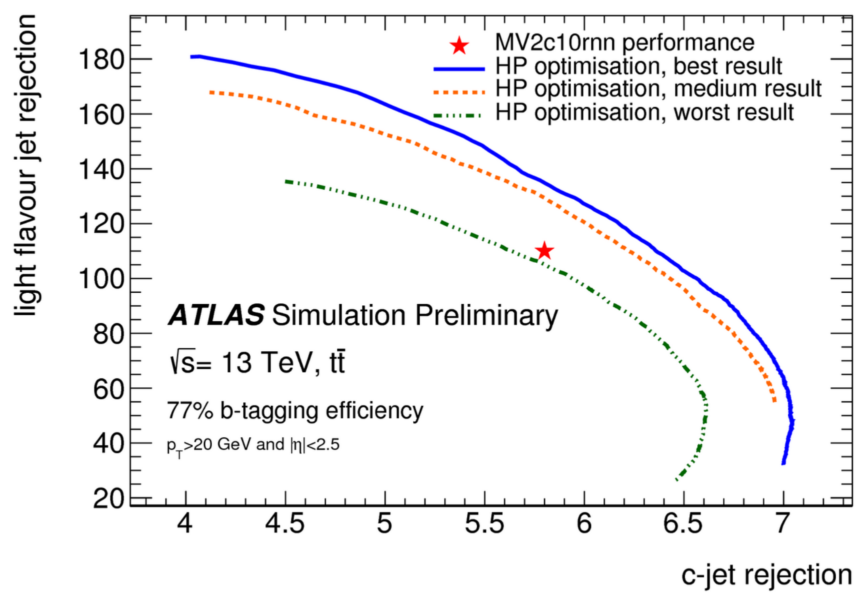
Flavour tagging of jets

- **Flavour tagging**
  - method to tag the origin of a jet of particles from quark hadronization : b-, c-, light quark
  - use specific properties : reconstruction of secondary vertices, soft-muons ...
  - can combine information through BDT, neural network, deep learning etc ...

- **Hyper Parameter (HP) scan**
  - embarrassingly parallel workload and can be split in several independent jobs each running on a GPU
  - optimisation is setup to scan 800 combinations spanning 6 HP dimensions (3 layers, learning rate, batch size and activation functions)
  - the workload has been split in 10 jobs each with 80 combinations. Each job run on the same training and validation data. The input files, small json files containing the configuration for each combination, were replicated to the sites with GPUs using rucio.

- **Results**

FTAG-2019-001
ATL-PHYS-PUB-2017-013



Other studies in ATLAS French groups

- **Increasing TileCal with ML,** LPNHE/UFJR Rio, W. Freund [*link*]
  - increase granularity without changing the mechanical structure of the detector
  - process of acquiring data is very cpu demanding
  - use a multianode 8x8 signals
  - evaluation of of CNN x NMF MLP on original dataset : similar performance;
  - increase stats with GAN
  - results (evaluated CNN only) suggest that a 2x granularity is feasible
    4x in the barrel? To be investigated

- **Discrimination of pile-up jets,** LAPP, P. Zamolodtchikov, N. Berger, E. Sauvan)
  summer internship 2018
  - use of Recursive Neural Network (RNN)
  - using CC-IN2P3 GPU platform for training

- **Analysis ttH(bb) with single lepton,** CPPM (Ziyu Guo, Y. Coadou)
  - BDT used to reconstruct top and Higgs + discriminate signal/background (ttbb)
  - aim to replace BDT by different neural networks
  - use GPU farm at computing department of Uiversity of Aix-Marseille
    → analysis time: 50890717 secondes
    (1 year 224 days 18 minutes 37 seconds)

Conclusion

- **General ATLAS usage of HPC and GPU**
  - usage of HPC resources is already a reality
    some site(s) even provide GPU farm on WLCG
  - will become crucial for HL-LHC
  - various usages already exist for GPU : trigger, simulation, ML

- **ATLAS usage by French groups**
  - many different use cases linked to Machine Learning
  - studies are done on different platforms available
    → all groups have expressed (future) interest
    in using the GPU farm at CC-IN2P3
    → need to collect all use cases and turn
    it into an official time request