Computing Performance Results and Issues: ATLAS

29th Geant4 Collaboration Meeting,
Catania
07-11 October 2024

Marilena Bandieramonte (University of Pittsburgh)
on behalf of the ATLAS Simulation Group

Geant4 Simulation in ATLAS

- For the last RUN3 Monte Carlo production (mc23e) ATLAS used:
  - Geant4 10.6.patch03.atlas13
    - It includes (new patches only):
      - c++20, gcc13 support
      - Make Woodcock tracking more robust against corner cases
      - Fix for some rare crashes in hadronic physics (`G4AntiNeutron` by default unstable)
      - Geant4 LTO-enabled
  - Physics lists: FTFP_BERT_ATL
  - Stepper: ATLASRK4
- Testing now Geant4 11.2.patch02
  - New project on streamlining the adoption of new Geant4 versions

Overview of Run3 MC Campaigns

***Run3 MC campaigns, mc23*:**

- **mc23a,** for 2022 data (supersedes mc21)
- **mc23d,** for 2023 data (supersedes mc23c)
- **mc23e,** for 2024 data

Optimizations put in production at the start of Run3 (mc21)

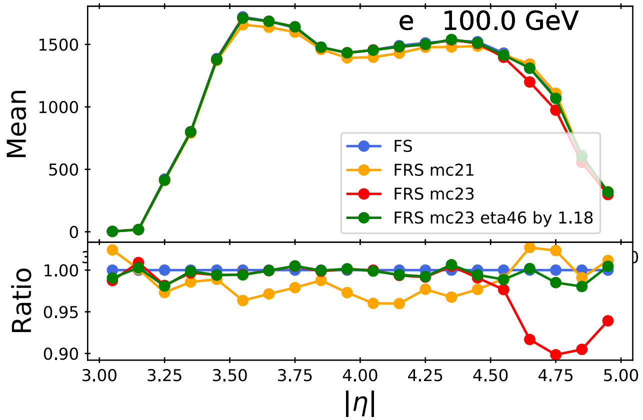
- **EM range cuts**
- **Neutron** and **Photon Russian Roulette**
- New **EMEC Slices** variant
- **BigLibrary**
- **G4GammaGeneralProcess**
- **Magnetic Field** tailored **switch-off**

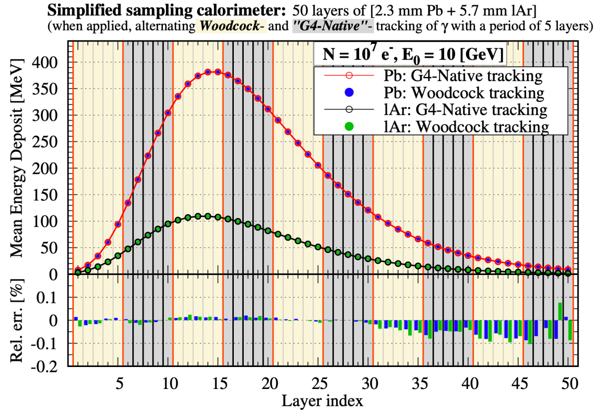
Optimizations added in mc23a/d:

- **VecGeom**: new vectorized geometry library, we replace only *polycons, cones and tubes (2-3%)*
- **Frozen Showers** update (tuned to the new Geant4 version)
- **Woodcock tracking** [1]: smart tracking for highly segmented detectors (~17.5%)

New Features:

- **Link Time Optimization (LTO):** Expands scope of inter-procedural optimization to encompass all objects visible at (final executable) link time => **5% speedup**
- Switch-off energy loss fluctuation in Geant4 (on hold)





Geant4 Simulation Optimizations for Run3

New Improvements that are used in production in the Run3 MC campaigns*

**since last year presentation, details in backup slides*

Link Time Optimization

Idea: test advanced compiler optimizations, that can lead to non-negligible speed-up factors

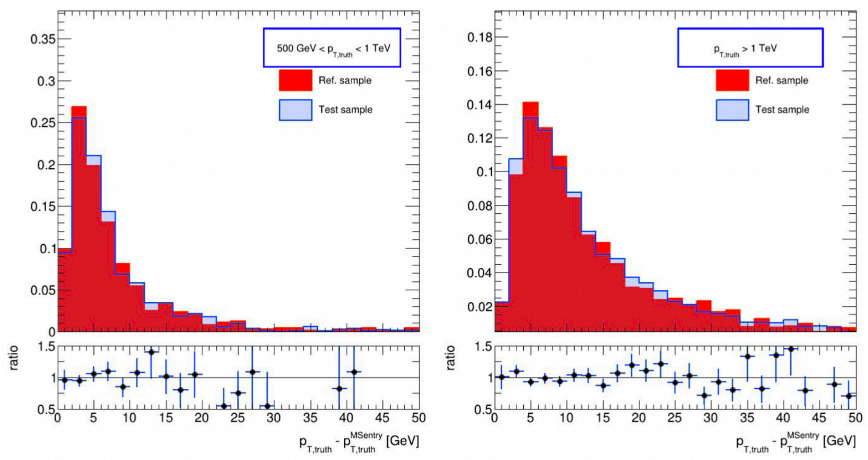
- CMS reported ≈ 10% speedup in their software with (LTO+PGO)
- LTO extends the scope of inter-procedural optimizations to include all objects accessible during the linking process.
- Traditionally, after compiling files separately into object files, the compiler merges them into a single executable during the linking stage.
- However, with LTO, the compiler can store its intermediate representation, allowing the various compilation units that contribute to the executable to be optimized collectively as a single module during the final linking phase
- Since Athena code with Geant4 dependencies is statically linked into one shared-object library, LTO can be applied to this library (instead of the executable).
- Implementing this required only modifications to the CMake configuration of the Athena build process.
- **Status**:
  - Put in production for the 2024 data taking MC campaign (mc23e): **5% speed-up**

Disabling Energy Loss Fluctuation in Geant4

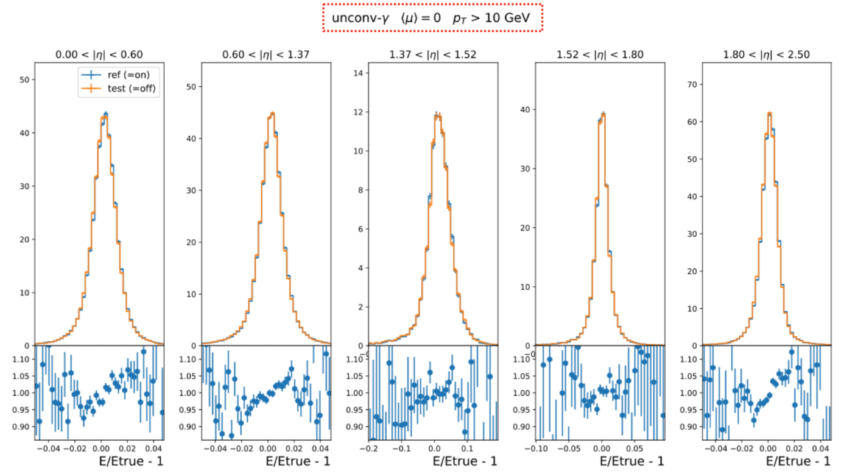
On hold

- ***Idea***: Switch-off energy loss fluctuation in Geant4 (**measured 3-4% speedup**)
- In Geant4 energy loss process calculates continuous and discrete losses **per step**
  - Both are stochastic processes, mean value of **continuous loss** can be calculated from (restricted) stopping power.
  - *Statistical fluctuation around this mean from Urban/PAI models in Geant4*
- Sampling these models adds (small) compute time per step, but fluctuation can often be ignored **if number of steps is large** in important volumes
- Statistical variation in energy loss over steps dominates per-step fluctuation

Muons with $p_T > 500$ GeV



Unconverted Photons



- When switching this process OFF ATLAS observed Changes on Egamma energy scale larger than calibration systematic and a higher catastrophic energy loss, with a 1% mismatch for muons with pT>500GeV
  - **NOT included in mc23e campaign**, but will wait for Geant4 update to enable the switch-off option depending on particle types / detector regions.

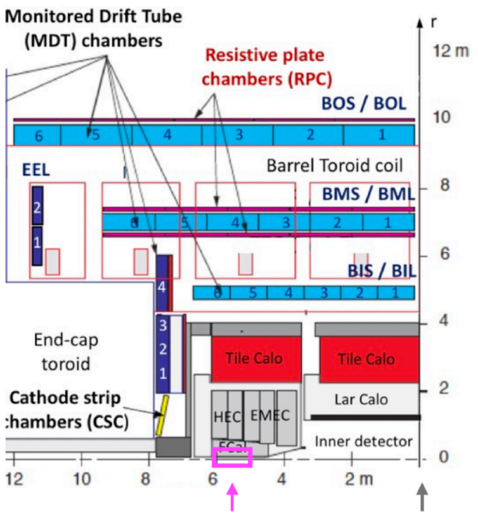
Ongoing effort

Targeting Run3 MC reprocessing, or future Run3 MC campaigns

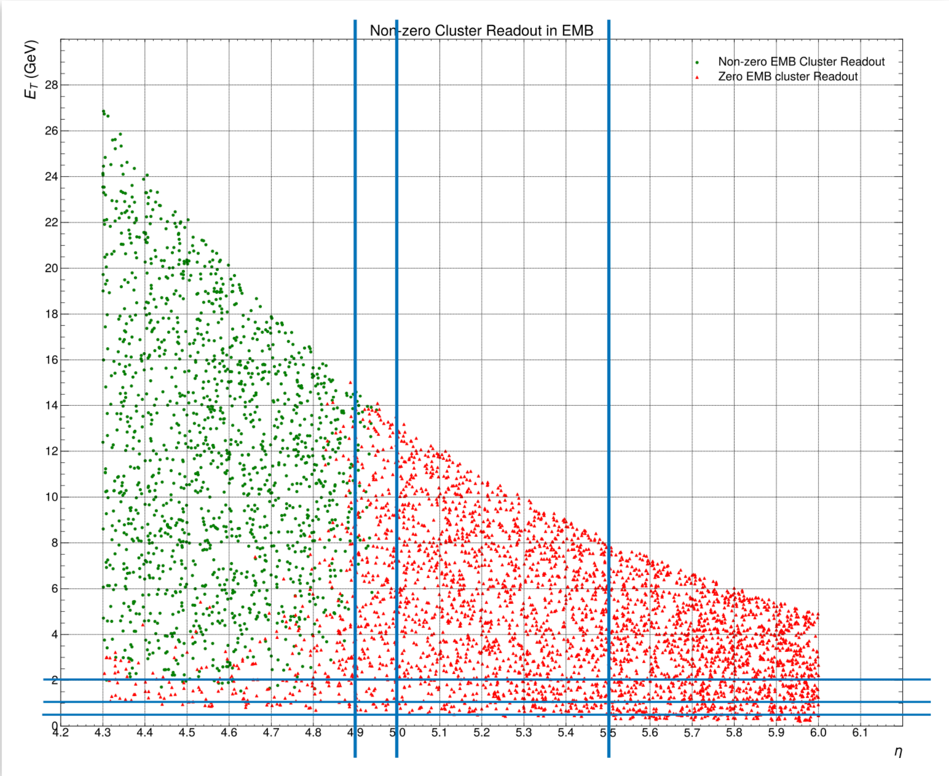
High-η particle rejection (ISF Particle Killer)

COMING SOON

- ***Idea***: Kill primary particles generating secondaries close to the beam-pipe at ~ 5-6 m
- Many particles in the collision are at high $|\eta|$ (no ID hits) with little energy compared to the calorimeter noise.
- Check if we can kill some particles early on with no or little effect on the simulated energy in the calorimeters to save CPU.
- Test sample:
    - 5000 events produced with PG (photon)
    - Energy (E): 1 MeV ~ 1 TeV
    - $\eta \in [4.3, 6]$



Reference [plots]



- Cutting $\eta > 5.0$ and $E_T < 0.5$ GeV looks promising.
- Further physics precision check is on-going before physics validation.

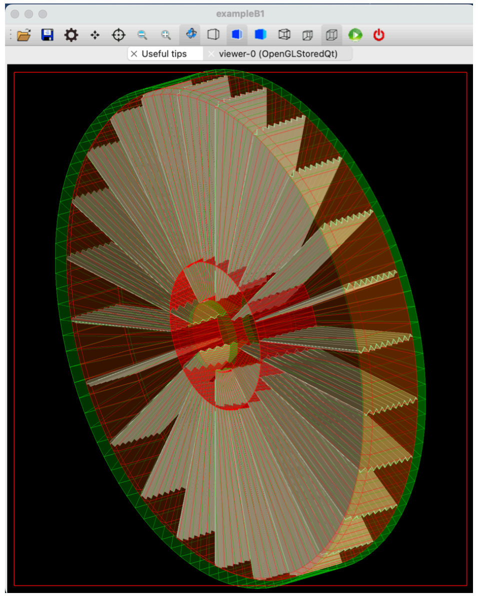
ATLAS EXPERIMENT

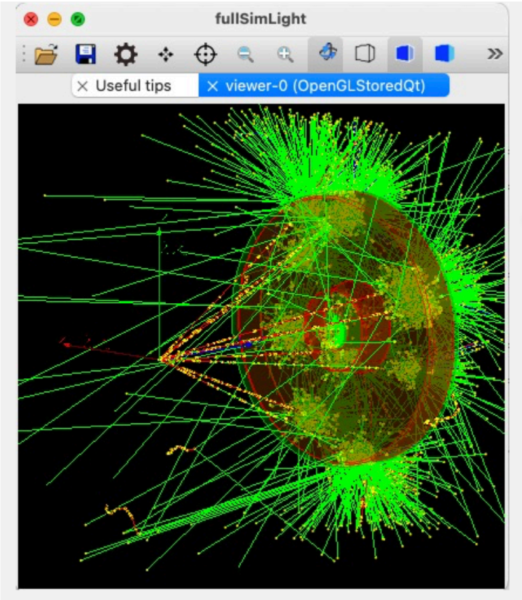
COMING SOON

GPU-friendly EMEC implementation

- ***Background:***
  - The as-built EMEC features a complex "Spanish Fan" geometry.
    - Early versions of Geant4 did not support an efficient description of this geometry using the available G4Solids.
    - To implement it, a custom solid was used with algebraic methods.

- ***Idea***: Re-implement the EMEC geometry starting from the technical drawings and with standard Geant4 shapes
  - This code will enable portability of the full ATLAS geometry allowing to offload the Simulation jobs to GPUs





Visualisation of a simulation run with the new EMEC

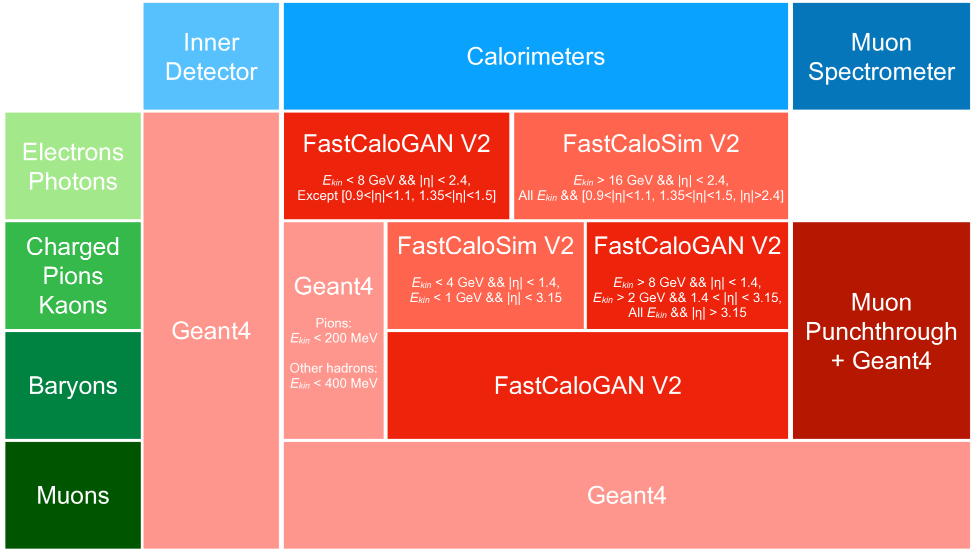
- **STATUS:** Implemented by using `G4GenericTrap`
  - Initial tests in standalone G4 example and in FullSimLight
    - Discovered a bug in the calculation of the Safety distance of the `G4GenericTrap` (fix available from G4 11.3)
  - Code now integrated in Athena.
    - The next step is to adjust the corresponding sensitive detectors code to align it with the new geometry implementation, enabling the evaluation of physics modelling.
  - Preliminary benchmarks show a **19% speedup**

Use AF3 in the EMEC in FullSim

COMING SOON

- ***Idea***: use AtlFast3 to handle low energy e/γ in the EMEC for the Full Simulation workflow

- **Motivation**: Even after MC23 optimizations, e/γ in EMEC are one of the biggest contributors to the total steps-per-event
  - Similar to Frozen Showers in ATLAS FCAL for low energy e/γ/neutrons
  - Goal is to find E/η region where outputs do not differ w.r.t. full Geant4



- **STATUS:**
  - AF3 appears to be precise enough to act as a substitute for G4
    - No significant impact on the shower shapes for E<8GeV, 1.5<η<2.5
    - Preliminary ~10% speed-up for sample in the same eta range

Specialised transport: G4HepEM/AdePT/Celeritas

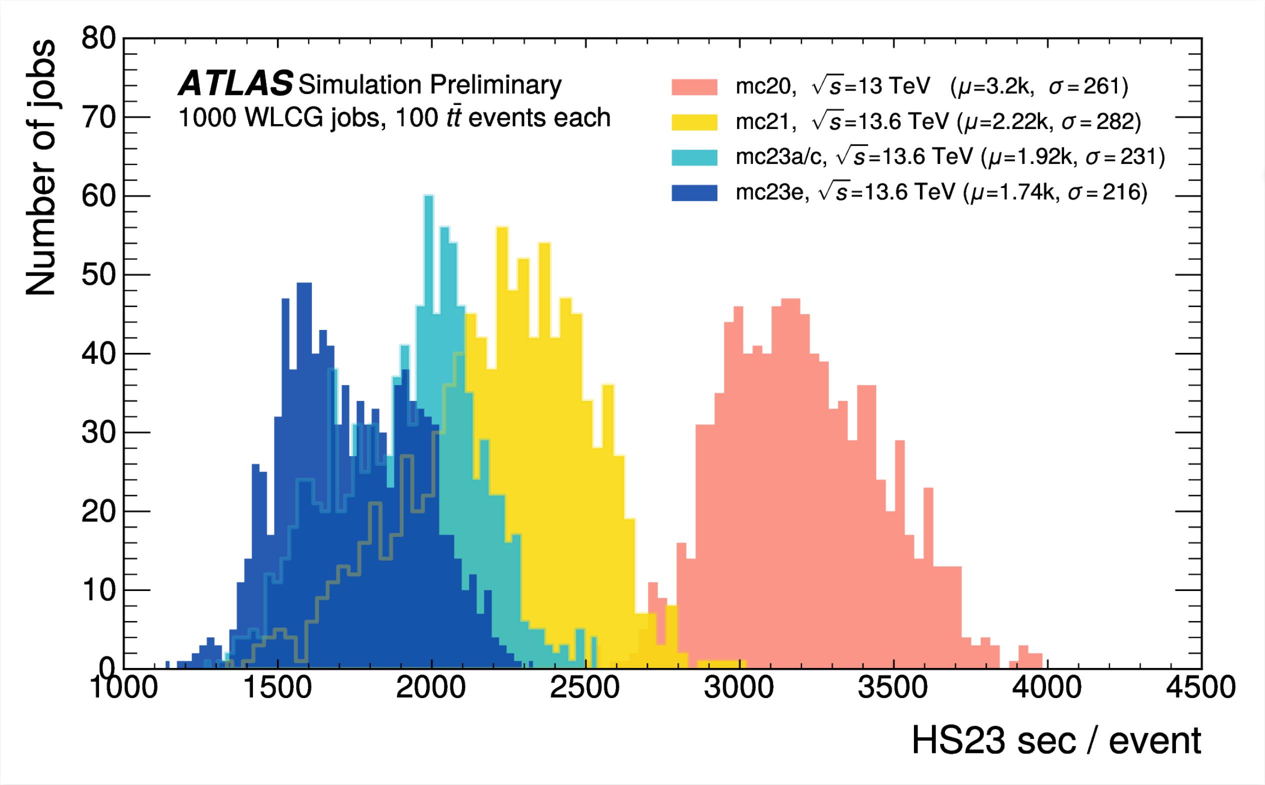
- **Geant4 11 Enhancements:**
  - **Per-particle Specialised Transport:**
    - Geant4 11 introduces the ability to customize particle transport based on particle type, allowing users to choose actions (such as stepping or parameterization) based on particle energy and location.
    - This adds greater flexibility, coherence, and potentially better performance compared to previous methods and hooks.
  - **G4HepEM:**
    - A compact Geant4 electromagnetic (EM) library with algorithms and memory layout optimized for high-energy physics (HEP) EM shower development and e-/e+/γ transport, validated against more general Geant4 EM models.
  - **AdePT and Celeritas:**
    These frameworks implement full stepping transport loops for e-/e+/γ particles on NVIDIA (and AMD) GPUs, tailored for HEP use cases involving geometry and physics processes.

- **Integration with Athena:**
  Work is underway to integrate these tools into Athena, using specialized transport to offload e-/e+/γ particles (e.g., in calorimeters) to the GPU asynchronously while maintaining event boundaries on the host side.
  - AdePT/Celeritas in Athena hackathon planned for next week
    - Goal: Integrate Celeritas and AdePT with Athena, so that GPU hits can be passed on to the ATLAS sensitive detectors to be processed and written out to a HITS file. The HITS file can then be processed by the rest of the atlas SW chain.
    - Runtime performance will be assessed on a range of host-device systems and under realistic production conditions.

Benchmarks

Performances of ATLAS Simulation in Run3

Public Plots



HepScore23 is the HEP wide benchmark for measuring CPU performance

Reference [1]

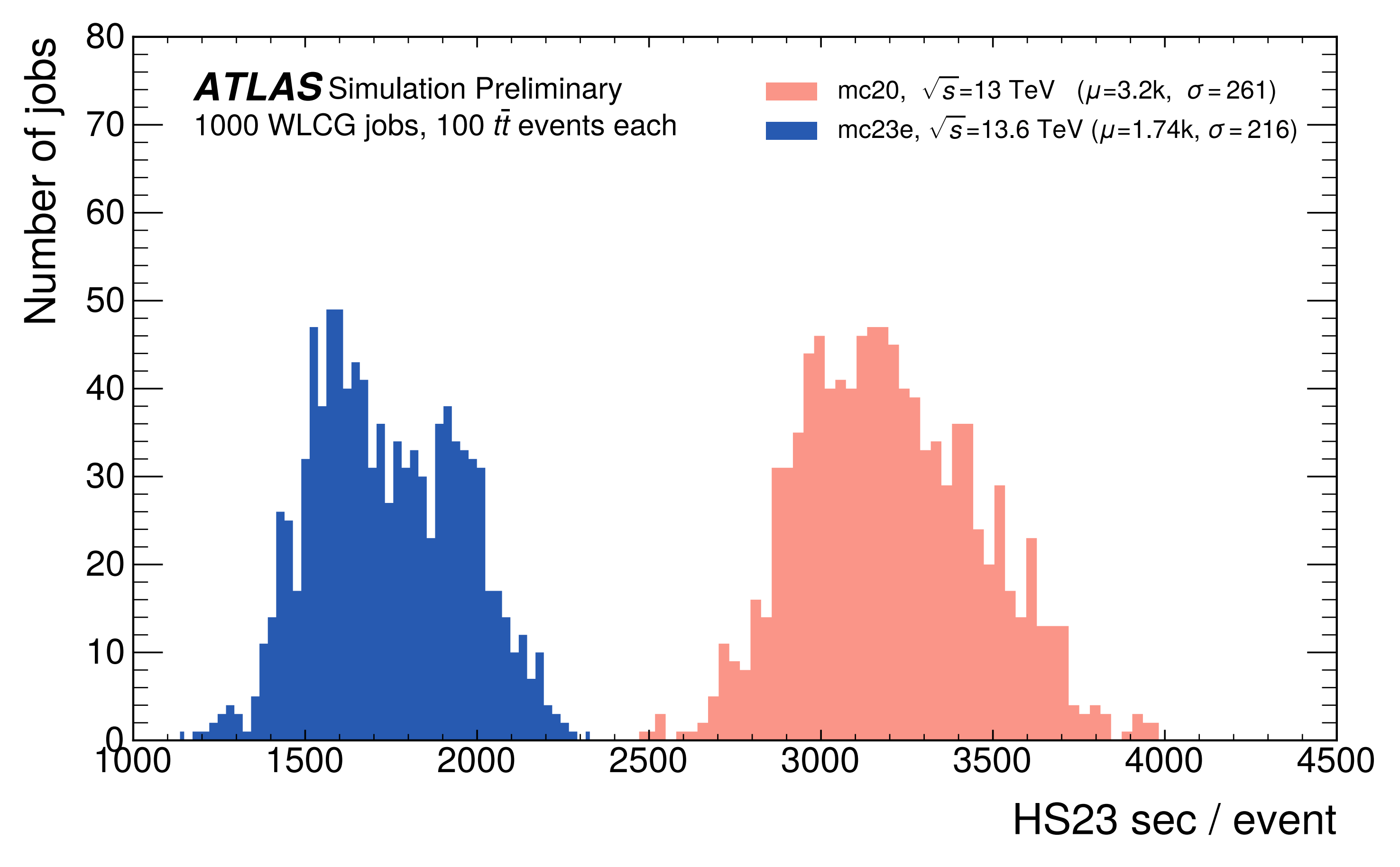
Throughput increased by x1.84 between mc20 and mc23e campaigns!

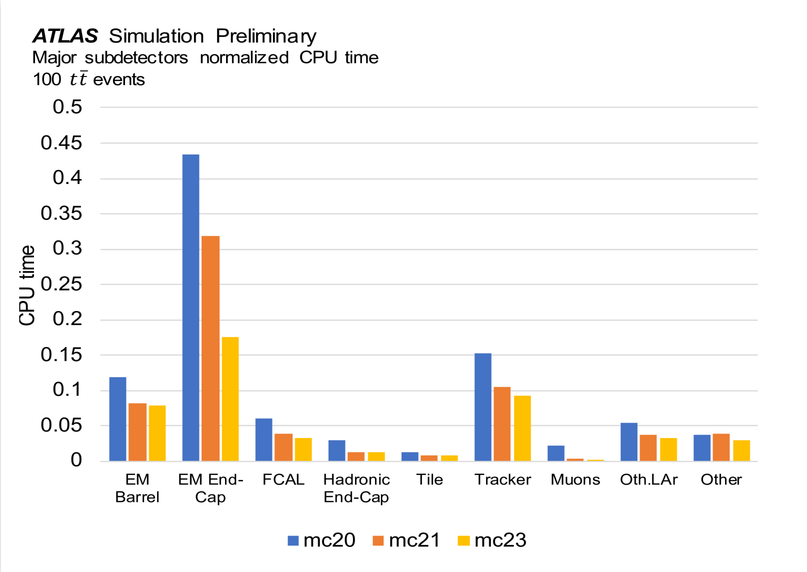
- mc20 uses Geant4 10.1
- mc21 & mc23x use Geant4 10.6
- Slowdown of ~5% between the 2 sw releases
- Slowdown of ~3% when moving from 13TeV to 13.6 TeV centre-of-mass energy

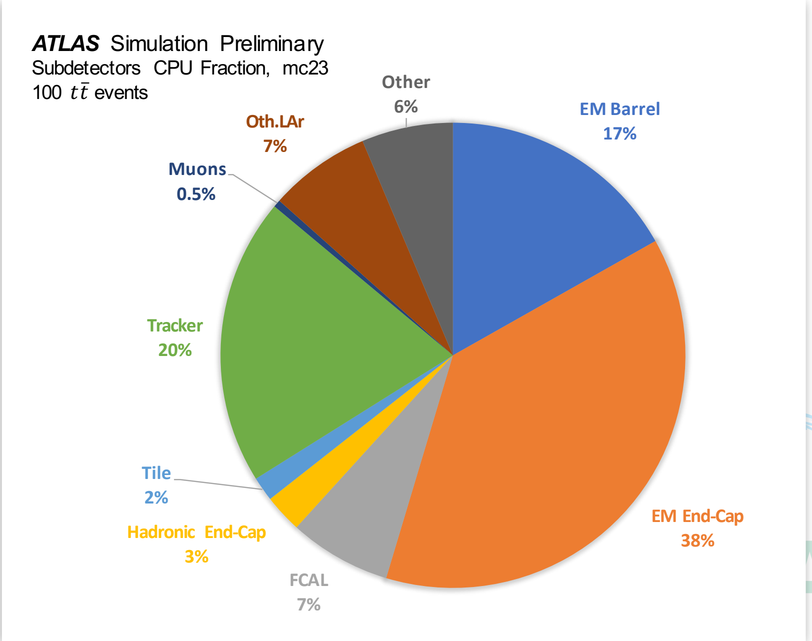
The mean CPU time per event simulated in Full Simulation comparing mc20 (Run2 MC Campaign) running at 13 TeV centre-of-mass energy (in red), mc21 (start of Run3 MC Campaign), running at 13.6 TeV centre-of-mass energy (in yellow), mc23c (Run3 MC campaign covering 2023 data taking), running at 13.6 TeV centre-of-mass energy (in light blue) and mc23e (Run3 MC campaign covering 2024 data taking), running at 13.6 TeV centre-of-mass energy (in dark blue) measured in standardised HS23 seconds.

Performances of ATLAS Simulation in Run3

Public [Plots](#)







Conclusions and outlook

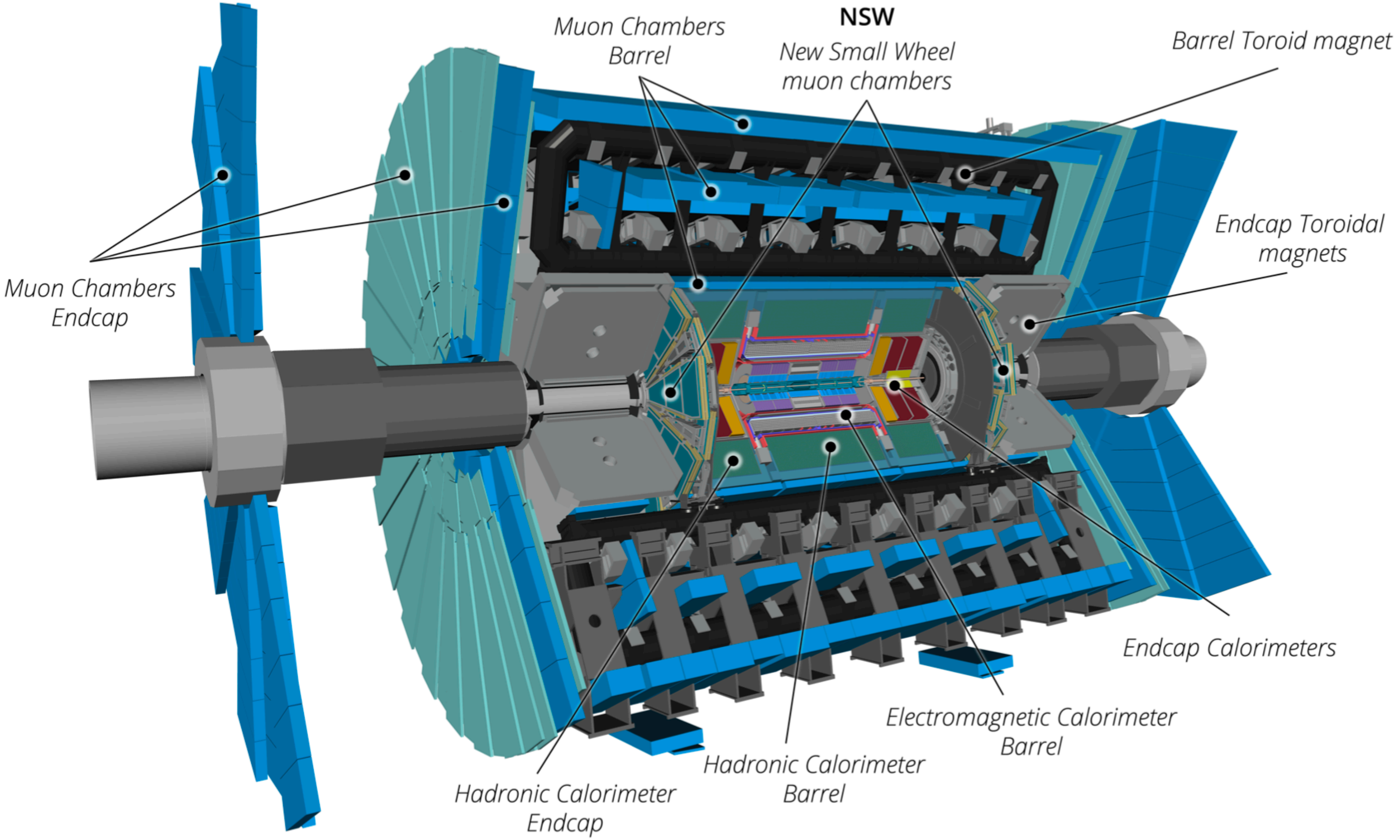
- **Run3 Full Simulation** runs 1.84 **times faster** than in Run2!
  - Many interesting ongoing and planned **projects for optimizations** in the Full Simulation Group:
    - Additional **5% speedup from LTO** in the latest MC campaign
  - Other **optimisations ongoing**
    - High-η particle rejection (ISF particle killer) **(few % expected)**
    - GPU-friendly EMEC implementation **(expected 19% speedup)**
    - G4HepEM library adoption **(preliminary tests in Athena ~7-8% speedup)**
    - Use AF3 in the EMEC in FullSim **(few % expected)**
    - **Energy Loss Fluctuation OFF (3-5%)**
  - Many interesting **longer-term developments** (please see backup slides):
    - Integration with AdePT/Celeritas
    - Quantized State Stepper (QSS) integration and testing
    - ML Correction for Aggressive Range Cuts
    - G4 Field parameter tuning

Thanks for your attention!

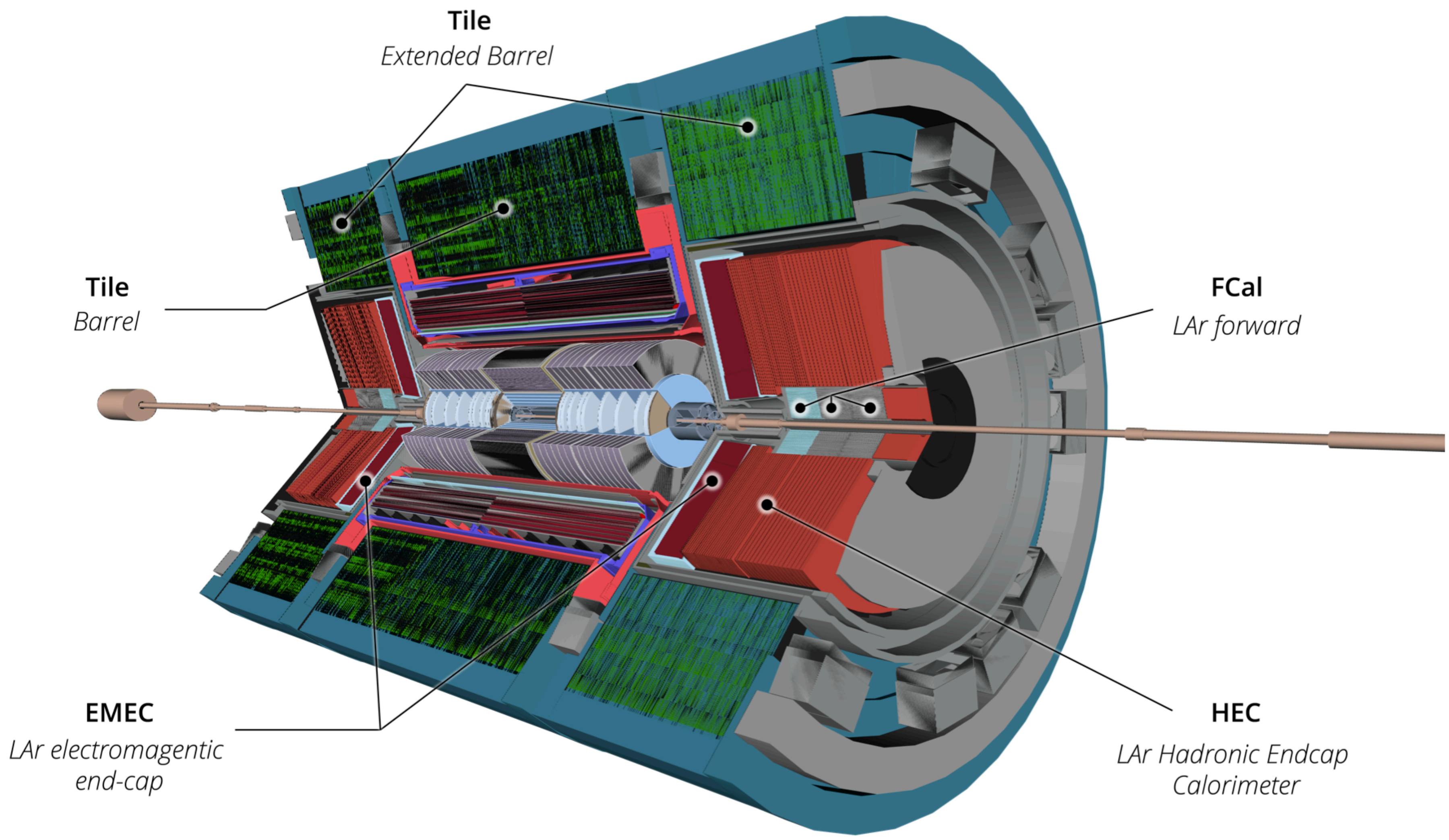
Marilena Bandieramonte
marilena.bandieramonte@cern.ch

Backup slides

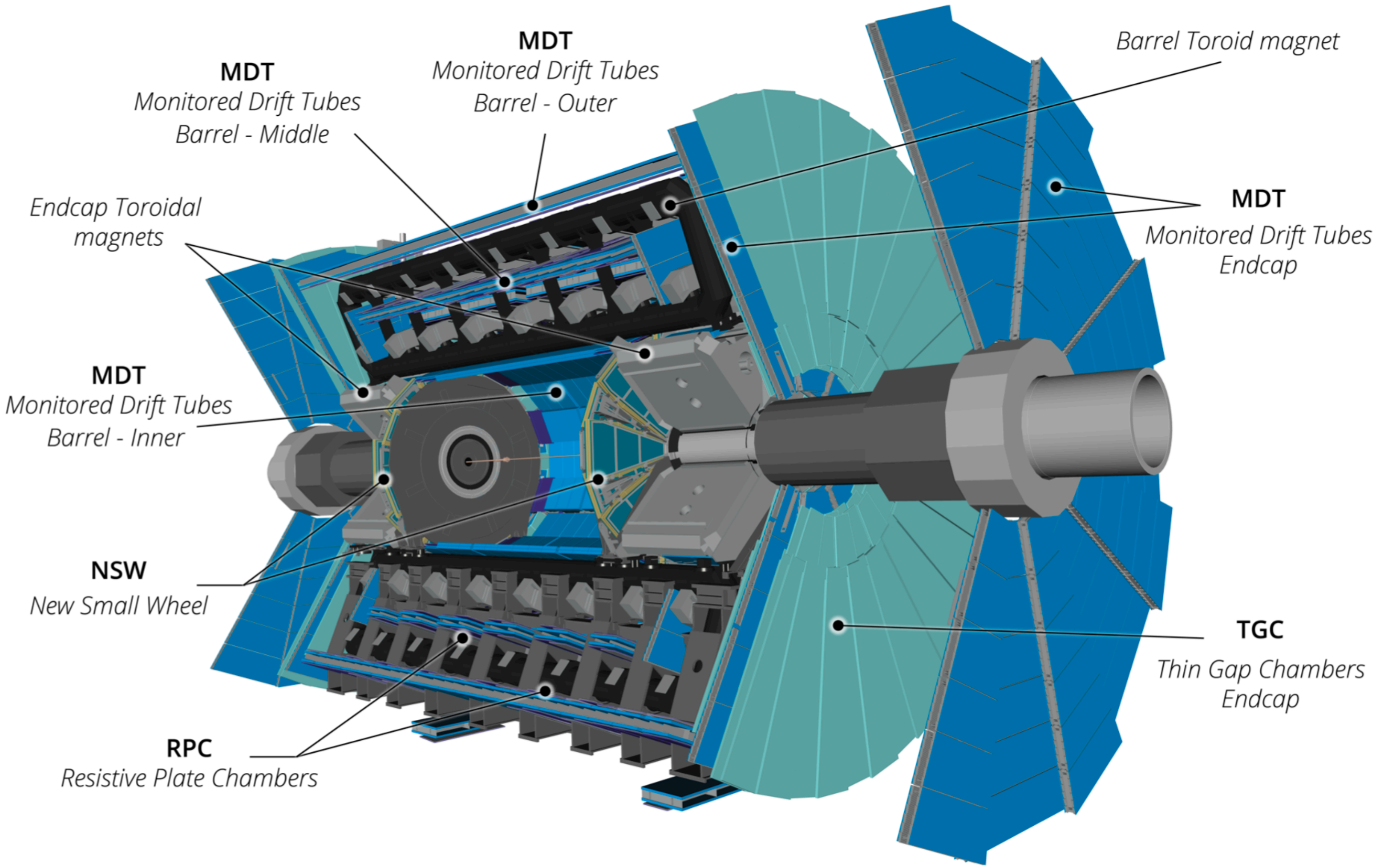
ATLAS detector in Run3



ATLAS detector in Run3 - Calorimeters

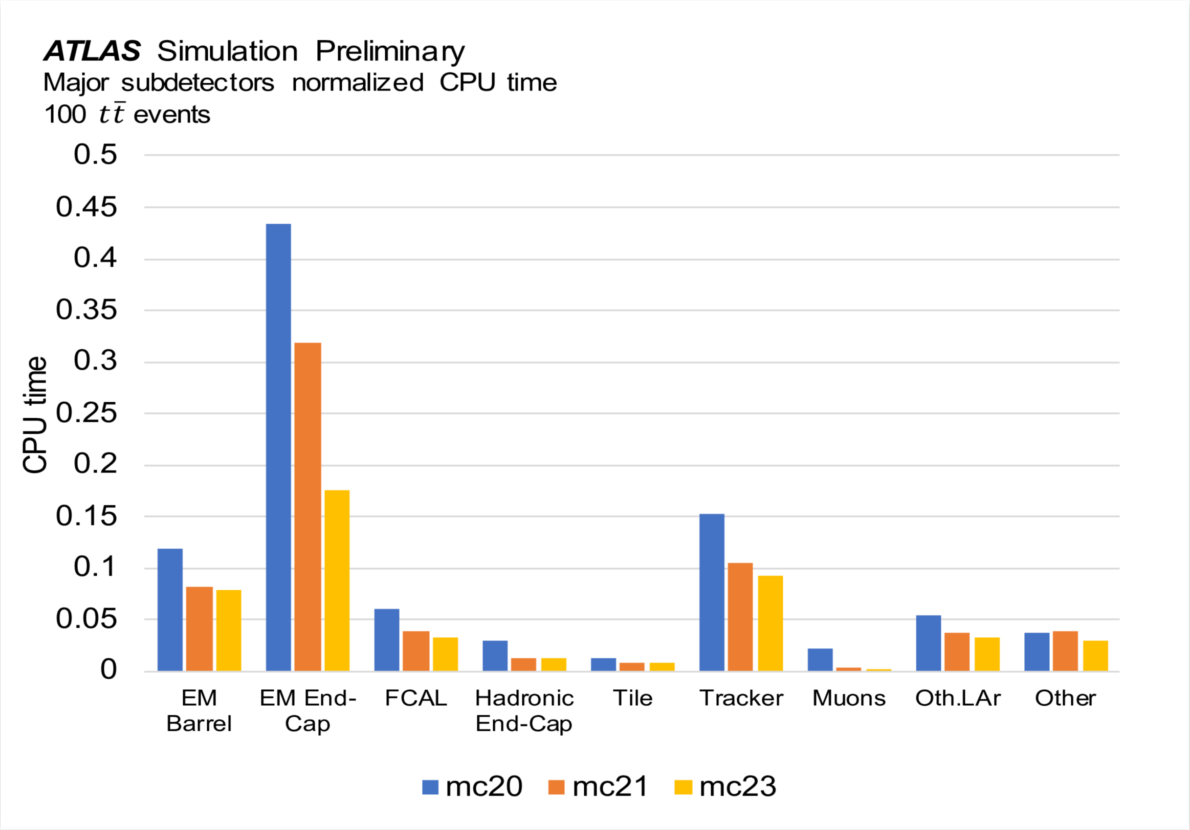


ATLAS detector in Run3: Muons System



Performances of ATLAS Simulation

- Looking from a different perspective (time spent per sub-detector and per particle type) gives us more insights:

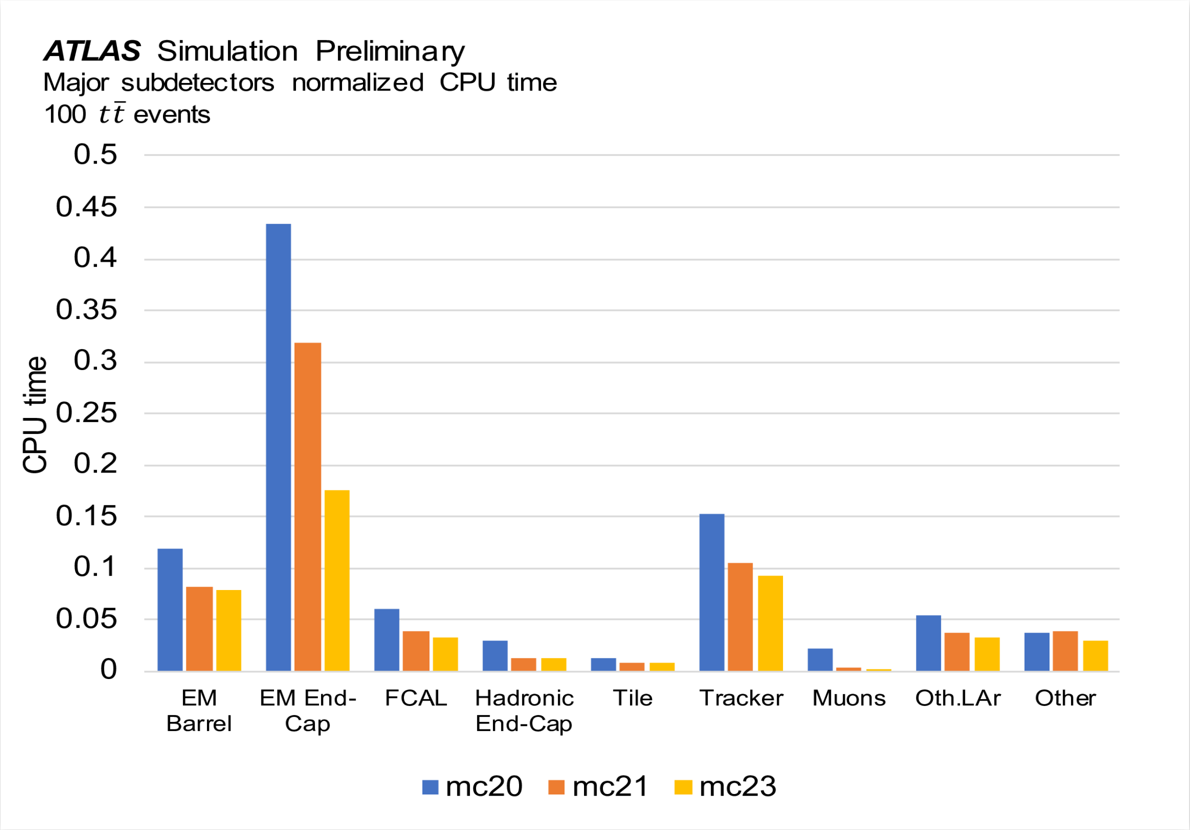


Improvements over different campaigns

- **mc20, MC reprocessing campaign for Run2:**
  - TRT range cut (sets the range cut for e-/e+ in the TRT volumes filled with Xenon)
  - Frozen Showers
- **mc21, MC campaign for start of Run3: (in addition to mc20)**
  - EM range cuts: expect to reduce the n. of low energy electrons
  - PRR: expect to reduce the n. of tracked photons
  - NRR: expect to reduce the n. of tracked neutrons
  - EMEC slices: improve the CPU time in the EMEC
  - Big Library: improve overall the Full Simulation Time
  - MagField Off: improve transport in magField
  - G4GammaGeneralProcess: improve time spent to transport gammas
  - BeamPipe killer: kills particles at eta>5.5, that do not make to the detector
- **mc23c (in addition to the mc21), MC campaign for Run3:**
  - Woodcock Tracking: improve timing of gammas in the EMEC
  - VecGeom: Improve CPU time for detectors that use Polycons, Tubes and Cones

Performances of ATLAS Simulation

- Looking from a different perspective (time spent per sub-detector and per particle type) gives us more insights:

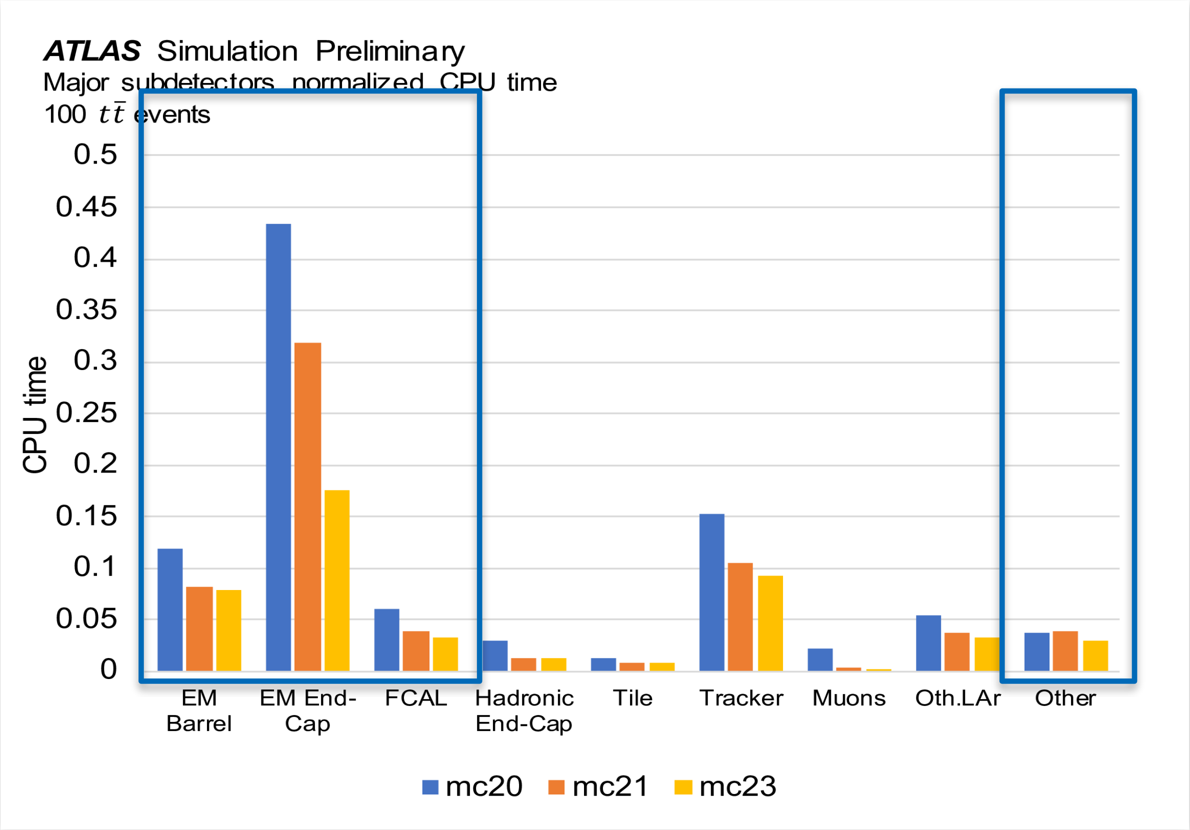


Changes affecting all subsystems

- **mc20, MC reprocessing campaign for Run2:**
  - TRT range cut (sets the range cut for e-/e+ in the TRT volumes filled with Xenon)
  - Frozen Showers
- **mc21a, MC campaign for start of Run3: (in addition to mc20)**
  - EM range cuts: expect to reduce the n. of low energy electrons
  - PRR: expect to reduce the n. of tracked photons
  - NRR: expect to reduce the n. of tracked neutrons
  - EMEC slices: improve the CPU time in the EMEC
  - Big Library: improve overall the Full Simulation Time
  - MagField Off: improve transport in magField
  - G4GammaGeneralProcess: improve time spent to transport gammas
  - BeamPipe killer: kills particles at eta>5.5, that do not make to the detector
- **mc23c (in addition to the mc21), MC campaign for Run3:**
  - Woodcock Tracking: improve timing of gammas in the EMEC
  - VecGeom: Improve CPU time for detectors that use Polycons, Tubes and Cones

Performances of ATLAS Simulation

- Looking from a different perspective (time spent per sub-detector and per particle type) gives us more insights:

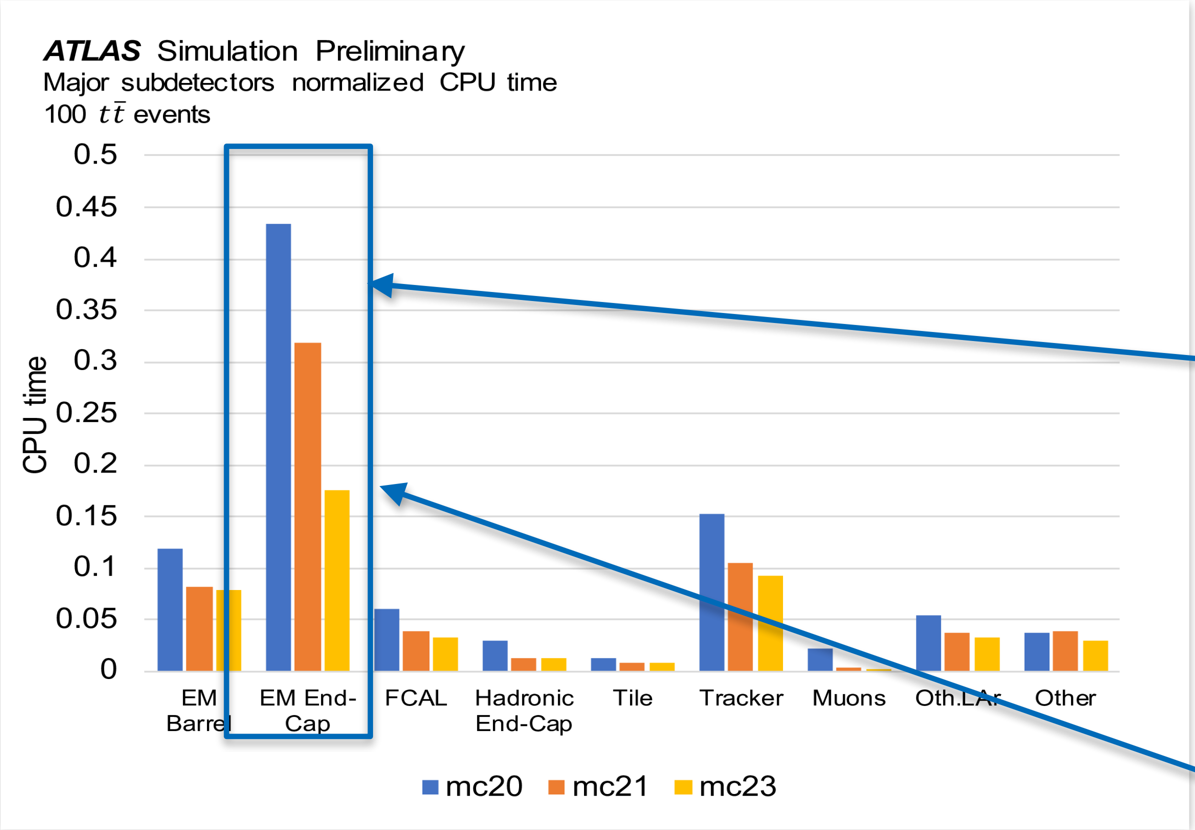


Changes impacting mainly EM calorimeters

- **mc20, MC reprocessing campaign for Run2:**
  - TRT range cut (sets the range cut for e-/e+ in the TRT volumes filled with Xenon)
  - Frozen Showers
- **mc21a, MC campaign for start of Run3: (in addition to mc20)**
  - EM range cuts: expect to reduce the n. of low energy electrons
  - PRR: expect to reduce the n. of tracked photons
  - NRR: expect to reduce the n. of tracked neutrons
  - EMEC slices: improve the CPU time in the EMEC
  - Big Library: improve overall the Full Simulation Time
  - MagField Off: improve transport in magField
  - G4GammaGeneralProcess: improve time spent to transport gammas
  - BeamPipe killer: kills particles at eta>5.5, that do not make to the detector
- **mc23c (in addition to the mc21), MC campaign for Run3:**
  - Woodcock Tracking: improve timing of gammas in the EMEC
  - VecGeom: Improve CPU time for detectors that use Polycons, Tubes and Cones

Performances of ATLAS Simulation

- Looking from a different perspective (time spent per sub-detector and per particle type) gives us more insights:



Changes targeting the EMEC

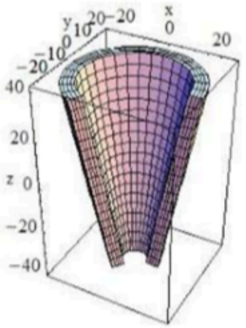
- **mc20, MC reprocessing campaign for Run2:**
  - TRT range cut (sets the range cut for e-/e+ in the TRT volumes filled with Xenon)
  - Frozen Showers
- **mc21a, MC campaign for start of Run3: (in addition to mc20)**
  - EM range cuts: expect to reduce the n. of low energy electrons
  - PRR: expect to reduce the n. of tracked photons
  - NRR: expect to reduce the n. of tracked neutrons
  - EMEC slices: improve the CPU time in the EMEC
  - Big Library: improve overall the Full Simulation Time
  - MagField Off: improve transport in magField
  - G4GammaGeneralProcess: improve time spent to transport gammas
  - BeamPipe killer: kills particles at eta>5.5, that do not make to the detector
- **mc23c (in addition to the mc21), MC campaign for Run3:**
  - Woodcock Tracking: improve timing of gammas in the EMEC
  - VecGeom: Improve CPU time for detectors that use Polycons, Tubes and Cones

Geant4 Simulation Optimizations for Run3

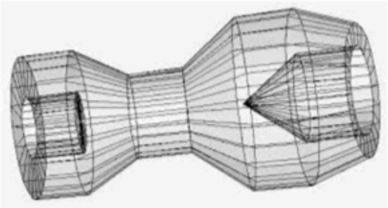
VecGeom: vectorized Geometry

New and optimised implementation of geometrical shapes, designed to take advantage of explicit and implicit vectorisation:

- we replace only *polycons, cones and tubes*



2-3%



measured on 500 ttbar events in the Atlas software framework (Athena)

VecGeom alone

BNL Cluster, 1000 Jobs 100 tt-bar events/job	Walltime,s	sigma	Speedup CPU time	Job ID
Athena 22.0.47 (baseline)	280	22.8	-	27857092
Athena 22.0.63 with VecGeom (no BigLibrary)	263	23	-6.1%	31379861
Athena 22.0.63 with VecGeom (no BigLibrary) +Run3Opt	195	15.9	-30.4%	31379871

VecGeom on top of the Big Library

BNL Cluster, 1000 Jobs 100 tt-bar events/job	Walltime,s	sigma	Speedup CPU time	Job ID
Athena 22.0.47 (baseline)	280	22.8	-	27857092
Athena 22.0.67 "out the box" (BigLibrary only)	270	21.3	-3.6%	31366397
Athena 22.0.67 "out the box" (BigLibrary only) +Run3Opt	189	15.7	-32.5%	31366406
Athena 23.0.7 "out the box" (BigLibrary + VecGeom)	252	21.6	-10%	31352919
Athena 23.0.7 "out the box" (BigLibrary + VecGeom)+Run3Opt	**185**	**15.8**	**-33.9%**	**31352932**

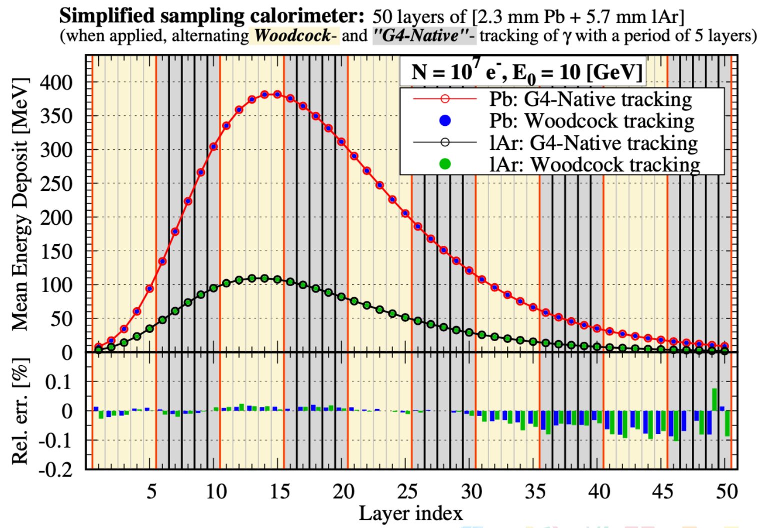
Ben Morgan

Woodcock tracking

- Reduce the **number of steps for gammas** in highly segmented detectors (e.g. EMEC), where the geometric boundaries limit the step, rather than the physics
- Perform the transport in a geometry with no-boundaries made of the densest material (i.e. Pb)
- The interaction probability is proportional to the cross-section ratio between the real material and the “fake" one
- Integrated in FullSimLight *and in* ***Athena***
  - Shows a ~**17.5%** performance gain when using *Woodcock-tracking* of gamma photons in the EMEC detector region on top of the gain already provided by the *Gamma-general* process
  - **Full physics validation successful and in production for Run3!**

		FTFP_BERT_ATL	_WDCK-*GammaGeneral*	_WDCK *Woodcock*(EMEC)
#secondary	γ	3.054e+05	3.06e+05	3.062e+05
	e^-	6.240e+05	6.204e+05	6.201e+05
	e^+	2.186e+04	2.19e+04	2.193e+04
#steps	charged	3.548e+06	3.548e+06	3.550e+06
	neutral	**8.501e+06**	**8.464e+06**	**4.215e+06**

FTFP_BERT_ATL	_WDCK-*GammaGeneral*	_WDCK *Woodcock*(EMEC)
base line	[2 - 4] %	[10 - 14] %

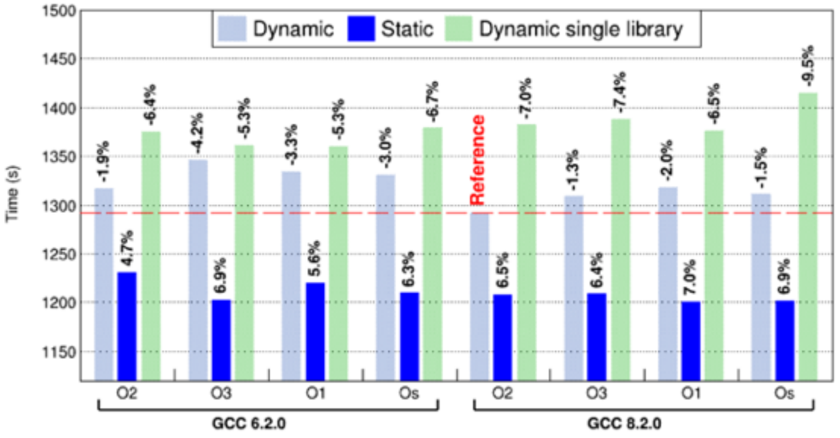


M. Novak, J. Apostolakis

Geant4 intrinsic improvements

Geant4 static linking (Big Library)

Combine all Athena libraries/components that use Geant4 into a ***single shared library*** linking to Geant4 ***static libraries***



-7%

measured in Athena

B. Morgan, C. Macron [ATLASSIM-3150]

G4GammaGeneralProcess

A single general process for gammas that interacts with the Stepping Manager:
- significantly reduce the number of operations needed

-4.3%

measured on 100 ttbar events in Athena

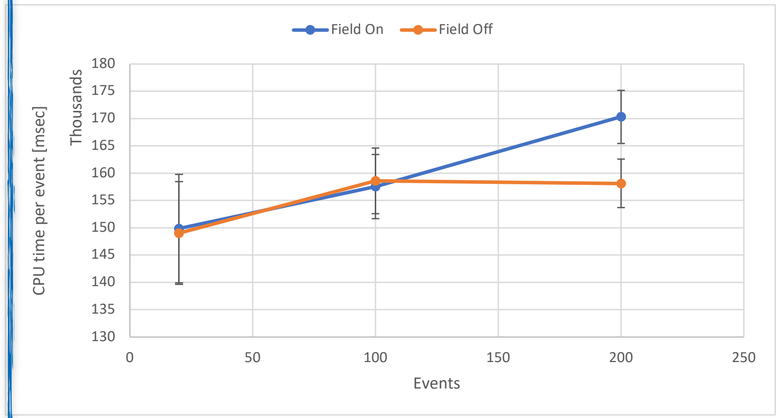
M. Bandieramonte [ATLASSIM-4791] [ATLPHYSVAL-818]

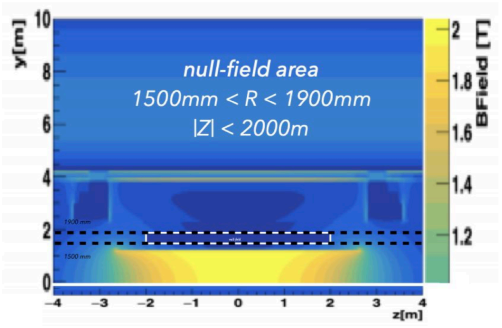
Reduce the number of steps/operations

Magnetic Field Tailored Switch-off

Speedup observed when switching off magnetic field in LAr calorimeter (except for muons) without affecting shower shapes

- **Detailed studies showed smaller null-field area needed**
  - ~1-2% speedup for full ttbar events





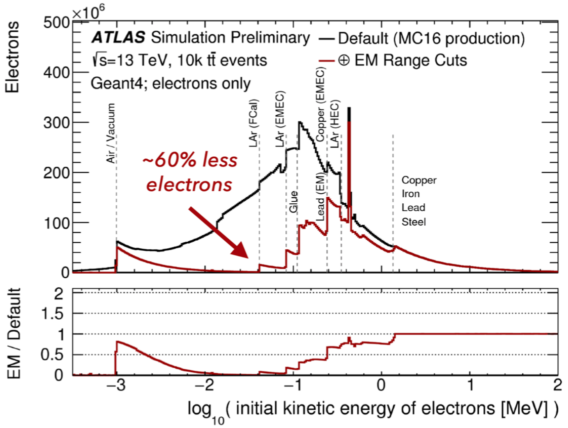
V. Kourlitis [ATLASSIM-4749] [ATLPHYSVAL-773]

EM range cuts

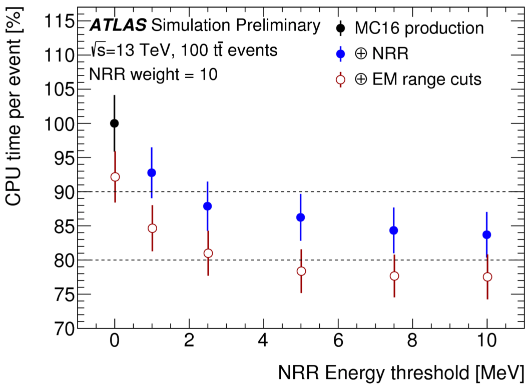
Explicitly activated for *Compton, photoelectric, pair-production*

Avoid creating secondaries and transporting a gamma if its energy is below a certain value

6-7% speedup



Photon and Neutron Russian roulette



Randomly discard particles below the energy threshold and weight accordingly the surviving ones

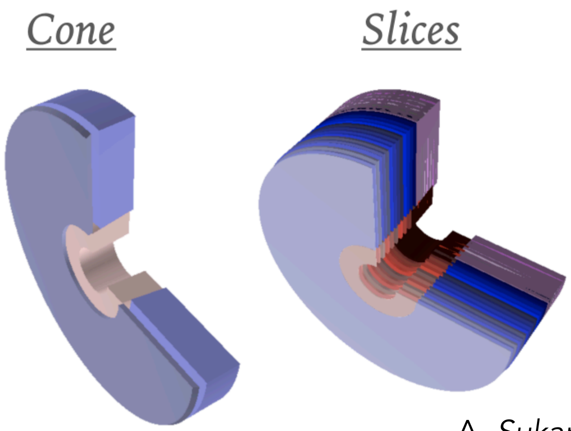
10% speedup with 2 MeV threshold

M. Muskinja [SIM-2019-001] [ATLASSIM-3924] [ATLPHYSVAL-632]

Geometry optimizations

New EMEC custom solid variants

- Wheel the default LArWheelSolid with G4Polycone
- Cone improved LArWheelSolid with G4ShiftedCone - outer wheel divided into two conical-shaped sections
- **Slices** new LArWheelSliceSolid — each wheel is divided into many thick slices along Z axis:
  - **best candidate: 5-6% speedup**



A. *Sukarev* [ATLPHYSVAL-785]

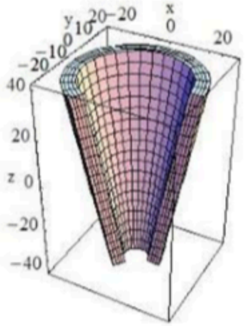
VecGeom

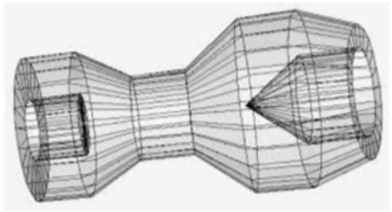
New and optimised implementation of geometrical shapes, designed to take advantage of explicit and implicit vectorisation:

- we replace only *polycons, cons and tubes*

-2-3% **

measured on 500 ttbar events in Athena





Ben Morgan [ATLASSIM-4750] [ATLPHYSVAL-831]

Longer terms effort

G4 Field parameter tuning

Goal: Find the optimal values of the in-field tracking parameters for physics performance and CPU savings
Inspired by CMS efforts to optimize the below parameters

- DeltaIntersection: accuracy of intersection with boundary volume
- Epsilong_max: relative accuracy for endpoint of 'ordinary' integration step (delta_one_step / step_size)
- DeltaChord: approximation of curve with linear sections
- MaxStep: maximum step length

Magnetic field parameters	**Strict**	**Intermediate**	**Loose**
DeltaIntersection (mm)	10^{-6}	10^{-4}	0.01
Epsilon_max	10^{-4} /step_size	10^{-3} /step_size	0.1/step_size
DeltaChord (mm)	10^{-3}	$2 \cdot 10^{-3}$	0.1
MaxStep (mm)	150	50	150

N. *Nitika* [ATLASSIM-6147]

Quantized State Stepper (QSS)

Quantized State System (QSS): numerical methods to solve the ordinary differential equations that describe the movement of particles in a field

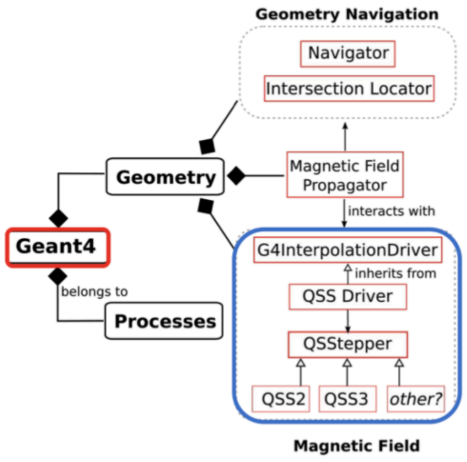
- QSS methods **discretize the system state** variables as opposed to the traditional methods that discretize the time

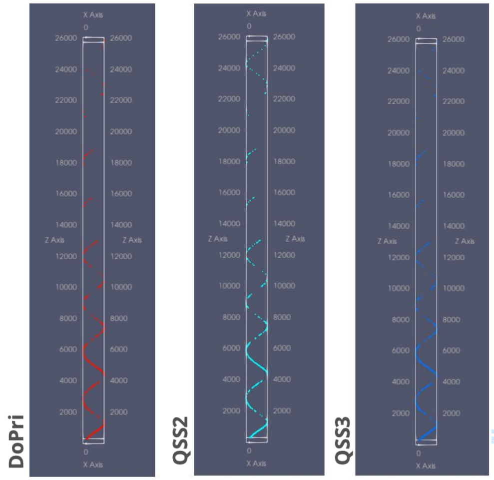
- This method handles very efficiently **discontinuities** in the simulation of continuous systems

Reference: Efficient discrete-event based particle tracking simulation for high energy physics

STATUS

- Successfully ported QSS stepper from Geant4 v10.5 to v10.7.2
    - Test within the ATLAS geometry:
        - Results using the N02 model qualitatively indistinguishable compared to those using the G4DormandPrince745
- Integrated in **FullSimLight**
    - ATLAS geometry & magnetic field map
    - Performance profiling ongoing



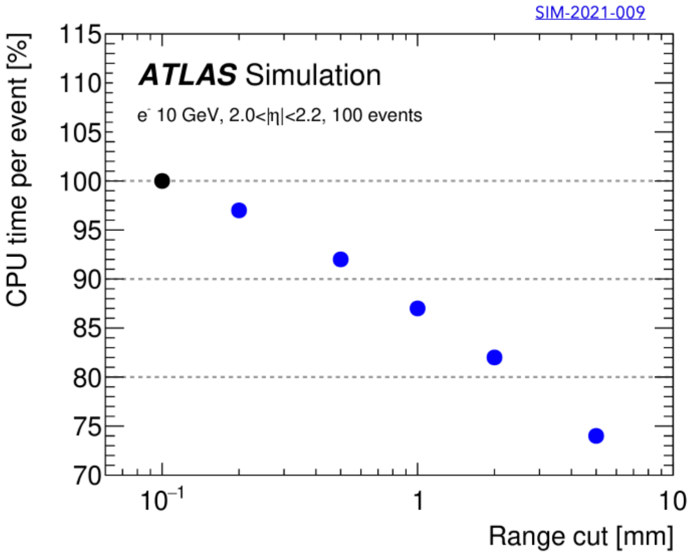


R. Castro, L. Santi,Leandro Gómez Vidal, Alejandro Mignanelli: [link]

ML Correction for Aggressive Range Cuts

Increased range cuts can **reduce the number of photons**, thus reduce the transportation steps and **increase computational performance**

EM calorimeters dominate the simulation load due to low-energy photons from electron scattering, ~90% of these are transportation processes



Side-effect:
"High" range cuts can degrade the accuracy of the simulation

The ML correction applied as a post-processing step utilizing **batch processing** and **accelerator hardware** achieving **~15% speed up** in example geometries – ML inference time negligible compared to simulation time reduction.

Solution to be implemented/tuned for the ATLAS EMEC

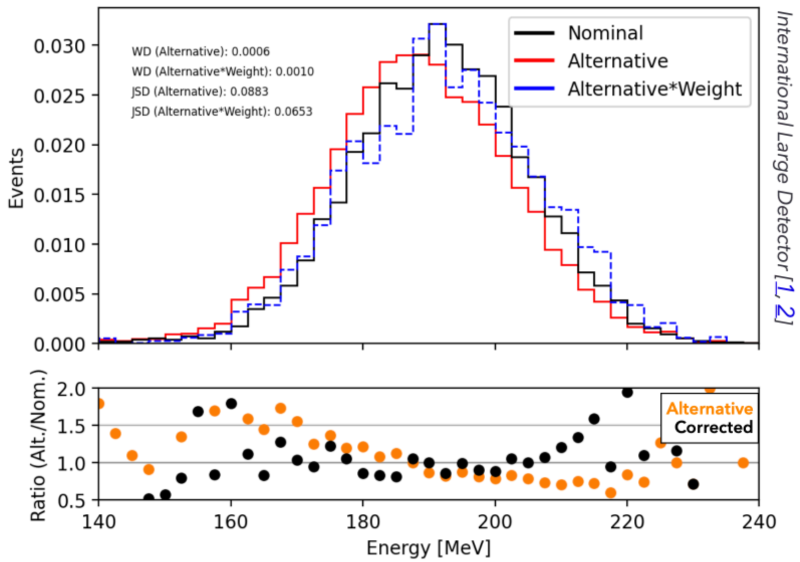
ML-based correction

Classification NN to learn correction weights [ref]

Re-weight the ***alternative*** *simulation to the* ***nominal*** *one by learning multi-dimensional weights considering all cell energy deposits*

$$r(\vec{x}) = \frac{p(\vec{x} \mid \theta_p)}{q(\vec{x} \mid \theta_q)}$$

θ be the range cut, $\boldsymbol{x}$ the energy deposits



V. Kourlitis, D. Mayilyan link