PROCEEDINGS OF SCIENCE

The Fast Simulation Program of ATLAS at the LHC

Martina Javurkova* **on behalf of the ATLAS computing activity**

University of Massachusetts Amherst,
Amherst, Massachusetts, USA

E-mail: martina.javurkova@cern.ch

The simulation of Monte Carlo (MC) events is a crucial task and an indispensable ingredient for every physics analysis. GEANT4 is the state-of-the-art tool used for detailed simulations of the ATLAS detector, which however requires large CPU resources. To reduce the CPU needs, which in turn enables the production of higher statistics MC samples, ATLAS has developed a strong program to replace parts of the simulation chain by fast simulation tools. These developments pave the way towards High Luminosity LHC when resources will be even scarcer. Among those tools is ATLFAST3, which utilises a combination of Generative Adversarial Networks (GANs) and sophisticated parametrisations for the fast simulation of showers in the electromagnetic and hadronic calorimeters. For the Run 3 MC campaign, various improvements of ATLFAST3 were developed, for example a refinement and extended usage of the GANs and a better model of the punch through of showers into the muon system. Consequently, the performance of ATLFAST3 in Run 3 is better than ever. ATLAS also aspires to use fast simulation in the inner detector. FATRAS is a tool that approximates particle interactions with the material through physics formalisms. An integration of FATRAS with the experiment-independent common tracking software (ACTS) is also in development. Track overlay is a technique to speed-up the production of MC samples that include additional interactions (pile-up) aside the hard-scatter interaction. The idea is to reconstruct pile-up tracks before they are merged with the hard-scatter, which reduces CPU needs. Machine learning techniques are used to ensure this method can even be applied in dense tracking environments. This talk will discuss the status of the development of these tools as well as their performance in terms of physics modelling and computing resources.

ATL-SOFT-PROC-2024-009
30 September 2024

42nd International Conference on High Energy Physics (ICHEP2024)
18-24 July 2024
Prague, Czech Republic

*Speaker

1. Introduction

The ATLAS experiment [1] at the Large Hadron Collider (LHC) relies heavily on Monte Carlo (MC) simulations to model detector responses and interpret physics events. Currently, around 70% of ATLAS's grid CPU time is consumed by MC production, mainly for full detector simulations using GEANT4 [2]. As the LHC enters the High-Luminosity era (HL-LHC), which will deliver unprecedented complex events with up to 200 proton-proton interactions per bunch crossing, the demand for both computing and storage will rise significantly. This trend is evident in the projected annual CPU consumption and tape storage needs, as shown in Figure 1. Fast simulations offer a more efficient alternative, balancing accuracy and speed, and are essential for meeting these challenges and ensuring sufficient event sample production in future physics analyses.

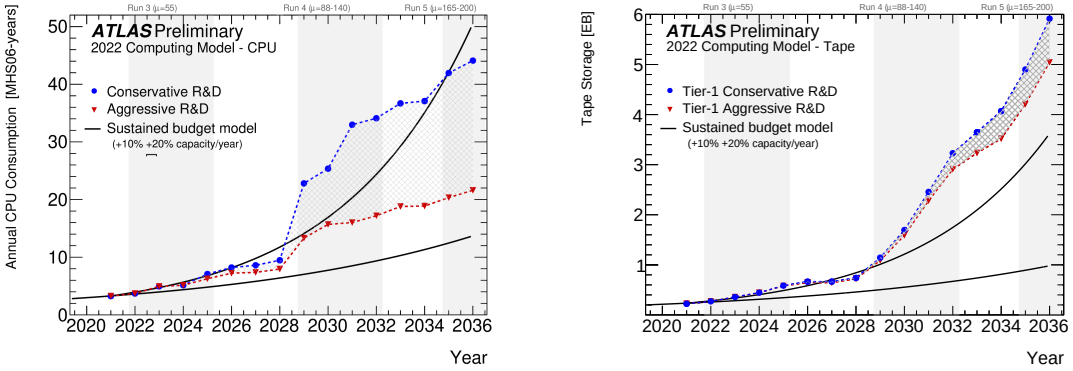

Figure 1: Projected evolution of compute (left) and tape (right) usage, under the conservative (blue) and aggressive (red) R&D scenarios. The black lines indicate the impact of sustained year-on-year budget increases. Taken from [3].

2. Fast Calorimeter Simulation

One of the most CPU-intensive components of the GEANT4 simulation in the ATLAS experiment is the calorimeter shower simulation, accounting for nearly 80% of the total simulation time. ATLFAST3 [4, 5] provides a fast solution that replaces the slow propagation and interaction of particles inside the calorimeters with the direct generation of energy deposits based on an underlying parameterisation. First introduced in Run 2 and further improved for Run 3, ATLFAST3 simplifies the complex geometry of the calorimeter and combines two complementary techniques: `FastCaloSimV2`, a parametric approach for shower development, and `FastCaloGANV2`, which leverages generative adversarial networks (GANs). Together, these tools aim to replicate the accuracy of the full GEANT4 simulation while achieving significant speed-ups.

The parametrisation used for `FastCaloSimV2` and the training of GANs is derived from simulations of single particles with GEANT4. These simulations encompass various particle types, including photons and electrons for electromagnetic showers, as well as charged pions for hadronic showers; in the case of `FastCaloGANV2`, protons are also included. To ensure comprehensive coverage, the parameterisation is obtained for 100 linearly spaced bins spanning up to $|\eta| < 5$. Additionally, energy ranges are sampled across 17 logarithmically spaced energy bins.

`FastCaloSimV2` employs a separate parameterisation for longitudinal and lateral shower development. Longitudinally, the energy deposited in layers is decorrelated using Principal Component Analysis (PCA), while laterally, the average shower profile is parameterised as 2D probability density functions. In contrast, `FastCaloGANV2` uses a more sophisticated approach, involving 600 GANs - one for each particle type and energy bin, conditioned on true momentum. These GANs are trained to reproduce energy deposition in the calorimeter's voxels, layers, and the total calorimeter energy in a single step. Voxels are small 3D grid units that group the calorimeter hits, and their granularity is optimised to be finer than the calorimeter cells, thereby improving model accuracy. The network architecture and hyperparameters are finely tuned for optimal performance. During simulation, hits are generated based on the selected technique, particle type, and energy, with additional corrections applied to ensure results align with GEANT4 accuracy.

ATLFAST3 integrates the strengths of both `FastCaloSimV2` and `FastCaloGANV2`, selecting the most appropriate algorithm based on the properties of the shower-initiating particles, as can be seen in Figure 2 left. GEANT4 continues to be used for simulating particles in the inner detector, for muons, and for very low-energy hadrons in the calorimeters. High-energy hadrons, which may interact late or not at all in the calorimeter, can create a spray of secondary hadrons that reach the muon spectrometer - a phenomenon known as punch-through, which is now modelled based on a deep neural network (DNN) tool.

The computational performance of ATLFAST3 was evaluated across six different physics processes. ATLFAST3 demonstrates significant speed advantages over GEANT4, being 3 to 15 times faster depending on the specific process (Figure 2 right). The most substantial improvements are observed in processes involving the highest energy particles.

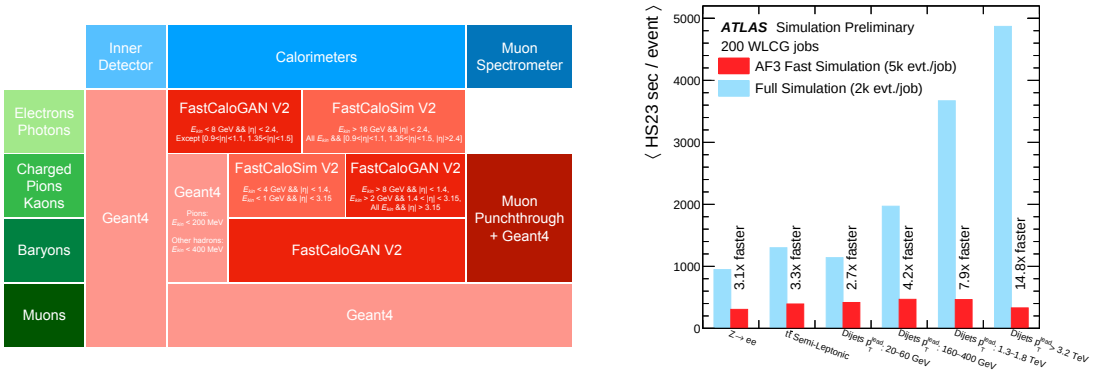

Figure 2: The collection of tools that constitute ATLFAST3 (left) and the mean CPU time per event for Fast Calorimeter Simulation (ATLFAST3) and Full Simulation (GEANT4) measured in standardized HS23 seconds (right). Taken from [5] and [6].

The physics performance of ATLFAST3 is assessed by comparing the modelling of reconstructed quantities and key kinematic variables between ATLFAST3 and GEANT4. Accurate modelling is achieved for various metrics, including the number of constituents for the leading jet (Figure 3 left), and variables commonly used in jet-tagging algorithms, such as the energy-correlation-function ratio D2 (Figure 3 right). Overall, ATLFAST3 and GEANT4 agree within a few percent for most observables used in physics analyses. Therefore, ATLFAST3 is suitable for a wide range of analyses, including both signal and background studies.

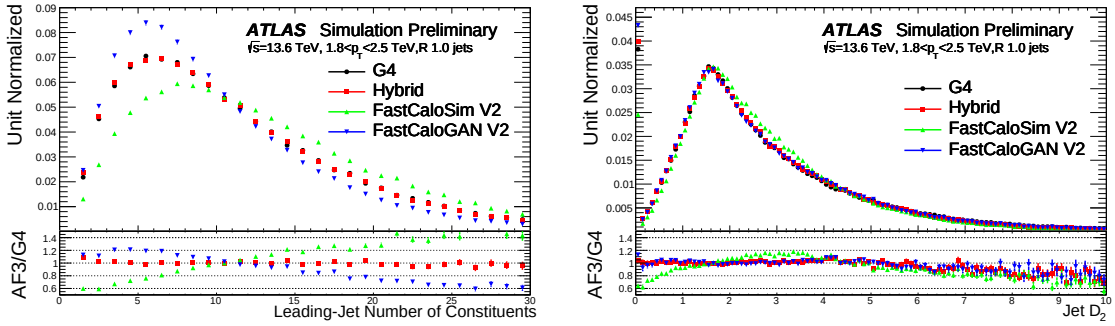

Figure 3: The number of constituents for the leading reconstructed jet (left) and D2 variable (right) in di-jet events with $1.8 < p_{\mathrm{T}} < 2.5$ TeV are compared across different simulation methods. The results are shown for samples simulated with GEANT4 (black circles), `FastCaloSimV2` (green upward-pointing triangles), `FastCaloGANV2` (blue downward-pointing triangles), and ATLFAST3 (red squares). Taken from [7].

3. Fast Track Simulation

When using ATLFAST3, most of the computational time is consumed by simulating the inner detector with GEANT4. To address this, efforts are focused on optimising this part of the simulation. FATRAS (Fast ATLAS Track Simulation) [8] plays a key role in this optimisation by using a simplified detector geometry, where the material properties of detector volumes are projected onto layer surfaces. Additionally, fast algorithms are employed to parameterise material effects. Current studies show that FATRAS can reproduce GEANT4 results with about 10% accuracy (Figure 4). While good agreement, mostly within statistical uncertainties, is achieved for electromagnetic processes (Figure 4 left), FATRAS yields substantially better resolution, differing from GEANT4 by about 10% across the pseudorapidity range (Figure 4 right). This discrepancy arises from FATRAS's inability to simulate rare hadronic interactions that produce tracks with large impact parameters. Ongoing improvements aim to reduce this discrepancy to 1%.

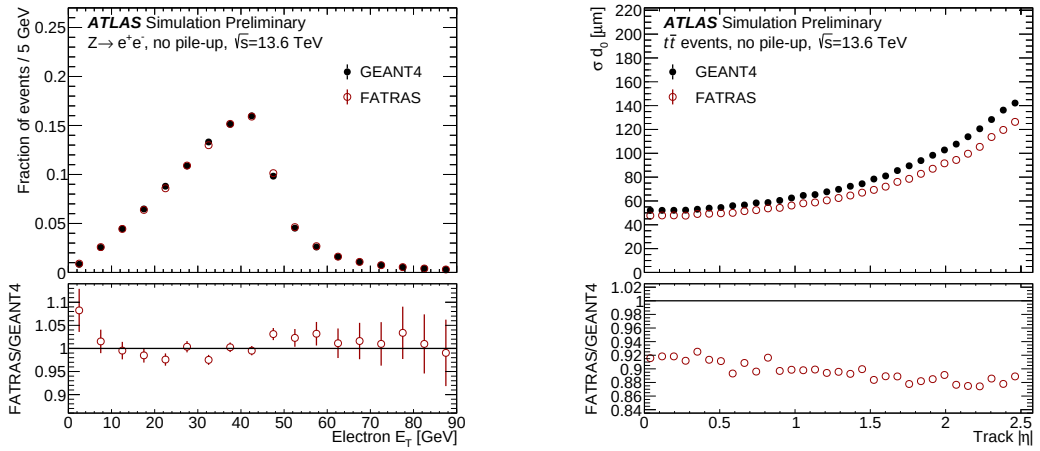

Figure 4: Transverse energy (E_{T}) of electrons (or positrons) in $Z \rightarrow ee$ events for tracking detector simulation (left) and transverse impact parameter (d_0) resolution of tracks as a function of pseudorapidity (η) in $t\bar{t}$ events (right) with FATRAS (open markers) and GEANT4 (full markers). Taken from [9].

Further progress is being made by integrating FATRAS into the ACTS (A Common Tracking Software) [10] framework. ACTS is an experiment-independent software package designed for particle reconstruction in high-energy physics. This integration is expected to make FATRAS thread-safe, enabling multi-threaded simulations across the entire ATLAS detector. At present, FATRAS combined with AtlFast3 is approximately three times faster (in semi-leptonic top-quark pair events) than AtlFast3 alone, as can be seen in Figure 5.

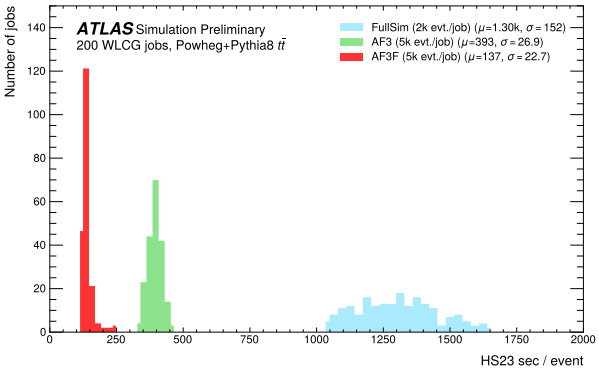

Figure 5: The mean CPU time per event for Fast Calorimeter Simulation (AtlFast3, abbreviated as AF3), Fast Calorimeter + Fast Tracking Simulation (AtlFast3 combined with FATRAS, referred to as AF3F) and Full Simulation (Geant4) measured in standardized HS23 seconds. Taken from [11].

4. Fast Reconstruction

The most CPU-intensive reconstruction algorithm is the inner detector track reconstruction, which identifies track candidates from tracking detector hits. This process is significantly slowed down by the presence of pile-up (PU) collisions. Currently, ATLAS uses the MC-overlay model, where PU collisions are simulated, digitised, and overlaid onto hard-scatter (HS) events during the digitisation phase. Track-overlay, a new method recently developed, is designed to address this issue by simulating, digitising, reconstructing, and overlaying PU collisions onto HS events during the reconstruction phase.

It has been shown that Track-overlay is suitable only in scenarios where hard-scatter (HS) track reconstruction is not influenced by pile-up (PU) hits, i.e., in processes with less dense environments. To address this limitation, a Deep Neural Network (DNN) has been developed to assess on an event-by-event basis whether Track-overlay can be applied, based on features such as the kinematics of generator-level particles, event topology (e.g., local track density), and PU information. After incorporating this ML-based decision, this PU model is referred to as Hybrid-overlay.

Preliminary results indicate that 86% of top quark pair events, 94% of QCD multijet events where the leading jet has $60 < p_{\mathrm{T}}^{\text{leading jet}} < 160$ GeV, and 35% of QCD multijet events with $1.8 < p_{\mathrm{T}}^{\text{leading jet}} < 2.5$ TeV could be processed with Track-overlay, while the remaining events are processed with MC-overlay. This approach results in negligible degradation in physics performance compared to MC-overlay, as shown in Figure 6, demonstrating that the Hybrid-overlay decision is working effectively. Moreover, Hybrid-overlay is expected to increase CPU efficiency in the inner detector reconstruction by a factor of approximately 1.8, which is promising. Therefore, Hybrid-overlay is nearing readiness for official MC production.

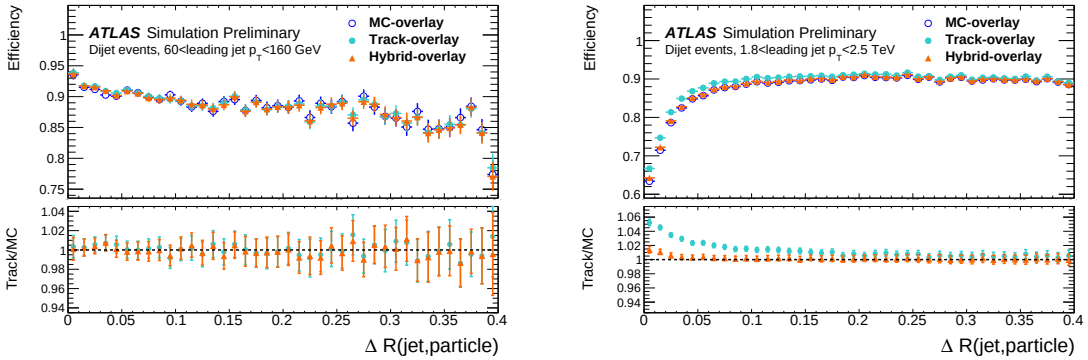

Figure 6: Track reconstruction efficiency as a function of the angular distance between the generator-level particle and the jet axis. This efficiency is examined for low-p_{T} QCD events (left) and high-p_{T} QCD events (right). All jets are constructed using EM-scale topological clusters and are reconstructed using the anti-k_t algorithm [12] with a radial distance parameter $R = 0.4$. Different PU models are compared, including MC-overlay (empty circles), Track-overlay (filled circles), and Hybrid-overlay (filled triangles). Taken from [13].

5. Fast Chain Workflow

With the availability of fast simulation tools, the focus now shifts to how these can be effectively integrated into large-scale production workflows. One promising approach is the Fast Chain workflow, which combines various simulation techniques, offering the flexibility to deploy various simulation scenarios tailored to the physics processes or reflecting current computational needs. This adaptable workflow not only optimises the simulation process but also enables the direct generation of outputs for physics analyses without the need to store intermediate files. As a result, the Fast Chain workflow can bring significant storage space savings and streamlines data handling in large production environments.

In addition to flexibility and storage efficiency, the Fast Chain workflow can deliver notable CPU savings. FATRAS, for example, is projected to save 1.2 MHS06 per year, potentially trading this for over 200 petabytes of tape space when skipping HITS[1] and re-running simulations annually. Track-overlay further is expected to contribute with 0.8 MHS06 per year, making the Fast Chain workflow a highly efficient solution for managing the simulation workload in the ATLAS experiment.

6. Conclusion

The Fast Simulation Program of ATLAS is designed to offer a more efficient alternative to the traditional MC production chain, optimising I/O and CPU resource management to meet the computational demands of future LHC runs. The AtlFast3 simulation, optimised for Run 3, provides high precision for various physics objects, achieving a 3-15 times speed-up in CPU performance, making it suitable for a broad range of physics analyses, including both signal and background studies. It is set to become the default simulation tool for the HL-LHC. FATRAS aims to further accelerate simulation times by approximately a factor of 3, with ongoing efforts to improve its

[1] A data format that records the detailed detector hits during particle interactions.

physics modelling performance. Additionally, the Hybrid-overlay method enhances CPU efficiency in inner detector reconstruction by about 1.8 times, with minimal impact on physics performance, is expected to be implemented later in Run 3. By integrating these advanced simulation tools into various Fast Chain workflow scenarios, ATLAS may be able to meet its resource requirements for CPU and storage in the future during the HL-LHC era, ensuring the continued success of its research endeavors.

References

[1] ATLAS collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*, 2008 JINST 3 S08003

[2] ATLAS collaboration, *ATLAS HL-LHC Computing Conceptual Design Report*, CERN-LHCC-2020-015, https://cds.cern.ch/record/2729668

[3] ATLAS collaboration, *ATLAS Software and Computing HL-LHC Roadmap*, CERN-LHCC-2022-005, https://cds.cern.ch/record/2802918

[4] ATLAS collaboration, *AtlFast3: The Next Generation of Fast Simulation in ATLAS*, Comput Softw Big Sci (2022) 6:7, arXiv:2109.02551 (2022)

[5] ATLAS collaboration, *Software and computing for Run 3 of the ATLAS experiment at the LHC*, arXiv:2404.06335 (2024)

[6] ATLAS collaboration, *CPU performance of ATLAS Fast Simulation (AF3) in Run 3*, https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/SIM-2023-005/

[7] ATLAS collaboration, *Performances of AtlFast3 for Run 3*, https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/SIM-2023-004/

[8] K. Edmonds, S. Fleischmann, T. Lenz, C. Magass, J. Mechnich, A. Salzburger, *The Fast ATLAS Track Simulation (FATRAS)*, ATL-SOFT-PUB-2008-001 (2008), https://cds.cern.ch/record/1091969

[9] ATLAS collaboration, *FATRAS Performance for ACAT 2024*, https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/SIM-2024-002/

[10] Ch. Gumpert, A. Salzburger, M. Kiehn, J. Hrdinka, N. Calace, *ACTS: from ATLAS software towards a common track reconstruction software*, J. Phys. Conf. Ser. **898** (2017) 042011

[11] ATLAS collaboration, *CPU Performance of Fast Simulations* , https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/SIMU-2024-06/

[12] M. Cacciari, G. P. Salam, G. Soyez, *The anti-k_t jet clustering algorithm*, JHEP04(2008)063, arXiv:0802.1189 (2008)

[13] ATLAS collaboration, *Track overlay validation for ACAT 2024*, https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/SIM-2024-001/