Simulation CWP : Executive Summary

In the following we describe the main challenges and opportunities to be faced over the next decade for Full and Fast simulation applications of HEP experiments that include both the modeling of detectors and beamlines, and then establish a plan to address them. The scope of the topics covered includes the main components of a HEP simulation application, which are MC truth handling, geometry modeling, particle propagation in materials and fields, physics modeling of the interactions of particles with matter, the treatment of pileup and other backgrounds, as well as signal processing and digitisation.

Challenges

The experimental programmes planned in the next decade are driving developments in the simulation domain; they include the High Luminosity LHC project (HL-LHC), neutrino and muon experiments, and studies towards future colliders such as the Future Circular Collider (FCC) and the Compact Linear Collider (CLIC). The requirement of improving precision in the simulation implies production of larger Monte Carlo samples, which scale with the number of real events recorded in future experiments, and this places an additional burden on the computing resources that will be needed to generate them. The diversification of the physics programmes also requires new and improved physics models.

The widely used detector simulation toolkit, Geant4, is at the core of simulation in almost every HEP experiment. Its continuous development, maintenance, and support to the experiments is of vital importance and needs to be strengthened. Programs to develop future incarnations of Geant4 target the HL-LHC experiments to start not earlier than 2026.

The main R&D challenges to be addressed if the required physics and software performance goals are to be achieved can be summarised as follows:

- reviewing the physics models assumptions, approximations and limitations in order to achieve higher precision, and to extend the validity of models up to FCC energies on the order of 100 TeV;
- redesigning, developing, commissioning detector simulation toolkits to be more efficient when executed on emerging computing architectures e.g. Single Instruction Multiple Data (SIMD) vectorisation, Non-uniform Memory Access (NUMA) hierarchies, Many Integrated Core (MIC) devices, Graphic Processing Unit (GPU) and Tensor Processing Unit (TPU) architectures;
- porting and optimising the experiment’s simulation applications, or developing new ones to allow exploitation of High Performance Computing (HPC) facilities;
- exploring different fast simulation options, including common frameworks for fast tuning and validation;

- developing, improving, optimising geometry tools that can be shared among experiments to make the modeling of complex detectors computationally more efficient, modular, and transparent;
- developing techniques for background modeling, including contributions of multiple hard interactions overlapping the event of interest in collider experiments (pile-up);
- revisiting digitisation algorithms to improve performance by means of vectorisation and sub-system parallelisation techniques, and exploring opportunities for code sharing among experiments;
- recruiting, training, retaining, human resources in all areas of expertise pertaining to the simulation domain, including software and physics.

Strategy

An important requirement for the future organisation of the work programme is the need to support on-going experiments through continuous maintenance and improvement of the Geant4 simulation toolkit. New or refined functionality continues to be delivered in the on-going development programme both in physics coverage and accuracy, whilst introducing software performance improvements whenever possible. Such improvements are already included in the most recent Geant4 release and are being evaluated by experiment collaborations. In addition the Geant4 collaboration is working closely with user communities to enrich the physics models validation system with data acquired during physics runs and test beam campaigns. The Geant4 simulation toolkit will continue to evolve over the next decade. This evolution may include contributions from various R&D projects that aim to exploit various new techniques in order to extend physics reach and software performance.

Improving the speed of simulation codes by an order of magnitude represents a huge challenge. The gains that can be made by speeding up critical elements of Geant4 can be leveraged for all applications that use it and therefore it is well worth the investment in manpower needed to achieve it. An ambitious R&D programme is underway to investigate ways of improving the performance of all components of the simulation software for the longer term. One of the most ambitious elements of this programme is a new approach to managing particle transport, which has been introduced by the GeantV project. The aim is to deliver a multi-threaded vectorised transport engine that has the potential to deliver large performance benefits. Its main feature is track-level parallelisation, bundling particles with similar properties from different events to process them in a single thread. This approach, combined with SIMD vectorisation coding techniques and use of data locality, is expected to yield significant speed-ups, which are being measured in a realistic prototype that is under development.

At the same time that this new transport engine is being developed, work is also on-going to exploit parallelisation techniques to improve the performance of the accompanying modules, including geometry, navigation, and the physics models. They are developed as independent modules in such a way that they can also be used together with the current Geant4 transport engine. Of course when used with Geant4 they will not expose their full performance potential, since transport in Geant4 is sequential, but this allows their full validation and comparison with the existing implementations. The benefit of this approach is that new developments can be delivered as soon as they are available and validated in the existing framework. There are recent examples which have demonstrated that this approach is successful, such as the introduction of the new vectorised geometry package (VecGeom).

The work on Fast Simulation is also accelerating with a view to producing a flexible framework that permits Full and Fast simulation to be combined for different particles in the same event. Various approaches to Fast Simulation are being tried all with the same goal of saving computing time under the assumption that it is possible to improve time performance without an unacceptable loss of physics accuracy.

It is obviously of critical importance that the whole community of scientists working in the simulation domain continue to work together in as efficient way as possible in order to deliver the required improvements. Very specific expertise is required across all simulation domains, such as the physics modeling, tracking through complex geometries and magnetic fields, and building realistic applications that accurately simulate highly complex detectors. Continuous support is needed to recruit, train, and retain the personpower with the unique set of skills needed to guarantee the development, maintenance, and support of simulation codes over the long timeframes foreseen in the HEP experimental programme.

Workplan

The full work plan for 2017 for Geant4 has been defined together with user communities[1]. Important items include plans for significant improvements that facilitate running on HPC systems, as well as significant updates to the electromagnetic (EM) and hadronic physics models. Assuming manpower levels can be maintained, it is foreseen to deliver the following within three years.

- A full set of neutrino interactions that includes the primary neutrino weak interaction with the nucleus simulated using the GENIE generator, followed by the full hadronization process using Geant4 with efficient event biasing options.
- Muonic atom and related physics models.
- Specialized EM physics models for liquid noble gases, such as liquid Xe and liquid Ar (currently known as NEST), fully integrated into Geant4.
- Revision and improvement of the high energy quark-gluon string model (QGS).
- New biasing options including cross-section biasing for charged particles.

[1] Geant4 workplan : http://geant4.cern.ch/support/planned_features.shtml

- First release of a thermal neutron model that could be spawned to run on GPU-based co-processors (hybrid simulation).
- Data distribution and reduction in Geant4-based HEP detector simulations through combination of MPI and MT for efficient execution on HPC systems.

In the longer term, within 5 years, it is intended to investigate alternative models for very high energy hadronic interactions that will be of particular interest for modeling cosmic ray experiments and for use in FCC studies. The EPOS (**E**nergy conserving multiple scattering **P**artons, parton ladders, and strings **O**ff-shell remnants **S**aturation) hadronic generator offers currently the most accurate simulations in this domain, and it is therefore envisaged to make a fresh C++ rewrite of the model. Additional work will be invested in the further improvement of the QGS string model, which is a more theory-based approach than the phenomenological FTF model, and therefore offers better confidence at high energies i.e. up to a few TeV. For the intra-nuclear cascade models, while Bertini and Binary are likely to remain stable, further development is expected for the INCL++ model.

On the kernel and geometry side, the adoption of C++11/14/17 is one of the highest priorities. Much of the Geant4 code was written almost 20 years ago without modern math functions or advanced compiler optimizers. A pilot project has already been applied to some solid shapes and resulted in up to a factor of two speed-up so that these solids are now faster than any alternative solid models. Such code modernization should be applied to kernel and geometry first, followed by the most frequently used physics models. Another large design revision is particle transportation. Currently particle transportation in detector geometries and the integration of electromagnetic fields are handled by a single class regardless of particle type. Splitting this transportation class into several dedicated classes optimized for particle types would significantly reduce performance overhead. This splitting also requires considerable revision in the Geant4 kernel that currently assumes a single transportation class. This kernel revision has to be made without changes to the APIs which experiments currently use. The third target is the optimization of data reduction in Geant4-based HEP detector simulations on HPC. This is particularly demanded by ATLAS. Geant4 has already demonstrated almost perfect linearity when using the whole three million threads of ANL Mira, but efficient reduction of simulation results for such a large number of threads is really a challenge.

Research and Development

The plan of work and deliverables are staged in two phases. The *alpha* release will be made available by the end of 2017, serving as a preview of the new particle transport engine and demonstrating many of its features. The *beta* release will be made available at the end of 2018 and is expected to contain enough functionality to build first real applications. This will allow performance to be measured and give sufficient time to prepare for HL-LHC running.

The *alpha* release will deliver comprehensive electromagnetic shower simulation for electrons, positrons and gammas in scalar mode, with most of the physics models, the geometry and the magnetic field propagation optimised for multi-particle vectorisation. The user interfaces will be fully operational, with an extended set of examples covering the most important use cases, such as trackers and calorimeters. This release will be integrated in parallel task-based workflows steered by experiment data processing frameworks, such as CMSSW and GAUDI. It will allow the new particle transport engine to be integrated with the experiments' frameworks and preliminary performance measurements to be made.

The *beta* release will provide new features and optimisations in terms of geometry and electromagnetic physics, a first version of a hadronic physics package, and input/output (I/O) and support for fast simulation. Most EM physics models, the geometry and the propagation in fields will be vectorised, while the global single thread performance should be close to the design goal of a factor of 2-5 speed-up as compared to Geant4. Milestones associated with the *beta* release are:

- Performance optimisation of the kernel to percolate multi-particle vector optimisations down to the compute-intensive physics models. Other benefits will be lower contention from better data handling and a significantly lower memory footprint.
- Complete physics modeling for electrons, gammas and positrons. Scalar versions of these improved physics models will be backported to Geant4 to allow the community to profit as soon as possible from components developed in the context of the R&D.
- Production-quality vectorised geometry.
- High Performance Bertini hadronic interaction package. Glauber-Gribov cross sections plus low-energy parameterisations, elastic scattering.
- Prototype interface for user/experiment scoring and actions, supported by several examples. Vectorised transport in electromagnetic fields
- Simulation reproducibility in a fine-grain multithreaded environment.
- Realistic demonstrator for efficient MC truth information usage.
- HPC demonstrators on many core architectures and General Purpose GPU (GPGPU), including workload balancing and optimisation.
- Demonstrators for efficient concurrent I/O and multi-event user data management.
- Testing, validation, and performance suite for comparison with the Geant4 benchmark.

Physics Models

Physics models are continuously being reviewed and in some cases reimplemented in order to improve accuracy and software performance, and to ensure long term maintainability. Discrepancies between measurements and simulation have been observed in detector response linearity and energy resolution for some particles, as well as in electromagnetic and hadronic shower shapes. For example, in ATLAS shower shapes in the eta-direction are consistently wider in data than simulation, both for electrons and photons, whilst in CMS discrepancies exist in the EM shower width in the endcap region. Investigations need to be made to understand whether their cause can be attributed to a problem in the underlying physics model or to some other cause, such as an issue with the modelling of detector materials. More details can be found in the full report.

The ongoing work programme foresees a number of improvements being made to the physics models. Within a timescale of 3 years a new implementation of one full set of hadronic physics models for the full LHC energy range and improved physics for liquid Argon detectors will be released that is both performant and production quality. In addition, a robust and high-physics-quality implementation of an alternative set for cascade and string hadronic models will be made, potentially at a substantial additional computing cost, i.e. 1.5-3 times slower with respect to the default physics list. This alternative list will include a full set of EM physics processes, including de-excitation options and alternative implementations for angular distributions at low energies, enabling detailed detector and physics simulation for high sensitivity studies. It will also be possible to blend full physics and fast (ML-based) simulation for systematic studies and production. To address the needs of cosmic frontier experiments optical photon transport must be improved and made faster.

An improved implementation of hadronic cascade and string models with a modular design will be released within 5 years. For example a cascade could be assembled choosing between sub-modules for cross-sections in nuclear matter, different models of nuclei (discrete nucleons or shells of different densities) and types of hadron-nucleon interactions (producing only two or any number of secondaries). At the same time, the full set of existing cascade and string models will be recast in this redesigned modular framework, which will include a documented set of model parameters to enable tuning.

Experiment Applications

In the first year, the LHC experiments will collaborate in a computing performance assessment and monitoring programme that includes an “apples-to-apples” comparison of the Geant4 module of their simulation application with the goal to understand needed resources in the HL-LHC era. Moreover, ATLAS and CMS will run their simulation application in multithreaded mode, benefitting from significant memory savings. The Neutrino and Muon experiments will perform computing performance tests to quantify more accurately their computing needs for running simulation.

The release of the VecGeom geometry package, the first deliverable of the R&D programme and compatible with Geant4 in scalar mode, offers the experiments a way to improve the time performance of their detector simulation applications and work is underway to evaluate its impact. Experiments will continue their on-going efforts to port their CPU intensive detector simulation applications to run on high performance computing systems (HPC).

After 3 years, most experiments will have acquired the ability to run their Geant4 detector simulation modules in multithreaded mode. The release of a revisited version of navigation and electromagnetic physics packages, compatible with Geant4 in scalar mode and optimised for GeantV in vector mode, will offer the experiments more accurate physics and improved time performance. The experiments will have the option to integrate GeantV's beta version into their applications and run early computing performance tests to verify the benefits of fine-grained parallelism (track-level parallelisation) and vectorised geometry, navigation, and physics libraries. Some experiments will be running vectorised versions of their readout and digitisation modeling code and verifying potential time performance improvements. Experiments will also be running production versions of their CPU intensive applications in HPC systems, benefiting from increased event throughput offered by these hybrid systems of CPUs, GPUs, and coprocessors. Training activities related to new software paradigms and emerging computing technologies, and support to the experiments on the use of newly available simulation tools should ramp-up at this time.

After 5 years, optimised versions of geometry and physics libraries will be available, improving even further the computing performance of detector simulation applications. The availability of a production version of the new track-level parallelisation and fully vectorised geometry, navigation, and physics libraries will offer the experiments the option to finalise integration into their frameworks, a task that will involve a significant personpower investment. They will then be able to focus on physics validation and computing performance tests, and measure whether the target goal of a factor of 2-5 improvement in time performance has been reached. If successful, the new engine would be in production on the timescale of the start of the HL-LHC run. High throughput tests can also be made by running optimised versions of CPU intensive applications in HPC systems.

In the longer term, as the improved libraries and simulation tools become mainstream and the applications for HPC facilities evolve, an iterative process will commence where interaction between tool developers and experiments will become of fundamental importance. Feedback from the experiments will be followed by fixes and optimisation by tool developers and the process will continue with an expected further improvement in performance. At this stage, experiment support and training will peak and be critical.

Fast Simulation

Fast simulation techniques have been employed since the LEP era and are likely to increase in importance for HL-LHC experiments. Many approaches to Fast Simulation take advantage of detector-specific features thus limiting the possibilities to develop common software. Some of them are applied to Geant4-based Full Simulation applications, such as Russian Roulette and shower libraries, to speedup simulation in a specific sub-detector. Many experiments, such as CMS, still label these applications as Full Simulation and Geant4-based. Some experiments have developed alternative Fast Simulation frameworks, based on fast track propagation through a simplified geometry and parametrized showers, which are less accurate and are typically used in detector upgrade studies and to produce “signal” samples that require scans of a large region of theory parameter space. There is also work to extend “hybrid” simulations, where some aspects of the event are treated with a fast approach and others with a detailed one. However, analysis-specific simulations would necessitate dedicated calibration and efficiency scale factors, making it impractical in many experiments.

The integration of fast and full simulation in the same product is a longer term R&D plan done in close collaboration with experiments, aiming to provide more generic solutions for common fast simulation use cases. This aims to a better coverage of experiment needs, reducing the need for separate frameworks or allowing for easier connection with existing solutions. Although, machine-learning algorithms will most likely need to be detector-specific, R&D is on the way to explore the possibility to reduce detector-specific effort by successfully making fast simulation match detailed Geant4 results.

Many ideas are being explored by experiments to improve the physics accuracy and speed of Fast Simulation in an effort to satisfy the need to generate a larger fraction of the total simulated events than before using these techniques. Machine Learning is one of the techniques being explored in these initiatives. One of the most ambitious is aiming to deliver, as part of the GeantV beta release, seamless integration of Fast Simulation in the Full Simulation flow, a standalone fast simulation tool based on Machine Learning, and a few examples including a detailed performance study for a calorimeter.

Fast Simulation physics tuning and validation is an ongoing task. The benefits of the machine learning approach will be evaluated during the first year. Assuming success, a ML-based fast simulation tune should be available for some physics observables and, by the end of year 5, it should be clear the extent to which a common infrastructure is applicable to the variety of detector configurations.

Pileup

Realistic simulation of proton-proton collisions simultaneous with, or almost simultaneous with, the collision of interest (pileup) will rise in importance, growing from an average of 50 in 2017 to 200 at the HL-LHC. The individual detector elements of the experiments are sensitive to pileup in different ways, and over different time ranges. In fact, the backgrounds to hard-scatter events have many components including in-time pileup, out-of-time-pileup, cavern background and beam-gas collisions. All of these components can be simulated but they present storage and I/O challenges related to the handling of the large simulated min-bias samples used to model the extra interactions. Alternatively, real zero-bias events can be collected, bypassing any zero suppression, and overlaid on the fully simulated hard scatters. This approach faces challenges related to the collection of non-zero-suppressed samples or the use of suppressed events, non-linear effects when adding electronic signals from different samples, and sub-detector misalignment consistency between the simulation and the real experiment. Another option is to “pre-mix” together the minimum bias collisions into individual events that have the full background expected for a single collision of interest. Although experiments have prioritized different approaches in the past, they will invest effort on improving their pre-mixing techniques, which allow the mixing to be performed at the digitisation level reducing the disk and network usage for a single event. They are also expected to invest in the development of the zero-bias overlay approach within the next 3 years.

Digitisation

Simulation toolkits do not include effects like charge drift in an electric field or models of the readout electronics of the experiments. Instead, these effects are normally taken into account in a separate step called digitisation. Digitisation is inherently local to a given sub-detector, and often even to a given readout element, so that there are many opportunities for parallelism in terms of vectorisation and multiprocessing or multi-threading, if the code and the data objects are designed optimally.

Recently, both hardware and software projects have benefitted from an increased level of sharing among experiments. The LArSoft Collaboration develops and supports a shared base of physics software across Liquid Argon (LAr) Time Projection Chamber (TPC) experiments, which includes to provide common digitisation code. Similarly, an effort exists among the LHC experiments to share code for modeling of radiation damage effects in silicon. As CMS and ATLAS expect to use similar readout chips in their future trackers, further code sharing might be possible.

It is expected that, within the next 3 years, common digitisation efforts are well-established among experiments and advanced high-performance generic digitisation examples, which experiments could use as a basis to develop their own code, become available. These prototypes may be based on vectorized software with SIMD friendly data structures and containers. It is desirable that this code is fully tested and validated within 5 years, in time for use by HL-LHC experiments and DUNE.

Pseudorandom Number Generation

The large number of random numbers required in the simulation of detector events presents challenges in the selection and use of pseudorandom number generators (PRNGs), in particular when running on infrastructures with a large degree of parallelism. Reproducibility is a key requirement, i.e. when repeating the simulation of an event with the same input particles, with the same state of a PRNG, the simulation must provide exactly the same results.

The future programme of work in the field of PRNGs must ensure that performant implementations of additional state-of-the-art PRNGs are made available, in particular PRNGs which have been shown to have none or minimal correlations between separate sequences or sub-sequences. The aim is to develop a single library containing sequential and vectorised implementations of the set of state-of-the-art PRNG, to replace the existing Root and CLHEP implementations within 3 years. This includes counter-based methods, the MIXMAX family of generators, improved implementations of RANLUX, implementations of CMRGs and other categories of sequential and parallel PRNGs. Once available, it will be necessary to promote a transition to the use of this library to replace existing implementations in ROOT, Geant4 and GeantV. At the same time, we will collaborate with the authors of the state of the art PRNG testing suites in order to extend testing to 64-bit variates, and expand the testing of correlations between sub-sequences. More generally we will follow and contribute to the evolution of the field of PRNGs for parallel and highly parallel applications, collaborating with researchers in the development of PRNG, seeking to obtain generators that address better our challenging requirements.