AI Denoising to Accelerate Detector Simulation

Principal Investigator: Kevin Pedro (FNAL)
Co-investigators: Javier Duarte (UCSD), Daniel Elvira (FNAL), Lindsey Gray (FNAL), Philip Harris (MIT), Scarlet Norberg (UPRM)

Introduction

The CMS detector simulation, which uses Geant4, consumed 40% of grid CPU at the beginning of Run 2 [1]. In the HL-LHC era, the CPU time to simulate an event will increase by a factor of 3 or more [2], because of the more complex detector geometry and the more detailed physics models needed to reproduce the precise measurements of the upgraded detector. The high level of physical accuracy in modern detector simulations is a crucial ingredient in high energy physics research, and its importance should not be understated. Statistical uncertainty arising from the limited sizes of simulated signal and background samples already poses a significant challenge for many searches and even some measurements. The experiment needs a solution that can preserve accuracy while decreasing CPU usage, in order to deliver simulated samples commensurate with the growing LHC datasets.

The CMS detector simulation already benefits from numerous technical optimizations and physics-preserving approximations, which improve its CPU efficiency by a factor of 4–6 compared to the default, as demonstrated by the PI [3]. Further improvements in the CPU usage of the full simulation, using existing computational techniques, are expected to be limited. This expectation is based on the concluded GeantV R&D project [4], whose integration in the CMS software was led by the PI [2], and recent in-depth profiling of Geant4 [5]. Therefore, the collaboration is motivated to search for more novel methods, such as artificial intelligence (AI), to optimize the detector simulation. The PI was appointed L3 convener of the CMS ML4Sim (machine learning for simulation) working group in mid-2020, in order to organize the various efforts related to this topic. The PI also serves as co-convener of the Detector Simulation working group in the HEP Software Foundation (HSF), in which role he gathers input from the entire HEP community on topics including ML for simulation.

Existing research into ML for simulation primarily utilizes generative adversarial networks. These networks are intended to learn the physics of the full simulation and then, given random input, produce physically valid output events. Though progress continues to be made at improving the results, the generative approach has several drawbacks. It requires very large samples of input data; the network training is not guaranteed to converge and may experience mode collapse; and the results may be invalid if input simulation parameters are extrapolated beyond the phase space of the training data.

Proposal

Denoising (related to inpainting and super-resolution techniques) is a promising and novel idea that directly addresses the drawbacks listed above and has not previously been employed

in high energy physics. It has been successfully used in industry to reduce the need for expensive Monte Carlo ray-tracing in computer animation [6]. This approach naturally avoids the reliability concerns that arise from attempting to generate wholly novel output. Instead, GEANT4 still simulates every event, using modified parameters to increase speed while reducing accuracy. The ML algorithm is trained by comparing the original detailed, high-accuracy simulated events to the lower-accuracy output. It thereby learns to "denoise" the latter, quickly producing a higher-quality final result. The fact that GEANT4 is still employed, in contrast to the fully generative approaches, provides a solid basis for the project to make significant progress in a one-year timescale. We target a speed improvement factor of 5–10 from the final GEANT4 parameter modifications.

The AI-based denoising approach is highly amenable to acceleration using coprocessors. Convolutional neural networks (CNNs) can be evaluated more than an order of magnitude faster using GPUs or FPGAs, compared to CPUs [7–9]. Therefore, this method will reduce the time spent on both the full simulation, which runs on CPU, and the ML inference step, which can run on heterogeneous resources. In particular, this presents a clear path to run detector simulation efficiently on the next generation of HPCs, which will rely heavily on GPUs. The capability to run inference on coprocessors within the CMS software is already available from the SONIC (Services for Optimized Network Inference on Coprocessors) project, developed by the PI [10]. The outcome of this project will apply SONIC to detector simulation for the first time in any HEP experiment.

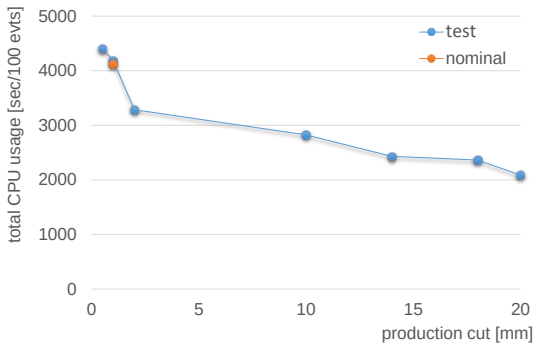

Figure 1: The computing time to process 100 $t\bar{t}$ events in GEANT4 for different values of the production cut.

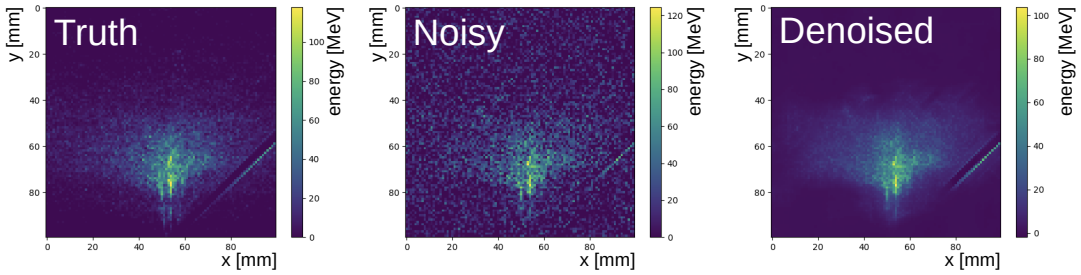

Figure 2: An example result from a denoising algorithm applied to a photon shower.

The project has four elements:

1. Determine GEANT4 parameter modifications;
2. Generate training data;
3. Design and train the denoising algorithm;
4. Implement the final product using SONIC in the CMS software.

Promising early results for these elements have been achieved. Figure 1 shows that modifying the production cut, a distance below which secondary particles are not produced, can decrease the GEANT4 computing time by a factor of two. In Fig. 2, the CNN architecture from Ref. [6] is applied to target the energy deposited in the CMS ECAL by photon showers as simulated in GEANT4, with noise added manually, producing a clearly denoised result.

The impacts of other modifications are being evaluated, such as parameters related to magnetic field integration and propagation, changes in the various physics models that are employed to simulate the interactions and energy losses of particles as they interact with the detector materials, and the thresholds and probabilities for the Russian Roulette algorithm that discards low-energy particles [11]. Neural network improvements are also being pursued, including a kernel-prediction architecture from Ref. [6], graph NNs, and neural architecture search [12]. The preliminary results above require several updates to become more realistic, including: moving from 2D to 3D images, varying the input particle parameters and subdetectors, a more realistic noise model, and including additional features, beyond the deposited energy, in the training data. The dedicated effort of a skilled postdoctoral researcher will be crucial to bring these elements together for a strong result. The successful candidate will have experience in machine learning and software development; additional experience in detector simulation, specifically calorimetry and particle showers, is preferred.

Timeline & Milestones

The postdoctoral researcher's contributions to the project are anticipated to begin in August 2021. Milestones for a full year of work are listed in Table 1 below. These deliverables build on the preliminary results discussed above.

The milestones listed in Table 1 are sufficient to produce results worthy of publication and to deliver a prototype of a viable solution to the computing challenges described in the introduction. The subsequent work to expand and improve the project will continue after the first year.

Mentoring & Supervision

The Fermilab CMS group will have several new postdoctoral associates starting in August 2021 who have some experience with AI and machine learning. Any of these new postdocs would be an ideal candidate for this project. In addition, there may be an opportunity to hire a postdoc specifically for this proposal.

Fermilab has an active detector simulation group in the Scientific Computing Division with significant expertise in GEANT4. There are also many AI experts at Fermilab and in the

Month	Milestone	Deliverables
3	Generate and validate training data for photon showers	• List of modified GEANT4 parameter settings • List of additional features • Presentation with CPU profiles of GEANT4 modifications and comparisons of physical variables
6	Initial neural network architecture	• Implementation of 3D kernel-prediction CNN • Optimized hyperparameters to achieve best performance
9	Implementation of prototype	• Integration in CMS software using SONIC • Publication of initial results in journal, including both physical and computational comparisons
12	Expansion of project scope	• Training data for different physical processes, detectors (e.g. HCAL, HGCal) • Exploration of more sophisticated ML techniques

Table 1: Timeline of milestones and deliverables for the proposal.

Chicago area (at Argonne, University of Chicago, etc.) who can provide useful feedback. The Fermilab group is one of the core institutions in the FastML (fastmachinelearning.org) research collective, which has pioneered the use of heterogeneous computing for AI acceleration in particle physics. As detector simulation is crucial to nearly every aspect of high energy physics research, there is significant community interest in overcoming the associated challenges posed by the HL-LHC. Accordingly, by partnering with institutions associated with the LHC Physics Center, we can leverage effort from graduate students or university postdocs based at Fermilab. This would provide the postdoc leading the effort with additional personpower and management experience.

The Fermilab CMS group has established a highly successful postdoc mentoring program that includes regular communication with both a direct supervisor and a mentor who looks out for the postdoc's broader career interests. Postdocs report on their work annually in a meeting attended by the entire group. The PIs on this proposal have experience supervising postdocs. Priority is given for postdocs to attend conferences and other major events, including HSF workshops, ensuring this work is presented to the broader community with high visibility. The biweekly CMS simulation and ML4Sim meetings will also provide opportunities to report on the progress and receive feedback from the collaboration. Given the key work in industry developing the techniques that form the basis for this project, direct communication with industry experts will also be facilitated as much as possible.

References

[1] HEP Software Foundation, "HEP Software Foundation Community White Paper Working Group - Detector Simulation", arXiv:1803.04165.

[2] CMS Collaboration, "Integration and Performance of New Technologies in the CMS Simulation", *EPJ Web Conf.* **245** (2020) 02020, doi:10.1051/epjconf/202024502020, arXiv:2004.02327.

[3] K. Pedro, "Current and future performance of the CMS simulation", *Eur. Phys. J Web Conf.* **214** (2019) 02036, doi:10.1051/epjconf/201921402036.

[4] G. Amadio et al., "GeantV: Results from the Prototype of Concurrent Vector Particle Transport Simulation in HEP", *Comput. Softw. Big Sci.* **5** (2021) 3, doi:10.1007/s41781-020-00048-6, arXiv:2005.00949.

[5] G. Amadio, "Geant4 code performance optimizations", 2019. https://indico.cern.ch/event/809405/contributions/3629814/.

[6] S. Bako et al., "Kernel-predicting convolutional networks for denoising Monte Carlo renderings", *ACM Trans. Graph.* **36** (2017) doi:10.1145/3072959.3073708.

[7] J. Duarte et al., "FPGA-accelerated machine learning inference as a service for particle physics computing", *Comp. Soft. Big Sci.* **3** (2019) 13, doi:10.1007/s41781-019-0027-2, arXiv:1904.08986.

[8] J. Krupa et al., "GPU coprocessors as a service for deep learning inference in high energy physics", arXiv:2007.10359. Accepted by *Mach. Learn.: Sci. Technol.*

[9] D. S. Rankin et al., "FPGAs-as-a-Service Toolkit (FaaST)", in *Proceedings of Sixth International Workshop on Heterogeneous High-performance Reconfigurable Computing (H2RC20)*. 2020. arXiv:2010.08556. doi:10.1109/H2RC51942.2020.00010.

[10] K. Pedro et al., "SonicCMS". [software], 2020. version 5.2.0. https://github.com/hls-fpga-machine-learning/SonicCMS.

[11] CMS Collaboration, "Upgrades for the CMS simulation", *J. Phys. Conf. Ser.* **608** (2015), no. 1, 012056, doi:10.1088/1742-6596/608/1/012056.

[12] M. F. Kasim et al., "Building high accuracy emulators for scientific simulations with deep neural architecture search", arXiv:2001.08055.