Fast simulation methods in ATLAS: from classical to generative models

John Chapman[1], *Kyle* Cranmer[2], *Stefan* Gadatsch[3], *Tobias* Golling[4], *Aishik* Ghosh[5], *Heather M.* Gray[6,7,*], *Tommaso* Lari[8], *Vincent R.* Pascuzzi[7], *John A.* Raine[4], *David* Rousseau[5], *Dalila* Salamani[4], and *Jana* Schaarschmidt[9]on behalf of the ATLAS Collaboration

[1]Cavendish Laboratory, University of Cambridge, Cambridge, United Kingdom
[2]Department of Physics, New York University, New York, NY, USA
[3]Simons Mobility Gmbh, Munich, Bavaria, Germany
[4]Facultè des Sciences, Département de Physique Nucléaire et Corpusculaire (DPNC), Université de Genève 24, Quai Ernest-Ansermet, CH-1211 Genève 4, Geneva, Switzerland
[5]IJCLab, Université Paris-Saclay, CNRS/IN2P3, 91405, Orsay,France
[6]Physics Department, University of California, Berkeley, CA, United States of America
[7]Physics Division, Lawrence Berkeley National Laboratory, Berkeley, CA, United States of America
[8]INFN Sezione di Milano; Milano, Italy
[9]Department of Physics, University of Washington, Seattle WA, United States of America

Abstract. The ATLAS physics program relies on very large samples of Geant4 simulated events, which provide a highly detailed and accurate simulation of the ATLAS detector. However, this accuracy comes with a high price in CPU, and the sensitivity of many physics analyses is already limited by the available Monte Carlo statistics and will be even more so in the future. Therefore, sophisticated fast simulation tools have been developed. In Run 3 we aim to replace the calorimeter shower simulation for most samples with a new parametrised description of longitudinal and lateral energy deposits, including machine learning approaches, to achieve a fast and accurate description. Looking further ahead, prototypes are being developed using cutting edge machine learning approaches to learn the appropriate calorimeter response, which are expected to improve modeling of correlations within showers. Two different approaches, using Variational Auto-Encoders (VAEs) or Generative Adversarial Networks (GANs), are trained to model the shower simulation. Additional fast simulation tools will replace the inner detector simulation, as well as digitization and reconstruction algorithms, achieving up to two orders of magnitude improvement in speed. In this talk, we will describe the new tools for fast production of simulated events and an exploratory analysis of the deep learning methods.

1 Introduction

The ATLAS physics program relies extensively on very large samples of Geant4 simulated events, which provide a highly detailed and accurate simulation of the ATLAS detector [1–3]. In addition to physics, these simulated events are also used for the design and optimization of the detector and trigger. The baseline simulation strategy [4] for the ATLAS experiment

[*] e-mail: heather.gray@berkeley.edu

uses the Geant4 simulation toolkit [5], which models the interactions of the particles with the matter in the detectors to a very high level of accuracy.

However, with time, as the dataset from the LHC grows in size and complexity the computing resources needed to produce these simulated samples continues to grow. In fact, the sensitivity of many physics analyses is already limited by the available Monte Carlo statistics. Figure 1 shows the projections for the CPU resources needed by the ATLAS experiment for data and simulation processing. The brown points show the amount of time needed based on existing software performance and uses the ATLAS computing model developed in 2017. The large jump in the requirements in 2026 corresponds to the anticipated start date of the High-Luminosity LHC (HL-LHC). The solid black line shows the amount of resources expected to be available under the assumption of a flat funding scenario. The clear discrepancy between the resources needed and the resources available demonstrates that reducing the amount of CPU needed by the ATLAS experiment is critical. At present, simulation takes approximately 34% of the total CPU time used by ATLAS, and 75% of the time in simulation is spent in the simulation of the calorimeter, which means that significantly improving the speed of the simulation can have a large impact on the total amount of CPU required. The blue points in Figure 1 show how much the total CPU could be reduced by extensively using the fast calorimeter simulation instead of Geant4 (blue down triangles), and by adding a fast method of tracking detector simulation and reconstruction (blue circles), and, finally, by increasing the speed of the generators by a factor of two (blue up triangles). Here we will outline the different fast simulation techniques currently under development by the ATLAS experiment, discuss recent improvements, and briefly review the current performance.

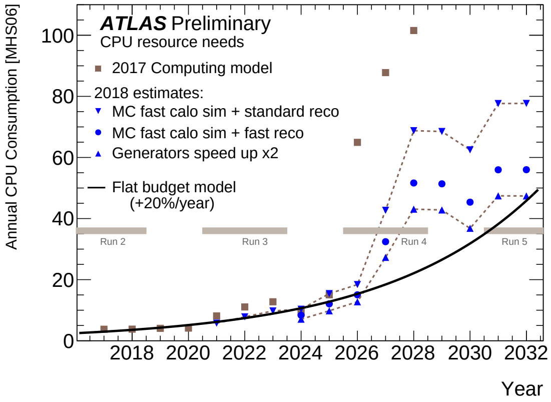


Figure 1. The estimated CPU resources needed by the ATLAS experiment for data and simulation processing. The brown points are estimates made in 2017, based on existing software performance estimates and using the ATLAS computing model parameters from 2017. The blue points show the improvements possible in three different scenarios: (1) top curve with the fast calorimeter simulation used for 75% of the Monte Carlo simulation; (2) middle curve using in addition a faster version of reconstruction, which is seeded by the event generator information for the tracks; (3) bottom curve, where the time spent in event generation is halved, either by software improvements or by re-using some of the events. The solid line shows the amount of resources expected to be available if a flat funding scenario is assumed, which implies an increase of 20% per year, based on the current technology trends. From Ref. [6].

2 Fast Calorimeter Simulation

The ATLAS detector is a multipurpose particle detector with a nearly 4π coverage in solid angle. It consists of an inner tracking detector surrounded by a thin superconducting solenoid providing a 2 T axial magnetic field, electromagnetic (EM) and hadronic calorimeters, and a muon spectrometer. The inner detector consists of silicon pixel, silicon microstrip, and transition radiation tracking detectors. The ATLAS electromagnetic calorimeter is a lead/liquid-argon sampling calorimeter with an accordion geometry. It consists of a barrel section covering the pseudorapidity region $|\eta| < 1.475$ and two endcap sections covering $1.375 < |\eta| < 3.2$. For $|\eta| < 2.5$, the EM calorimeters are segmented into three layers with different granularities. A thin presampler layer in front of the calorimeter covers $|\eta| < 1.8$. The ATLAS hadronic calorimeter consists of an iron/scintillator calorimeter for $|\eta| < 1.7$, two copper/liquid-argon calorimeters for $1.5 < |\eta| < 3.2$ and two tungsten/liquid argon forward calorimeters up to $|\eta| < 4.9$. The ATLAS calorimeters are used to reconstruct jets, photons, electrons, and in the determination of the missing transverse energy.

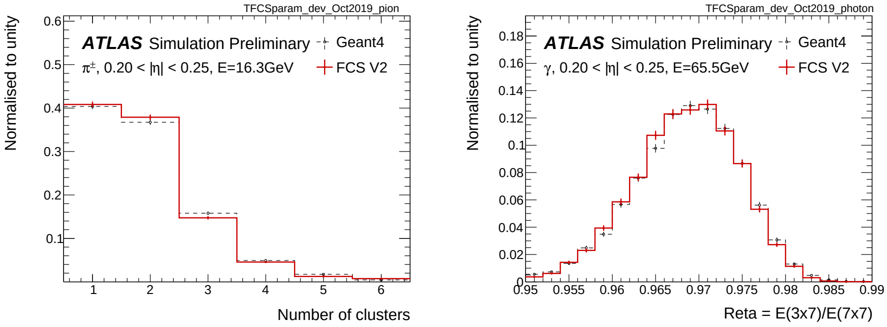
The complexity of the ATLAS calorimeter geometry with its accordion structure and varying cell sizes are the reason that the time to simulate a single event with the ATLAS detector is of the order of minutes. This is why fast simulation methods for the calorimeter have attracted significant attention. We will discuss the two methods under development by ATLAS that can be used to parametrise the calorimeter response. The first, known as FastCaloSim v2 or FCSv2 [7], uses principal component analysis (PCA), as discussed in Section 3 and the second uses neural networks, as discussed in Section 4. In both cases, the datasets used to derive these parametrisations are million of events containing single photons, electrons, and pions simulated using Geant4 with the simulation beginning at the surface of the calorimeter. To account for slight differences in the calorimeter response for positively and negatively charged particles, these are simulated separately.

3 Calorimeter Shower Parametrisation with Principal Component Analysis

The fast calorimeter simulation (FastCaloSim) has been used for fast simulation in ATLAS since Run 1 [8]. However, it is currently only used for a limited range of physics analyses due to limited performance in particular ranges of phase space, such as high-momentum jets or forward pseudorapidity. A new version of FastCaloSim, FastCaloSim v2, is currently under development with the goal of improving the physics performance to enable it to be used for an even wider range of analyses for Run 3. The parametrisation of the showers is derived separately for the amount of energy deposited in each calorimeter layer, which describes the longitudinal development of the shower, and the lateral shape of the shower in each layer. As there are large correlations between the energy deposits in the different layers, FastCaloSim relies heavily on PCA to convert the correlated input variables into a set of linearly uncorrelated variables by an orthogonal transformation of the coordinate system. The TPrincipal class from ROOT [9] is used to perform the PCA.

The FastCaloSim v2 parametrisation is derived in 17 logarithmically spaced energy bins ranging from 60 MeV to 4.2 TeV for photons and electrons and from 256 MeV for pions, and 100 uniform bins in $|\eta|$ ranging from 0 to 5 units of pseudorapidity. The inputs to the PCA are the fractional energy deposits in each layer of the calorimeter, as well as the total energy. The PCA is used to transform these into decorrelated inputs. The PCA component with the largest eigenvalue is divided into five equally populated bins and then the PCA chain is repeated on the showers within each bin.

The lateral shower shapes are modelled using two dimensional histograms binned in polar coordinates in the plane tangential to the calorimeter surface. These histograms are used as probability distribution functions and a certain number of hits are drawn to model the statistical fluctuations.



(a) (b)

Figure 2. (a) Number of clusters produced by a 16 GeV pion in the range $0.20 < |\eta| < 0.25$ in FastCaloSim v2 (red solid line) compared to Geant4 (black dashed line). (b) Fractional deposited energy in the η direction for the second barrel layer of the electromagnetic calorimeter for a 65 GeV photon reconstructed cluster in the range $0.20 < |\eta| < 0.25$. The 3x7 and 7x7 refers to the rectangle of cells considered around the cluster centre. FCSV2 (red solid line) is compared to Geant4 (black dashed line). In both cases, the FastCaloSim v2 parametrisation from October 2019 is used. From Ref. [10].

Recent improvements to FastCaloSim v2 include a new technique to derive the energy fluctuations for pions, by evaluating the stochastic and constant term from the Geant4 samples instead of those evaluated from beam tests. This results in a significant improvement in the modelling of the number of clusters and hence the description of jet substructure. The treatment of the cross-talk between the cells has also been improved, which leads to the fractional energy deposits to be better described for the photons (see Ref. [11] for further details). The current performance of FastCaloSim v2 is demonstrated in Figure 2. In both cases, the predictions from Geant4 (black dashed) are compared to those from FastCaloSim v2 (red). Figure 2 (a) shows that the number of energy clusters in the calorimeter for a 16 GeV pion is very well described, which depends critically on the accuracy to which fluctuations are modelled. Figure 2 (b) shows that the shape of the deposited energy in the η-direction is also very well-described for 65 GeV photons.

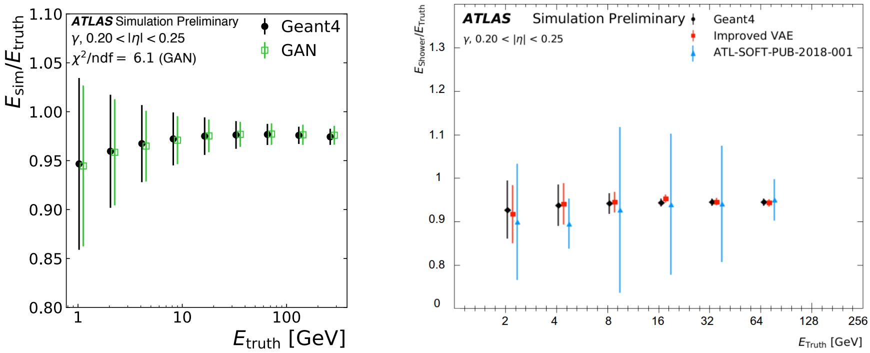
4 Calorimeter Shower Parametrisation with Neural Networks

Neural networks have proven themselves to be very powerful tools and are used extensively in reconstruction and analysis throughout high-energy particle physics. ATLAS is investigating whether deep neural networks can be used for simulation by training a network to approximate the showering from Geant4. This would provide an alternative parametrisation to FastCaloSim v2 to describe calorimeter showers. Two different approaches are being explored and both are based on unsupervised learning algorithms. The first uses Generative Adversarial Networks (GANs) [12] and the second uses Variational Auto-encoders (VAEs) [13, 14]. The inputs to the neural networks are the energy deposits in the calorimeter cells from single particle Geant4 events. At present, only photons in the central calorimeter barrel with $0.2 < |\eta| \leq 0.25$ have been studied.

The GANs consist of a generating neural network and a discriminating network, which tries to discriminate between the generated showers and those simulated by Geant4. After training, the generating network is used to simulate physics events. The VAE uses two stacked neural networks each containing four hidden layers. The first encodes the representation of the Geant4 showers into the latent space with reduced dimensionality. The second decodes the latent representation and is used to produce simulated events. See Ref. [15] for further details about initial design and architecture used for the GANs and the VAEs.

New and improved architectures for both sets of networks have been developed with respect to Ref. [15]. The GAN has been conditioned on the position of the incident particle and an additional discriminating network has been added to ensure that the total energy is well-modelled. In addition, the generator architecture has been optimised. This results in significant improvements to the description of the mean and the width of the showers. Figure 3 (a) compares the mean and width of the relative total energy distribution between the GAN and Geant4. Excellent agreement is demonstrated for all energy values. Good performance is also observed when interpolating between the energy points.

Recent improvements to the VAE include moving from cell energy to energy ratios to simplify the learning process for the network. In addition, five energy fractions are provided to the network in order to learn the correlation between the energies across the layers and the total energy. These fractions re-normalize the energies from the ratios. Weights were added to the reconstruction term of the loss function representing the importance of cell reconstruction with respect to the distance from the shower center. They are derived from the width of the energy ratio distributions. A weight per input feature is the inverse of the standard deviation. Figure 3 (b) compares the relative total energy distributions between the improved VAE, the previous VAE and Geant4. Previously the VAE significantly overestimated the width of the energy distribution, but now the mean and the width are in good agreement with Geant4.



(a) (b)

Figure 3. (a) Energy response of the calorimeter as a function of the true photon energy for particles with $0.20 < |\eta| < 0.25$. The calorimeter response for Geant4 is shown in black full markers used as reference and is compared to the one from the generative adversarial network (GAN), shown in green open markers. The GAN is shown with a small artificial shift towards the right for better visibility. The error bars indicate the resolution of the simulated energy deposits. From Ref. [16]. (b) Energy response of the calorimeter as function of the true photon energy for particles in the range of $0.20 < |\eta| < 0.25$ comparing Geant4 (black circles), the baseline VAE model (Ref. [15]) (cyan triangles) and the new improved VAE model (red squares). The shown error bars indicate the resolution of the simulated energy deposits. For clarity, a small offset is applied to the different simulation choices. From Ref. [17].

As both neural networks will produce parametrisations of the energy response similar to FCSv2, the CPU time when using the neural networks in simulation is expected to be comparable, but potentially with a reduced memory footprint.

5 Towards Even Faster Simulation

While the fast calorimeter simulation shows great promise, even greater gains in the CPU time are needed to simulate and reconstruct events, requiring additional techniques and new paradigms. The FastChain [18] is a generic fast simulation framework that incorporates FastCaloSim2, but provides additional fast simulation modules. The main target for improved speed is through the introduction of fast simulation (and reconstruction) techniques for the inner tracking detector and, to a lesser extent, the muon spectrometer. Two possibilities are currently being explored. The first, known as Fatras [19–21], simulates events using a simplified description of the detector geometry and parametrisations of physics processes. The second avoids both simulation and reconstruction of tracks by producing hits based on the true particle and fitting those to determine the track parameters. While the second is significantly faster, it is expected to have worse physics performance, as the impact of fake tracks is neglected.

Figure 4 shows the CPU time required to simulate 500 $t\bar{t}$ events in ATLAS using three different simulation techniques. The average time required using Geant4 is ~O(4 × 10^5) ms. With FastCaloSimv2, the time is reduced by more than an order of magnitude to ~O(3 × 10^4) ms. With FastChain, and using Fatras, the time is again reduced by more than an order of magnitude to ~O(10^3) ms.

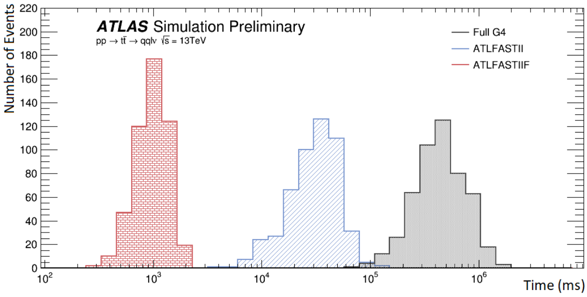


Figure 4. Comparison between Geant4 and two fast simulators, the fast calorimeter simulation (labelled as ATLFASTII) and the fast chain (labelled as ATLFASTIIF), in the CPU performance of event processing time. Simulations were performed in Athena release 21.3.8 on semi-leptonic $t\bar{t}$ events. Simulation benchmarks were performed using the BNL USATLAS Tier-3 Cluster which consists of 300 nodes, each with 8 2.6GHz CPUs and 16 GB of memory. 500 events were produced in a single run. No pile-up is simulated. From Ref. [22].

6 Conclusion

Large samples of Geant4 simulated events play a critical role within the ATLAS physics program, but their production requires large amounts of CPU. This is why the development of fast simulation techniques is critical for maintaining, and indeed increasing, physics performance, especially when looking ahead to the large datasets of the HL-LHC. We have reviewed the fast simulation methods currently under development by the ATLAS experiment. Two methods focusing on replacing the expensive calorimeter simulation with a parametrised response were presented with FastCaloSim v2 aiming to become the default simulation method for Run 3. In addition, FastChain seeks to speed up the simulation even further through new methods for the inner detector simulation and reconstruction. The methods presented are at various levels of development and highlights of recent progress were presented.

References

[1] ATLAS Collaboration, JINST **3**, S08003 (2008)

[2] ATLAS Collaboration, *ATLAS Insertable B-Layer Technical Design Report*, ATLAS-TDR-19 (2010), https://cds.cern.ch/record/1291633

[3] B. Abbott et al., JINST **13**, T05008 (2018), 1803.00844

[4] ATLAS Collaboration, Eur. Phys. J. C **70**, 823 (2010), 1005.4568

[5] S. Agostinelli et al. (GEANT4), Nucl. Instrum. Meth. A **506**, 250 (2003)

[6] ATLAS Collaboration, *Computing and Software Public Results*, https://twiki.cern.ch/twiki/bin/view/AtlasPublic/ComputingandSoftwarePublicResults

[7] ATLAS Collaboration, *The new Fast Calorimeter Simulation in ATLAS*, ATL-SOFT-PUB-2018-002 (2018), https://cds.cern.ch/record/2630434

[8] ATLAS Collaboration, *The simulation principle and performance of the ATLAS fast calorimeter simulation FastCaloSim*, ATL-PHYS-PUB-2010-013 (2010), https://cds.cern.ch/record/1300517

[9] R. Brun, F. Rademakers, Nucl. Instrum. Meth. A **389**, 81 (1997)

[10] ATLAS Collaboration, *Plots on Fast Simulation for NEC*, https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/SIM-2019-006/

[11] S.J. Gasiorowski (ATLAS Collaboration), Tech. Rep. ATL-SOFT-PROC-2020-027, CERN, Geneva (2020), https://cds.cern.ch/record/2712930

[12] I.J. Goodfellow et al., *Generative Adversarial Networks* (2014), 1406.2661

[13] D.J. Rezende, S. Mohamed, D. Wierstra, *Stochastic backpropagation and approximate inference in deep generative models* (2014), 1401.4082

[14] D.P. Kingma, M. Welling, *Auto-Encoding Variational Bayes* (2014), 1312.6114

[15] ATLAS Collaboration, *Deep generative models for fast shower simulation in ATLAS*, ATL-SOFT-PUB-2018-001 (2018), https://cds.cern.ch/record/2630433

[16] ATLAS Collaboration, *Energy resolution with a Generative Adversarial Network for Fast Shower Simulation in ATLAS*, https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/SIM-2019-004/

[17] ATLAS Collaboration, *VAE for photon shower simulation in ATLAS*, https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/SIM-2019-007/

[18] A. Basalaev, Z. Marshall (ATLAS), J. Phys. Conf. Ser. **898**, 042016 (2017)

[19] K. Edmonds et al., *The fast ATLAS track simulation (FATRAS)*, ATL-SOFT-PUB-2008-001 (2008), https://cds.cern.ch/record/1091969

[20] S. Hamilton et al., *The ATLAS Fast Track Simulation project (FATRAS)*, in *IEEE Nuclear Science Symposuim Medical Imaging Conference* (2010), pp. 311–316

[21] J. Mechnich et al., J. Phys. Conf. Ser. **331**, 032046 (2011)

[22] ATLAS Collaboration, *ATLAS Simulation CPU Performance*, https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/SIM-2019-002/