Simulation Project

Alberto Ribon
CERN EP-SFT

LHCC meeting, CERN, 19 November 2024

Outline

Two parts:

1. **Preamble**
   - Approach, challenges

2. **Status**
   - Towards Geant4 version 11.3
     - The main topic
   - Outlook for next year and beyond
     - Short and intermixed with the main topic

Preamble

Goal, Direction, Strategy

- Goal: provide **better** Geant4 detector **simulations**, in all aspects
  - More precise; faster; smaller (in memory); more robust; easier to use; capable of more things; covering more use-cases
  - Exploit advances in both computing technologies and physics modeling
- Direction: priorities driven by **user needs**
  - For the EP-SFT Simulation team, the top priority is the support of CERN experiments and projects – in particular the LHC experiments
- Strategy for physics development: calibrate on **thin**-target data, validate on **thick**-target
  - Development of physics models is meant to reduce the discrepancies between simulation and thin-target data
  - Simulation of electromagnetic and hadronic showers in calorimeters is the most important and challenging aspect (for both physics and computing) for HEP applications

Accuracy *vs* Speed

- Effort for improving the accuracy of simulations, regardless of the speed, is always worth, even if not used in production
  - To investigate disagreements between simulation and experimental data
  - To tune or train fast simulation approaches
- Ways for improving the computing performances of simulations
  - By reviewing implementations, algorithms and approaches
    - Without compromising the physics accuracy
  - By using fast simulation solutions, *i.e.* simplifications with reduced accuracy
    - Many approximations are possible, matching different requirements and use-cases
    - Traditionally, based on shower parameterisations;
      now, growing interest and applications of ML-based solutions
  - By exploiting new technologies
    - *E.g.* GPU accelerators

Parallelism opportunities in detector simulations

- Run: a large number of similar and independent events
  - This is the “embarrassingly parallel” part of Monte Carlo simulation
  - **Event-level parallelism** : already solved by multi-threading (G4 10.0 in 2013)
- Event: a large number of similar but not fully independent tracks
  - They are independent only once they are created, but they are created randomly during the transport of previously generated particles
  - **Track-level parallelism** : difficult, on-going effort in the last 10+ years
- Challenges of track-level parallelism for generic HEP simulations
  - Stochastic nature of Monte Carlo methods
  - Very complex geometries
  - Hundred different particle types, with energies from TeV down to at least keV
  - Several dozens of different physics models, with non-trivial algorithms of various complexity
  - Many physics tables to access randomly at run-time
  - Event reproducibility (essential for debugging)

GPU accelerators for HEP simulations

More on this later – AdePT & Celeritas

- There are some HEP software applications (*e.g.* reconstruction) that have demonstrated substantial speed-up on GPUs
- Also some specialized, low-energy Monte Carlo particle transport simulations have successfully achieved large speed-up on GPUs
  - Optical photons; low-energy neutrons in nuclear reactors;
    low-energy electron and gamma in voxelized water phantoms (medical applications)
- Can GPUs - in a hybrid CPU-GPU workflow - speed-up HEP simulations?
  - Not obvious, and not by large factors – but worth to try out!
    - Especially with the growing presence of GPU : there is also energy efficiency to consider, besides throughput
  - The above low-energy simulations, where GPU works well, have all essentially one particle type to transport, with a simple physics and no or few secondaries, and simple geometries, so there are essentially independent primary tracks undergoing relatively few branches
  - Generic HEP simulations are completely different: many more secondaries than primary particles, of many different types and energies, with a rich variety of physics models, in complex geometries, therefore high thread divergence seems inevitable…
    - GPU might potentially bring some benefit in the simulation of EM showers in calorimeters, but not large factors in the overall simulation

Status : G4 11.3 and beyond

Geometry

- VecGeom : towards version 2
  - Main development is focused on the new surface modeler targeting GPUs
  - Once fully validated, the (portable) surface model will be the model used on GPUs, while the solid model will be kept only for CPU (removing its CUDA part)
- Parallelised the geometry initialisation (voxelisation)
  - Parallelised the voxel optimisation over logical volumes, using threads/tasks
  - Requested by CMS; first version included in 11.3.beta release (June 2024)
- Revision of *G4GenericTrap*
  - Main Geant4 solid used to replace the ATLAS custom solid for the EMEC
- Added new classes for field configuration
  - New UI commands to allow users to configure field transportation
    - Originally developed in ALICE

Kernel

- First prototype of task-based sub-event level parallelism
  - Event split in sub-events, with automatic merging of the hits at the end of the event
    - In “Phase I” (*i.e.* for the coming release G4 11.3): all tasks have the same physics processes and see the same detector geometry
      - Useful for large events, *e.g.* heavy-ion collisions
    - In “Phase II” (*i.e.* after G4 11.3): each task has only the necessary physics processes and see the limited detector geometry which are needed for that particular task
      - Useful for heterogeneous simulation

Note: the threading model of Geant4 has been stable since its introduction with G4 10.0 (2013), from sequential to multi-threading (always keeping backward-compatibility with pure sequential mode) and with the addition of tasking in G4 11.0 (2021), as an alternative parallelism approach.
The introduction of sub-event parallelism is not going to change the threading model, and will be optional.

Electromagnetic physics

- Revised initialisation of EM tables and data structures
  - To make them thread safe, as a necessary condition towards the parallelisation of the initialisation of physics, expected for next year (G4 11.4)
- EM physics for transporting exotic charged particles
  - As required by ATLAS and LHCb, ionisation and multiple scattering (as well transportation in magnetic field and decay) are included for charged particles with well defined PDG code but unknown to Geant4
- Positron annihilation into 3 gammas, and positronium production and decay
  - In-flight positron annihilation is relevant for HEP: it may affect EM shower shapes at the per-mille level
  - Creation of positronium at-rest is relevant for medical applications (PET)
- Extension of models and examples for channeling

- On-going extensions
  - To provide the full functionality of native Geant4 electromagnetic physics
    - Configuration per detector region (*e.g.* multiple scattering for CMS)
    - Gamma / electron / positron – nuclear interactions
  - To gain extra computing performances
    - Data re-structuring (*e.g.* of the macroscopic cross sections)
    - “General Process”-like handling of macroscopic cross sections
    - Woodcock tracking for gamma particles
  - To maintain the GPU support of the library for AdePT

Hadronic physics

- Improved simulation of low-energy neutrons
  - Included the possibility (off by default) to treat the Unresolved Resonance Region (URR) by Particle Tables (PT) for low-energy (< 20 MeV) neutrons, relevant for more precise simulations of nuclear reactor criticality and shielding applications
    - Making Geant4 another step closer to MCNP and TRIPOLI (reference codes for neutronics)
- New alternative nuclear de-excitation model (NuDEX)
  - Optional, more sophisticated model, useful for precise simulations of nuclear reactions in particular concerning the emissions of gammas and internal conversion electrons
- New hadronic datasets
  - Updated with latest ENSDF (Evaluated Nuclear Structure Data File) data from March 2024
    - More consistent treatment of nuclides with incomplete information, and with fewer (hopefully none) unphysical nuclear states

Fluka-Cern and Geant4 integration

- Since G4 11.2 (2023), interface to Fortran Fluka-Cern available for Geant4 applications to get inelastic cross sections and final states
  - Understood this year some disagreements between Geant4 and Fluka-Cern
    - Due to the different **quasi-elastic** treatment: considered as *inelastic* (for Geant4) vs. *elastic* (for Fluka)
  - Under discussion possible extensions (elastic, ion-ion, gamma/lepton-nuclear, *etc.*)
- Areas of common interest − where resources could be shared
  - Physics validation
  - Hadronic datasets
  - Low-energy neutrons
  - Very high-energy hadronic models (above a few TeV in the Lab frame)
- Progress towards Fluka-Cern 5 (C++, compatible with Geant4)
  - Ready new C++ point-wise low-energy neutron treatment
  - On-going migration of the nuclear de-excitation models

geant-val.cern.ch

- Key validation tool for Geant4
  - *geant-val* is essential for validating Geant4 simulations across multiple areas, including electromagnetic and hadronic physics
- Sustained development efforts
  - Continuous work is underway to maintain, improve, and run *geant-val*
    - The project currently depends on a single maintainer, so ensuring the continuation of this activity is crucial
- Recent test integrations
  - CMS HGCal calorimeter test-beam
    - This test expands *geant-val*'s reach in calorimeter validation, as well as supports GPU R&D initiatives AdePT and Celeritas
  - Low-energy neutron benchmarks
    - New tests are being integrated, primarily through comparisons with the reference codes MCNP and TRIPOLI
    - These demonstrate the value of non-experimental tests

ML fast simulation (1/2)

- Significant progress in generative ML models for fast simulation
  - Used in production by ATLAS in Run 3 (based on GAN)
  - Steady progress by CMS and LHCb
- Community developments
  - Many developments in fast simulation have been experiment specific
  - **CaloChallenge** (2022) provided a set of common datasets and benchmarks to enable the comparison of various ML models in an experiment-neutral way
  - In preparation for the next community challenge: **Open Data Detector** (ODD)
- *Par04* example in Geant4
  - Example (introduced in G4 11.0) to demonstrate how to use the ML inference to create energy deposits as a fast simulation model using **ONNX** runtime, **LibTorch** and LWTNN
    - In G4 11.2, added the possibility to run the inference on GPUs with ONNX runtime
  - Two out of the three datasets of CaloChallenge were produced with Geant4 *Par04*

ML fast simulation (2/2)

- R&D efforts: towards more generic fast simulation ML models
  - Aim to reduce the computational resources required for developing ML fast sim models
  - Idea: explore a "**foundation model**" approach
    - Train the model once on a large dataset, consisting of several different detector geometries
    - Provide it to users for fast adaptation to specific use case
  - **CaloDiT** : promising generative diffusion model with transformers
    - On-going collaborative development by EP-SFT, OpenLab and IBM research
    - Good performance in terms of physics observables
    - Adapting to new geometry faster than training from scratch
    - Slow: investigating ways to reduce the number of diffusion steps (distillation to speed-up)
- Directly collaborating with experiments (ATLAS, LHCb, Future Colliders, ...)
  - Towards the next generation of fast calorimeter simulation (AtlFast4)
  - Detailed physics validation of CaloChallenge-like geometries within Gaussino framework
  - Common library (DDFastShowerML in Key4hep) and physics benchmarks

AdePT (**A**ccelerator **d**emonstrator of **e**lectromagnetic **P**article **T**ransport)

- Progress on the surface model
  - **All solids relevant for LHC are supported**
    - CMS (2018, 2026), CMS HGCal test-beam, LHCb ECal & HCal, LHCb upgrade, ATLAS EMEC, *etc*.
  - Treatment of geometry overlaps (which are more problematic for surface model)
  - Surface model still slightly slower than the solid model
    - On-going optimization of relocation, safety calculations and investigation of mixed precision
- Various developments
  - Common interface between AdePT and Celeritas for integration in Geant4 applications and frameworks
  - Refactoring of AdePT into a library
  - New method of integration on Geant4 (based on specialised tracking manager), and new scoring
  - Asynchronous kernel scheduling
- Integration with experiments
  - Recent hackathon dedicated to the integration of AdePT & Celeritas in the ATLAS software framework
  - On-going integration of AdePT in LHCb Gaussino
- Geant4 delta-assessment of AdePT & Celeritas postponed to March 2025
  - Check-point of these projects, likely the last one by Geant4

Celeritas

- Goal: offloading to GPU the simulation of electrons, gammas and positrons
  - Using VecGeom (with ORANGE as alternative surface-based geometry model)
  - Their own implementation of EM physics on GPU (future option to use G4HepEm)
  - Offering a common interface (as AdePT) for using inside experiments' sw frameworks
  - Reconstructed "hits" sent back to CPU to be used in "sensitive detector" code
- Benchmarks in several detector set-ups and hardware
  - ATLAS FullSim Light, ATLAS Tile calorimeter test beam, CMS Run 2 & 3, CMS HGCal
  - ¼ of a Perlmutter (NERSC) GPU node (16 cores of AMD EPYC 7763, 1 Nvidia A100)
- Draw positive conclusions and have an ambitious program for the future
  - "*For realistic HEP test problems, Celeritas offers … 2 x of better overall speedup, 2 x energy efficiency*" (CHEP 2024)
  - "*High-performance surface-based geometry; Muon physics and µCF models; Platform-portable optical physics; Neutron physics; …*" (CHEP 2024)