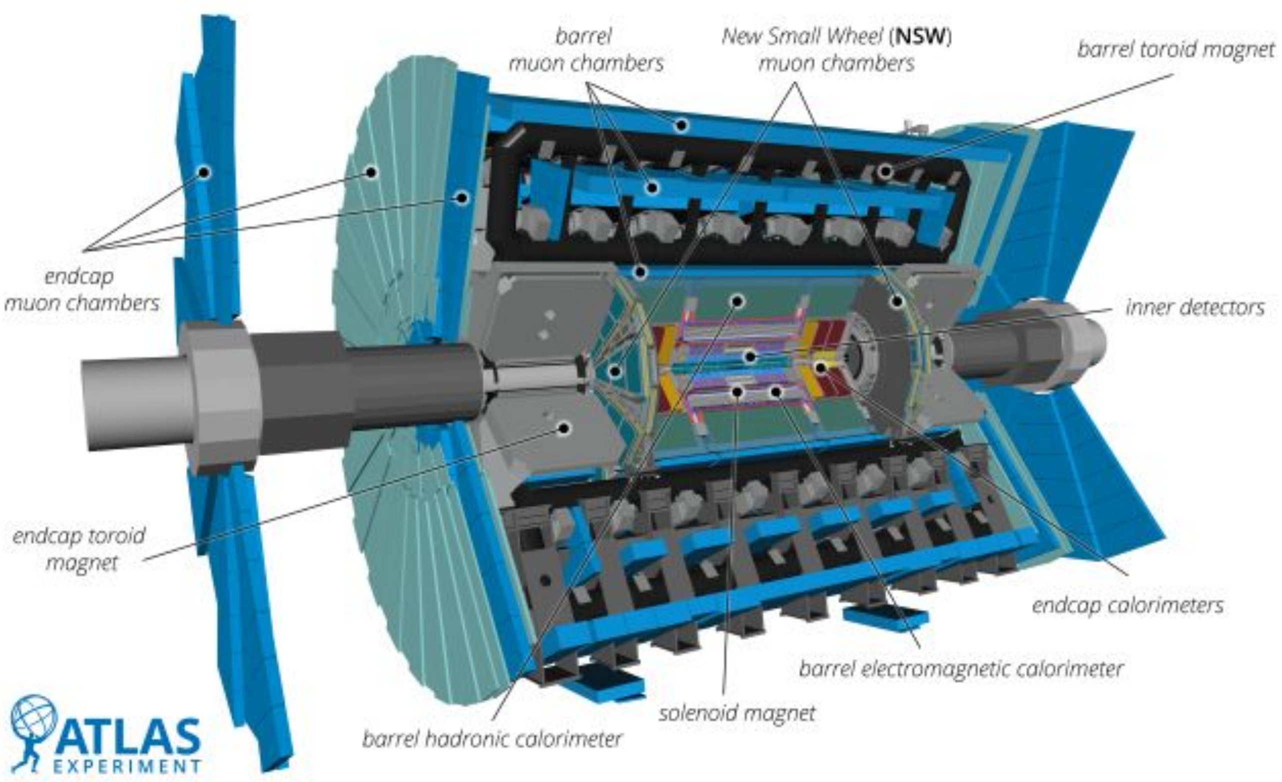
Optimizing the ATLAS Geant4 detector simulation

CHEP 2024, 21st October 2024

Ben Morgan (**The University of Warwick**) on behalf of the ATLAS Collaboration

Simulation of the ATLAS Experiment

- Simulation of interactions within the ATLAS Experiment at the LHC is performed using **Geant4** integrated into the ATLAS Offline Software framework (**Athena**).
- In Run 2 ATLAS used Geant4 10.1 (**mc20**).
- In Run 3 ATLAS uses Geant4 10.6 (**mc23**).



Increasing technical performance in MC21-23

- **A:** Avoid simulating of uninteresting particles:
  - *Neutron/Photon Russian Roulette (**Speed-up: ~10%**)*
  - *Increase secondary production threshold (range cut) for Gamma processes **(Speed-up: 6-10%)***
- **B:** Simulate interesting particles more efficiently:
  - *Use of VecGeom Tube/Cone/Polycone **(Speed-up: 2-7%)***
  - *EM Endcap Calorimeter (EMEC) Geometry Hierarchy Optimisation (**Speed-up: 5-6%)***
  - *Tailored magnetic field switch-off in LAr Calorimeters (**Speed-up: 3%)***
  - *G4GammaGeneralProcess to reduce number of MFP evaluations **(Speed-up: 3%)***
  - *Big/Static Library for Simulation to reduce PLT/GOT instructions **(Speed-up: 5-7%)***
  - *This talk covers "**Full**" Simulation workflow, but ATLAS also has a "**Fast**" Simulation workflow*
    - ***Advancements in the ATLAS Fast Chain for HL-LHC: Towards Efficient MC Production***
    - ***AtlFast3: Fast Simulation in ATLAS for LHC Run 3 and beyond***
- **In this latter category, highlight two topics today:**
  - ***Woodcock Tracking in the EMEC***
  - ***Link Time Optimization of the Big Library***

Woodcock Tracking in the EMEC

ATLAS EXPERIMENT

- Woodcock Tracking is a technique for highly segmented detectors where geometric boundaries rather than physical interaction lengths limit the simulation steps.
  - *Is the case for photons in ATLAS EMEC region: transportation process dominates*
- Photons don't interact *along* a step (no continuous energy loss), thus safe to perform tracking of photons in a simplified EMEC geometry (i.e. without boundaries) made of the densest material from the standard EMEC geometry (Pb).
  - ***Interaction then occurs with probability equal to ratio of cross-sections of the true material and Pb.***
  - ***Only statistical changes in output.***
- In ATLAS, Woodcock Tracking is applied on top of the G4GammaGeneralProcess.
  - *Added as an ATLAS patch on top of Geant4 10.6.patch03*
- **Speed-up: 17.5%**
  - *Via 50% reduction in number of steps for photons in EMEC.*

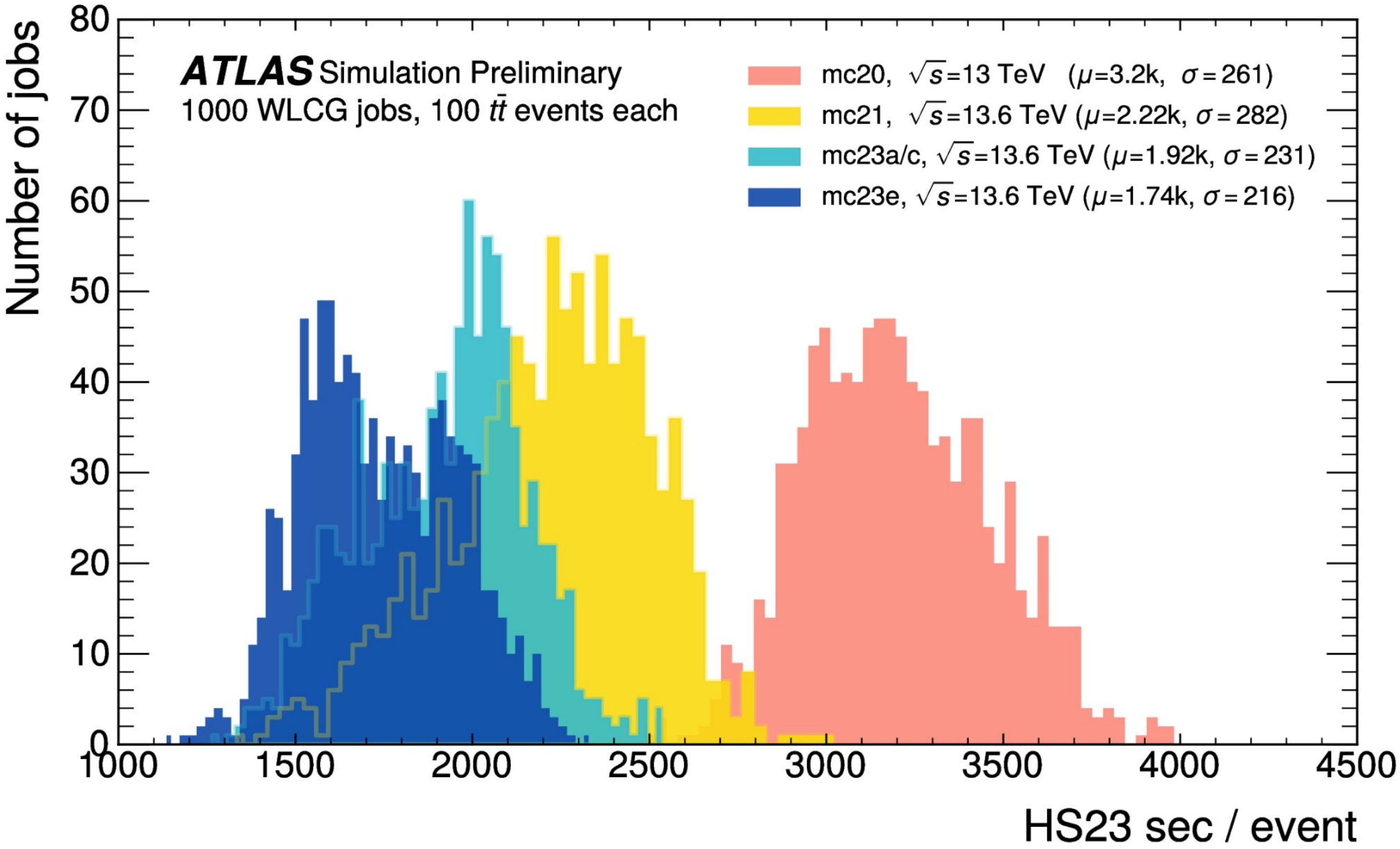
Link-Time (Interprocedural) Optimisation

- Set of methods to optimize **across**, rather than **within**, translation units
  - *GCC/Clang: extra data attached to object files to assist linker view "whole program"*
- AtlasGeant4 "Big Library" is a **shared** library, but links in **static** Geant4 libraries and **does not export their symbols**
  - *LTO not generally applicable to shared libraries given reuse/runtime loading*
  - *Here, non-export of symbols means linker can treat these like a "program"*
- Only requires appropriate compiler/linker flags, enabled in Athena via use of CMake **`CMAKE_INTERPROCEDURAL_OPTIMIZATION`** option

- **Speed-up: ~5%**
  - *Simulation output bitwise identical after this technical change, as expected.*

Simulation runtimes on the WLCG

- Throughput increased by **x1.84** between mc20 and mc23e campaigns!
  - *mc20 uses Geant4 10.1*
  - *mc21 & mc23x use Geant4 10.6*



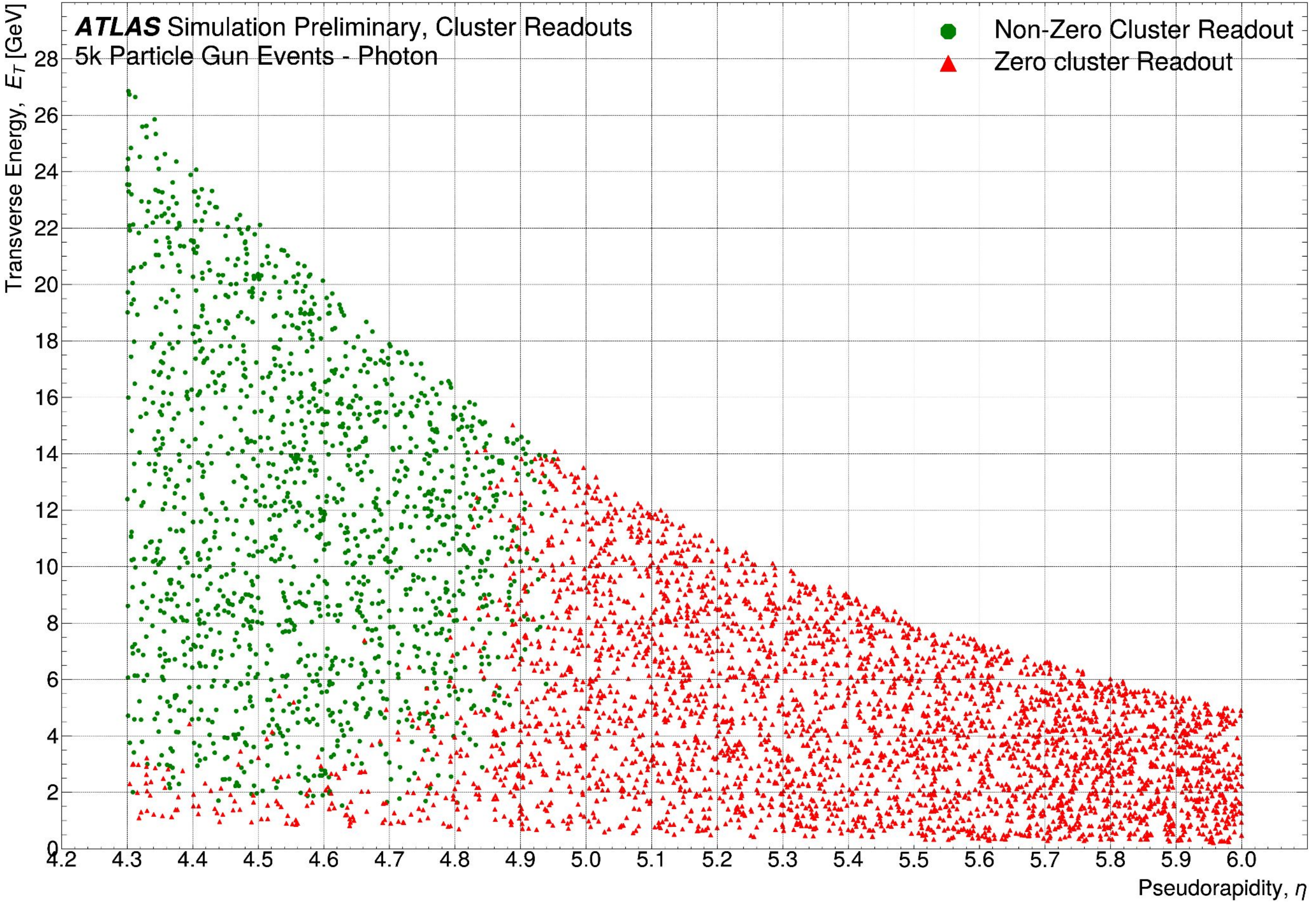
https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/SIMU-2023-06/

Future optimizations under development

- Several areas building on existing optimizations under development:
  - *Switching off EM Energy Loss fluctuations*
    - *Physics output-changing. Postponed until next major MC campaign.*
  - *Parameter Tuning of In-Field Tracking*
    - *Customize G4 tracking parameters based on particle type, properties and location region to optimize CPU performance without compromising precision.*
  - *G4NeutronGeneralProcess*
    - *Super-process for neutron physics (c.f. G4GammaGeneralProcess)*
  - Advanced Compiler Methods - PGO / AutoFDO
    - *Next step after LTO - use profile driven feedback to further optimize the big library.*
- **Today, we'll highlight five topics:**
  - ***High-η particle rejection***
  - ***Re-implementation of EMEC geometry***
  - ***Use of AF3 for low energy particles in the EMEC***
  - ***Streamlining adoption of new Geant4 versions***
  - ***Specialized transport per particle type with G4HepEM/AdePT/Celeritas***

High-η particle rejection

- **Idea:** Kill primary particles generating secondaries close to the beam-pipe at ~5-6 m from the IP.
- Many particles in the collision are at high $|\eta|$ (no Inner Detector hits) with little energy compared to the calorimeter noise.
- Check if we can kill some particles early on which will have no or little effect on the simulated energy in the calorimeters to save CPU.
- **Cutting η > 5.0 and E_T < 0.5 GeV looks promising.**

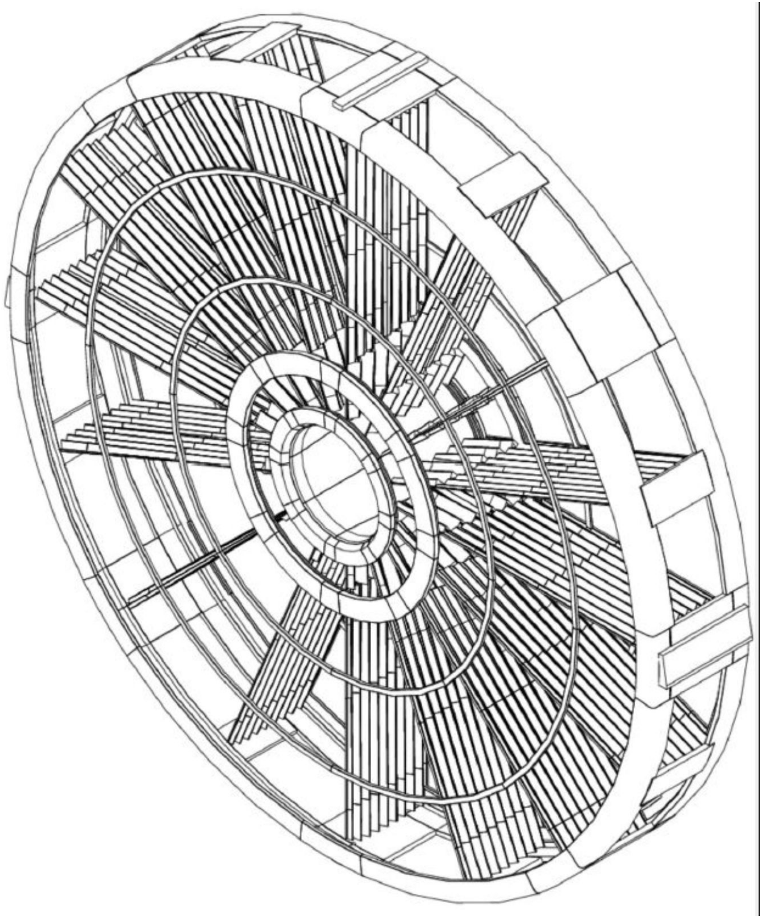


https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/SIMU-2024-05/

Re-implementation of the EMEC geometry

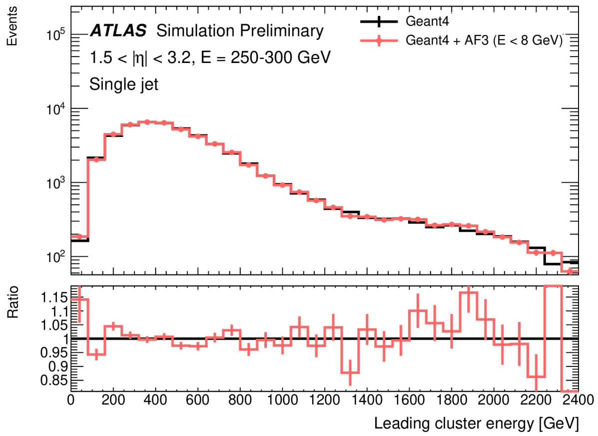
ATLAS EXPERIMENT

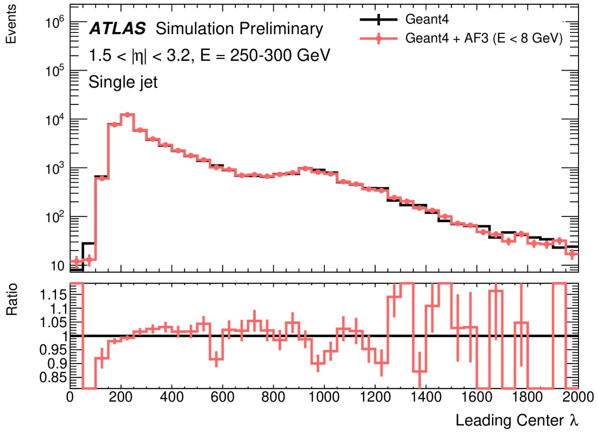
- Background:
  - *The as-built EMEC has a complicated "Spanish Fan" geometry.*
  - *Efficient description using the G4Solids available in early versions of Geant4 was not possible*
  - *Custom solids used to implement the geometry*
- New implementation using data from the technical drawings
  - *EMEC mother volume*
  - *Inner and Outer Wheel envelopes*
  - *Inner and Outer Wheel Absorbers and Electrodes defined as G4GenericTrap solids*
  - *Option to subdivide the geometry into slices to aid voxelization.*
- Code now available in Athena
  - *Preliminary benchmarks indicate ~19% speedup*
  - *Work underway to validate physics output*



Using AtlFast3 for low energy e/γ in the EMEC

- Even after MC23 optimizations, e/*γ* in EMEC one of the biggest contributors to the total steps-per-event.
- **Idea**: use AtlFast3 to handle low energy e/*γ* in the EMEC for the Full Simulation workflow
  - *Similar to [Frozen Showers in ATLAS FCAL](#) for low energy e/γ/neutrons*
  - *Find E/η region where outputs do not differ w.r.t. full Geant4*
- Preliminary studies show promise for particles with *E<8GeV*, *1.5<η<2.5*, with ~10% speed-up in this region compared to full Geant4.





Streamlining adoption of new Geant4 versions

- Using a new Geant4 version is a physics modelling update
  - *These are expensive in people/compute resource, typically once/twice per LHC Run*
- Nevertheless, we want to test new versions in Athena development:
  - *Be ready to take advantage of latest improvements to physics and performance*
  - *Quicker reporting of ATLAS-specific issues to Geant4*
  - *Minimize diff(s) between latest/production versions to ease understanding of physics modelling differences if observed*
- Work underway to automate more of this process
  - *Use Athena CI to regularly build against given Geant4 version(s), as done for Gaudi/ACTS*
  - *Run regular high level validations of these builds as early warning of discrepancies to help understanding before committing to a full physics modelling validation*
  - *Determination of sampling fractions, Birks corrections*

Specialized transport: G4HepEM/AdePT/Celeritas

ATLAS EXPERIMENT

- Geant4 11 added the capability to customize transport per particle type
  - *Choose what to do (stepping, parameterization) based on particle energy, location*
  - *Increased flexibility, coherence, and potentially performance, compared to other methods/hooks*
- G4HepEM
  - *A compact Geant4 EM library with memory layout/algorithms optimized for the HEP EM shower development and e-/e+/γ transport use case, validated against the more general Geant4 EM models.*
- AdePT, Celeritas
  - *Implement Action-based full stepping transport loop for e-/e+/γ on NVIDIA(AMD) GPUs, targeting HEP use case for geometry and physics processes*
    - *See presentations in next parallel session in this track*
  - *Integration in Athena underway using specialized transport to offload of e-/e+/γ in, say, calorimeters, to the GPU asynchronously whilst retaining host-side event boundaries*
  - ***AdePT/Celeritas allow use of host-side Sensitive Detectors for scoring, so output hits can be compared directly in downstream reconstruction/validation pipelines***
  - *Runtime performance to be evaluated in range of Host-Device systems and realistic production workflows*

Summary

- The ATLAS Simulation group aims to:
  - *provide Monte Carlo samples with consistent physics for entire LHC runs for analysers*
  - *include new optimizations without changing physics modelling between yearly sub-campaigns.*
  - *include physics modelling improvements between campaigns.*
- Multiple optimizations were introduced between the Run 2 (mc20) campaign and the latest Run 3 campaign (mc23e - matching 2024 data)
  - ***Throughput increasing by a factor of 1.84!***
- Healthy programme of on-going development of further optimizations for the future both from adopting new Geant4 features, improving Athena code, and exploring use of GPUs for stepping loops
  - *Close collaboration with the Geant4 Collaboration continues to be key here*
- Further performance improvements expected for the sub-campaign for 2025 data.

Additional Information

WARWICK
THE UNIVERSITY OF WARWICK

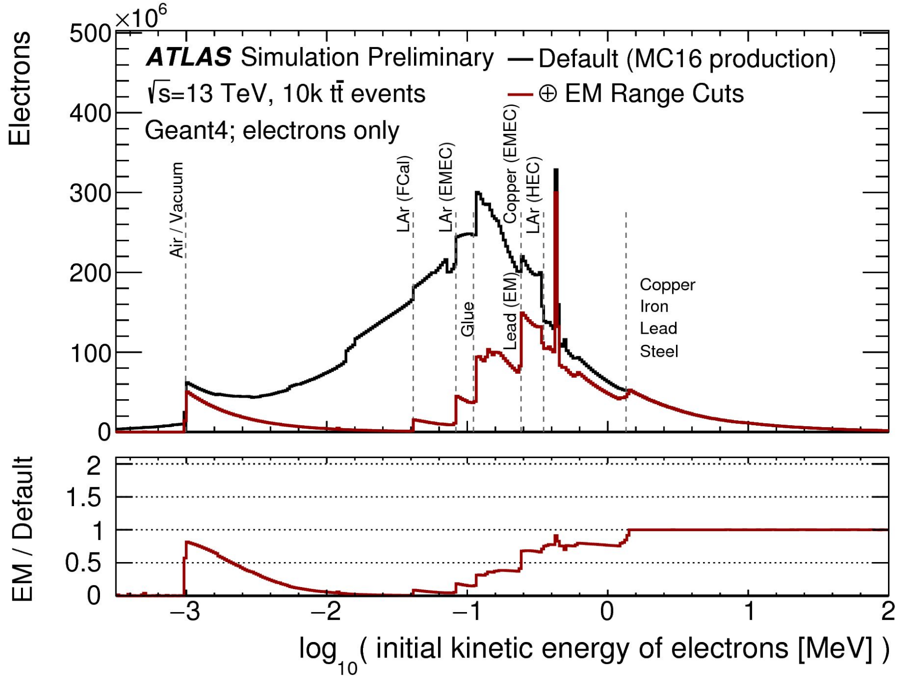
ATLAS EXPERIMENT

Neutron/Photon Russian Roulette

- Neutron Russian Roulette
  - *Low energy neutrons take quite some CPU time in simulation, usually with many steps that are not really correlated with the point of their creation.*
  - *Randomly kill neutrons below some threshold energy with a probability 1/w and apply a corresponding weight (w) to the remaining neutrons. The remaining neutrons would then deposit w-times the energy.*
    - *Based on an idea already used in CMS Simulations.*
  - *Parameters used in production: Threshold = 2.0 MeV, w = 10.*
- Photon Russian Roulette:
  - *Applied to photons produced in the LAr EM Calorimeters in ATLAS*
    - *Avoids unwanted effects in the Inner Detector.*
  - *Parameters used in production: Threshold = 0.5 MeV, w = 10.*
- Both Implemented in Athena in a configurable G4UserStackingAction.
- **Speed-up: ~10% overall**

EM Range Cuts

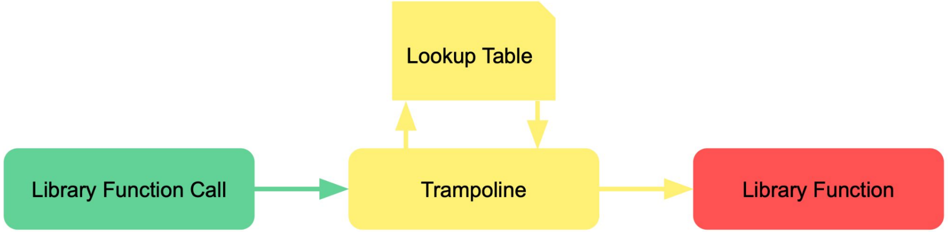
- Range cuts are a built-in way of increasing Geant4 performance
  - *Secondary particles that are expected to travel less than the range cut are not created and their energy is immediately deposited by the parent particle.*
  - *For each material-volume pair, range cuts can be specified in distance units (mm).*
- By default Geant4 does not apply range cuts for the conversion, photo-electric or Compton-scattering gamma processes.
  - *Option provided by Geant4 to activate range cuts for these processes:*
  - *G4 UI command: '/process/em/applyCuts true'*
- Range cut of 0.1 mm used, same as for electron processes.
- **Speed-up: 6-10%**
  - *Due to far fewer being particles created.*



https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/SIM-2019-001/

Big/Static Library for simulation

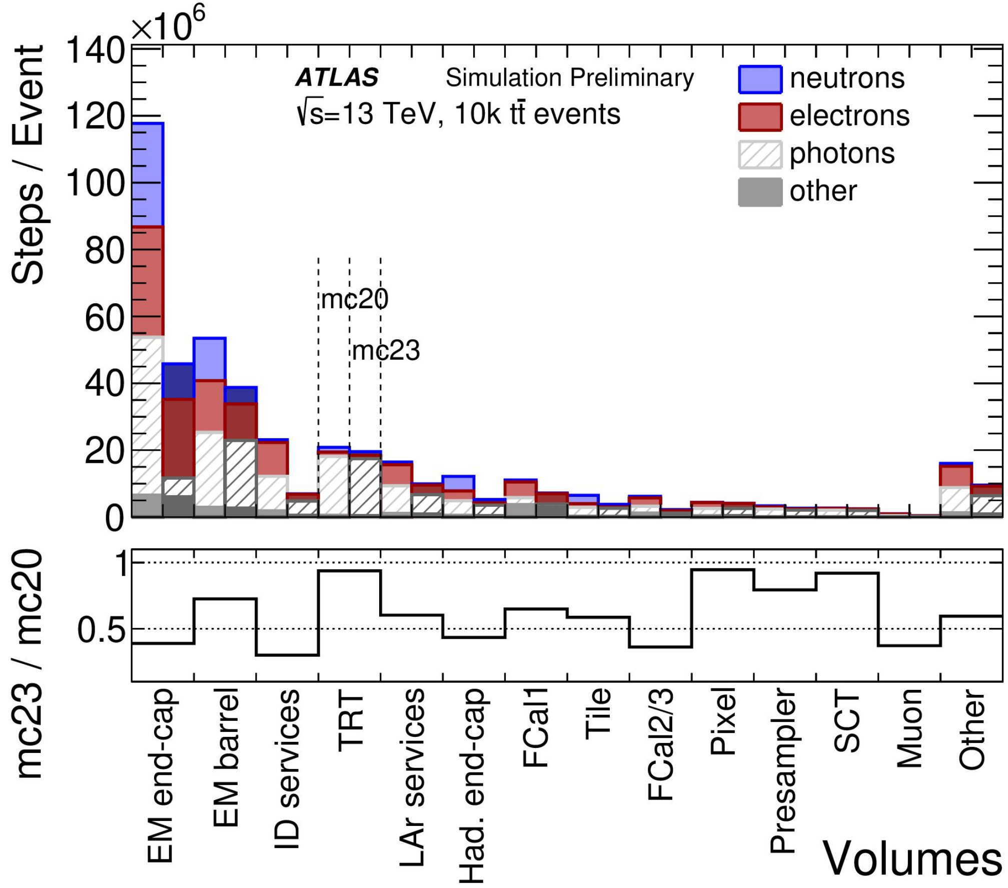
- Function calls in a **shared** library go through the Procedure Linkage Table (PLT), leading to extra instructions per call (so called “trampolines”):



- Remove some calls by **statically** linking Geant4 into a **shared** Athena library
  - *Compile Geant4 external in static mode with PIC (e.g. create .a archives).*
  - *Create a single “big” **shared** library in Athena linked together from **all packages linking to Geant4** (single point of linkage) plus the Geant4 **static** libraries*
  - *Use `--exclude-libs,ALL` linker flag to not export symbols from statically linked libraries, removing trampolines as these are now internal only to the “big” library*
- **Speed-up: 5-7%**
  - *Simulation output bitwise identical after this technical change, as expected.*

CPU Time: Steps per event per subdetector

ATLAS EXPERIMENT

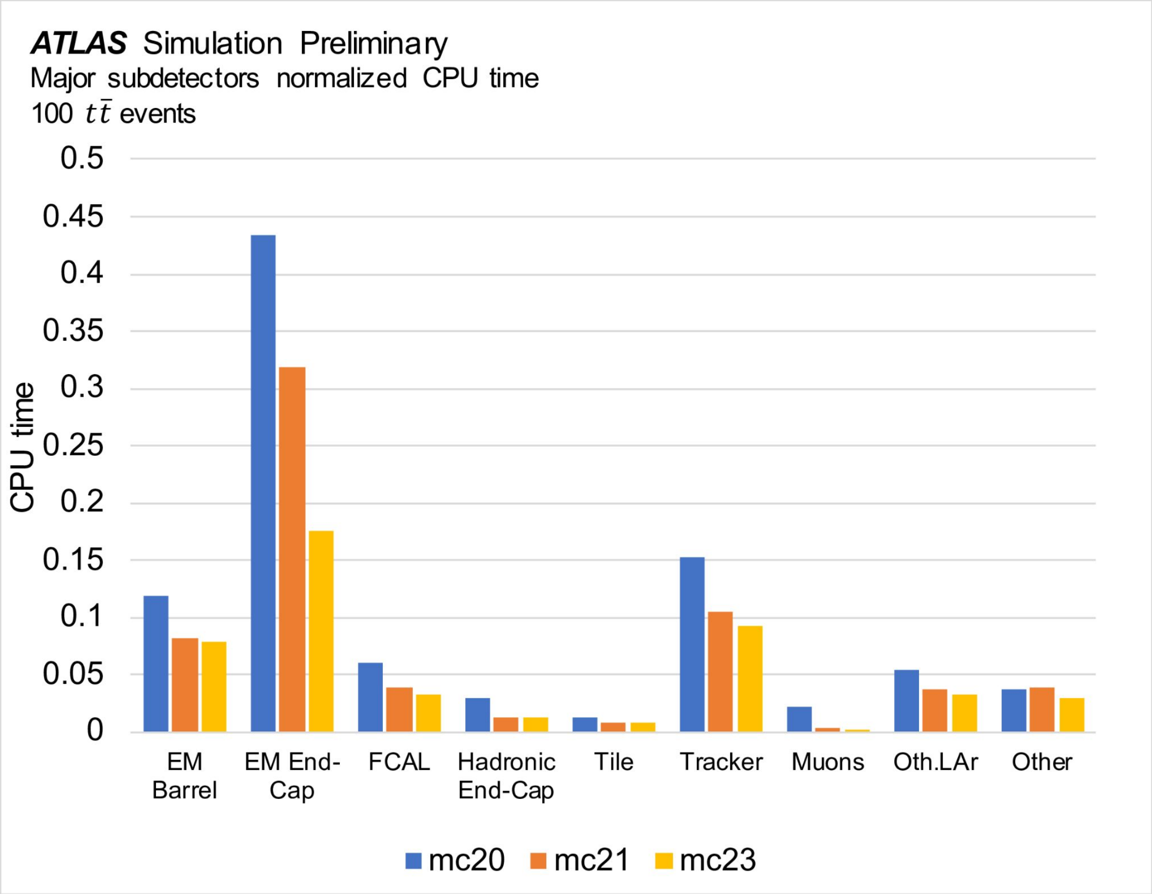


- Number of Geant4 steps per event for various ATLAS detector volumes. The left column in each section represents the Run 2 (mc20) setup and the right column represents the setup during Run 3 (mc23 = mc23c).
- 'FCal1' includes the first (electromagnetic) module of the forward calorimeter and 'FCal2/3' includes the subsequent two hadronic modules.
- 'ID services' includes ID services and the beam pipe.
- 'LAr services' includes LAr services and LAr cryostats.
- 'Other' includes all other particles and all other volumes that are simulated.

https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/SIM-2024-003/

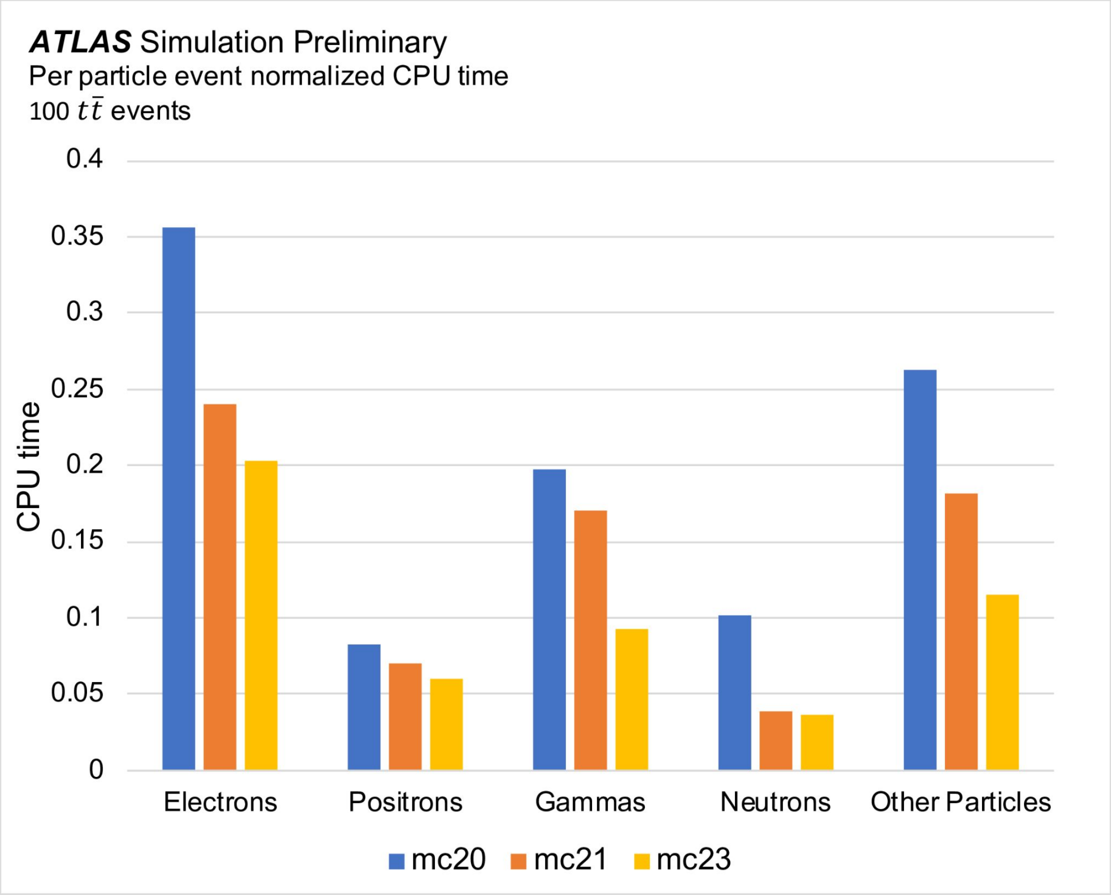
CPU Time: Subdetector fractions per event

ATLAS EXPERIMENT



- Time spent per event simulating 100 $t\bar{t}$ events, normalized to the total time spent to simulate events in mc20, for each of the major subdetectors.
- The different coloured bars represent different simulation configurations: the left most bar, for each subdetector, shows the simulation time for the Run 2 (mc20) configuration; the middle bar displays the time for an optimization used for the first Run 3 simulated samples (mc21); while the right bar shows the time spent on simulating events with the latest Run 3 configuration (mc23 = mc23c).

CPU Time: Particle type fractions per event



- Time spent per event simulating 100 $t\bar{t}$ events, normalized to the total time spent to simulate events in mc20, for each particle type.
- The different coloured bars represent different simulation configurations: the left most bar, for each subdetector, shows the simulation time for the Run 2 (mc20) configuration; the middle bar displays the time for an optimization used for the first Run 3 simulated samples (mc21); while the right bar shows the time spent on simulating events with the latest Run 3 configuration (mc23 = mc23c).

https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/SIM-2024-003/