FERMILAB-PUB-22-145-CMS-PPD

Detector simulation challenges for future accelerator experiments

John Apostolakis [1], Marilena Bandieramonte [2], Sunanda Banerjee [3], Nazir Bartosik [4], Gloria Corti [1], Gabriele Cosmo [1], V. Daniel Elvira [5,*], Thomas Evans [6], Andrei Gheata [1], Simone Pagan Griso [7], Vladimir Ivantchenko [1], Christopher Jones [5], Markus Klute [8], Charles Leggett [7], Ben Morgan [9], Tadej Novak [10], Kevin Pedro [5], and Harald Paganetti [11]

[1] *European Organization for Nuclear Research (CERN), Organisation Européenne pour la Recherche Nucléaire (CERN), F-01631 Prévessin Cedex, France, CH-1211 Genève 23, Geneva, Switzerland*
[2] *Department of Physics and Astronomy, University of Pittsburgh, 3941 O'Hara Street., Pittsburgh, PA 15260, United States*
[3] *Department of Physics, University of Wisconsin, Room 2320 Chamberlin Hall, 1150 University Avenue, Madison, WI 53706-1390, United States*
[4] *Sezione di Torino, Istituto Nazionale di Fisica Nucleare (INFN), via P. Giuria, 1, 10125 Torino, Italy, Turin*
[5] *Fermi National Accelerator Laboratory (FNAL), P.O. Box 500, Batavia, IL 60510-0500, United States*
[6] *Oak Ridge National Laboratory, 1 Bethel Valley Rd., Oak Ridge, TN 37831, United States*
[7] *Physics Division, Lawrence Berkeley National Laboratory (LBNL), 1 Cyclotron Road, Berkeley, CA 94720-8153, United States*
[8] *Karlsruhe Institute of Technology (KIT), Kaiserstrasse 12, 76131 Karlsruhe, Germany*
[9] *Department of Physics, The University of Warwick, Coventry, CV4 7AL, United Kingdom*
[10] *Deutsches Elektronen-Synchrotron DESY, Notkestr. 85, 22607 Hamburg, Germany*
[11] *Department of Radiation Oncology, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, United States*

Correspondence*:
V.D. Elvira, Fermi National Accelerator Laboratory (FNAL), P.O. Box 500, Batavia, IL 60510-0500, United States
daniel@fnal.gov

ABSTRACT

Detector simulation is a key component for studies on prospective future high-energy colliders, the design, optimization, testing and operation of particle physics experiments, and the analysis of the data collected to perform physics measurements. This review starts from the current state of the art technology applied to detector simulation in high-energy physics and elaborates on the evolution of software tools developed to address the challenges posed by future accelerator

This manuscript has been authored by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics.

programs beyond the HL-LHC era, into the 2030-2050 period. New accelerator, detector, and computing technologies set the stage for an exercise in how detector simulation will serve the needs of the high-energy physics programs of the mid 21^{st} century, and its potential impact on other research domains.

Keywords: High Energy Physics, Particle Physics, Radiation, Simulation, Monte Carlo, Software, Computing, High Performance Computing

1 INTRODUCTION

Simulation is an essential tool to design, build, and commission the sophisticated accelerator facilities and particle detectors utilized in experimental high energy physics (HEP). In this context, simulation refers to a software workflow consisting of a chain of modules that starts with generation of initial particles, for example, final state particles from a proton-proton collision. A second module simulates the passage of these particles through the detector geometry and electromagnetic fields, as well as the physics interactions with its materials. The output contains information about times, positions, and energy deposits of the particles when they traverse the readout-sensitive components of the detector. In most modern experiments, this module is based on the Geant4 software toolkit [1, 2, 3] but other packages such as FLUKA [4, 5] and MARS [6] are also widely used, depending on the application. A third module generates the electronic signals from the readout components in response to the simulated interactions, outputting this data in the same format as the real detector system. As such, the datasets generated through simulation may be input to the same algorithms used to reconstruct physics observables from real data. Simulation is thus not only vital in designing HEP experiments, it also plays a fundamental role in the interpretation, validation, and analysis of the large and complex datasets collected by experiments to produce physics results, and its impact here should not be underestimated [7].

With many unanswered questions remaining in particle physics and the end of the Large Hadron Collider (LHC) program expected in the late 2020s, plans and ideas for the next big facilities of the 2030s-2050s are gaining momentum. As these facilities intend to explore ever higher energy scales and luminosities, the scale of simulated data samples needed to design the detectors and their software, and analyze the physics results will correspondingly grow. Simulation codes will thus face challenges in scaling both their throughput and accuracy to meet these sample size requirements with finite but ever evolving computational facilities [8]. The LHC era has already seen a significant evolution of simulation methods from ”full” detailed history-based algorithms to a hybrid of full and ”fast” parameterized or machine-learning based algorithms for the most computationally expensive parts of detectors [9]. A hybrid simulation strategy, using a combination of full and fast techniques will play a major role for future collider experiments, but full simulation will still be required to develop and validate the fast algorithms, as well as to support searches and analyses of rare processes. The goal of this article is to discuss how detector simulation codes may evolve to meet these challenges in the context of the second and third elements of the above simulation chain, that is the modeling of the detector, excluding the generation of initial particles (An overview of the computational challenges here may be found in [8]).

Section 2 presents the design parameters of future accelerators and detectors relevant to their simulation such as colliding particle types, beam parameters, and backgrounds. Challenges in the description and implementation of complex detector geometries and particle navigation through rapidly varying magnetic fields and detector elements of different shapes and materials are discussed in Section 3, while the physics models needed to describe the passage of particles through the detector material at the energy ranges associated with the colliders under consideration will be discussed in Section 4. Beam backgrounds from particle decay or multiple hard collisions are another important topic of discussion, particularly in the case of beams with particles that decay or emit synchrotron radiation, and will be discussed in Section 5. Section 6 focuses on readout modeling in the context of the opportunities and challenges posed by new detector technology, including novel materials and new generation electronics. Section 7 looks forward to the computing landscape anticipated in the era of future colliders, and how these technologies could help improve the physics and computing performance of detector simulation software, and even shape their future evolution. Section 8 will discuss the evolution of simulation software toolkits, including how they might adjust to new computing platforms, experiment software frameworks, programming languages, and the potential success of speculative ideas, as well as the features that would be needed to satisfy the requirements of future collider physics programs. For decades, HEP has collaborated with other communities, such as medical and nuclear physics, and space science, on detector simulation codes, resulting in valuable sharing of research and resources. Section 9 will present examples of application of detector simulation tools originating in HEP, in particular to the medical field, and how the challenges for future HEP simulation may overlap.

This article is one of the first reviews on the role and potential evolution of detector simulation in far future HEP collider physics programs. We hope it contributes to highlight its strategic importance both for HEP and other fields, as well as the need to preserve and grow its priceless community of developers and experts.

2 FUTURE ACCELERATORS AND DETECTORS IN NUMBERS

There are several designs for future particle accelerators, each with its strengths and challenges. This chapter focuses on the accelerator and detector design parameters and issues relevant for software modeling. In particular, we survey a number of the most mature proposals, including the high luminosity LHC (HL-LHC), the high energy LHC (HE-LHC), the Large Hadron-electron Collider (LHeC) and its high luminosity upgrade (HL-LHec), the Future Circular Collider (FCC) program of ee (electron-positron), hh (hadron-hadron), and eh (electron-hadron) colliders, the Circular Electron Positron Collider (CEPC), the Muon Collider, the International Linear Collider (ILC), the Compact Linear Collider (CLIC), and the Cool Copper Collider (CCC). Table 1 summarizes the parameters of these proposed future accelerators, including design values for maximal energy, peak luminosity, and integrated luminosity, and references for each proposal. There are other potential future colliders that are still being designed, including the Super Proton-Proton Collider (SPPC) [10], an electron-muon collider [11], a muon-proton collider [12], and a muon-ion collider [13].

Modern particle physics accelerators operate with bunched beams and reach peak luminosities higher than 1–2 $\times$ 10^{34} $\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$, exceeding the initial LHC design specification. The luminosity for future hadron colliders, such as the High Luminosity LHC (HL-LHC), is limited by the maximum number of simultaneous proton-proton collisions, or pileup, under which the detectors can operate effectively. For circular lepton colliders at higher energies, the luminosity is limited by the beamstrahlung (deflection-induced synchrotron radiation), and “top-up” or “top-off” schemes to inject additional particles during beam circulation are expected to be necessary to extend beam lifetime [14]. For linear machines, design parameters like the beam size, beam power, beam currents, and repetition rates drive the peak luminosity.

Proton-proton collisions offer the greatest energy reach, but they are limited by construction costs and the availability of high-field magnets. The largest proposed energy comes from the FCC-hh at 100 TeV. Lepton colliders can also push the energy frontier to multiple TeV. The muon collider requires R&D in order to reduce the transverse and longitudinal beam emittance via cooling and to accelerate to collision energies all within the muon's $2.2\,\mu$s lifetime [15]. However, it offers an exciting path to collision energies up to a few tens of TeV by suppressing synchrotron radiation relative to electrons. The beam-induced background (BIB) created by beam muons decaying in flight places new and unique demands on simulation [16]. Wakefield acceleration also offers a possibility for reaching high energies more compactly in the further future [17].

The proposed detector technologies for the next generations of experiments at colliders are growing in breadth, as indicated by the summary in Table 2. These increases in technological variety are driven by both physics goals and experimental conditions. In addition, new detectors will be increasing complex and granular. The interplay between instrumentation and computing is therefore increasingly important, as detectors become more challenging to simulate. One example is the upcoming High Granularity Calorimeter (HGCAL) at the Compact Muon Solenoid (CMS) experiment [18]. With roughly six million channels, it will be the most granular calorimeter built to date. This massively increases the geometry complexity, leading to a $\sim$40–60% increase in the time to simulate the detector [19]; in addition, the increased precision of the detector is expected to require correspondingly more precise physical models, which may further double the simulation time in existing software [20]. The incorporation of precision timing information may also place more demands on the accuracy of the simulation.

The HL-LHC is the nearest-future collider surveyed here, and most further-future colliders aim at higher precision measurements or present even more difficult environments. Therefore, detector complexity should be expected to continue to increase, in order to facilitate the physics programs and measurements for these new colliders. More than ever before, increasingly energetic and potentially heavier particles will interact with the detector materials, and massive increases in accumulated luminosity will enable physicists to explore the tails of relevant kinematic distributions very precisely. New technologies will pose their own challenges, such as the muon collider BIB, or new materials whose electromagnetic and nuclear interactions may not be fully characterized. This motivates the continued development of detector simulation software, to ensure its computational performance and physical accuracy keep up with the bold next steps of experimental high energy physics.

3 CHALLENGES IN GEOMETRY DESCRIPTION AND NAVIGATION

Geometric modeling is a core component of particle transport simulation. It describes both the material properties of detector components, which condition the particle interactions, and their geometric boundary limits. Particles are transported through these geometries in small spatial steps, requiring fast and accurate computation of distances and finding the geometry location after crossing volume boundaries. This task uses a significant fraction of total simulation time even for the current LHC experiments [8], making performance a general concern for the evolution of geometry modeling tools. As discussed in section 2, future detectors will have both higher granularity and interaction rates than the LHC, requiring the geometry modeling and navigation software to increase the throughput of the above calculations given this increased complexity. Providing the navigation precision necessary to achieve the required physics accuracy will likely be challenged by the presence of very thin detectors placed far away from the interaction point.

The detector geometry description of a HEP experiment goes through several processing steps between the initial computer-aided designs (CAD) [21] to the in-memory representation used by the simulation. These transformations primarily reduce the complexity and level of detail available in the CAD model to increase computing performance without compromising the required physics accuracy. Although the geometry models at the core of today's HEP detector simulation were designed in the 60's, Geant geometry implementations [22, 1] have enjoyed continuous success over many generations of CPU architectures because of a number of features that reduce both the memory footprint and algorithmic complexity. Multiple volume placements, replication using regular patterns, and hierarchies of non-overlapping 'container' volumes enable efficient simulation of very complex setups comprising tens of millions of components. However, creating the model description for such setups is often a long and arduous process, and the resulting geometry is very difficult to update and optimize.

The most popular 3D models used in simulation nowadays are based on primitive solid representations such as boxes, tubes, or trapezoids, supporting arbitrarily complex Boolean combinations using these building blocks. Different simulation packages use different constructive solid geometry (CSG) flavours [23], providing a number of features and model constraints to enhance the descriptive power and computation efficiency. However, performance can be highly degraded by overuse of some of these features, such as creating unbalanced hierarchies of volumes or creating overly complex Boolean solids. Using such inefficient constructs in high occupancy detector regions near the interaction point generally leads to significant performance degradation.

The current geometry implementations have a very limited adaptive capability for optimizing such inefficient constructs, mainly due to the high complexity of the model building blocks. The geometry queries can only be decomposed to the granularity of solid primitives, so user-defined constructs cannot be internally simplified. This calls for investigating surface models as alternatives to today's geometry representations. Adopting boundary representation (BREP) models [24] composed of first and second-order algebraic surfaces, would allow decomposing navigation tasks into simple surface queries. An appropriate choice of the BREP model flavor allowing surface queries to be independent could greatly favor the highly-parallel workflows of the future.

Developing automatic conversion tools from CAD surface-based models to the Geant4 simulation geometry proved to be too challenging in the past. Supporting surface representations directly in the simulation geometry would make such conversions possible. This would provide a simpler transition from the engineering designs to the simulation geometry, having fewer intermediate representations. It would also make it easier to implement transparent build-time optimizations for inefficient user constructs.

Successive upgrades to adapt to new computing paradigms such as object-oriented or parallel design have not touched the main modelling concepts described above, which served their purpose for decades of CPU evolution but are quickly becoming a limiting factor for computing hardware with acceleration. Recent R&D studies [25, 26] have shown that today's state-of-the-art Geant-derived geometry codes such as VecGeom [27] represent a bottleneck for vectorized or massively parallel workflows. Deep polymorphic code stacks, low branch predictability, and incoherent memory access are some of the most important reasons for performance degradation when instruction execution coherence is a hardware constraint. This is intrinsic to the model being used, combining in the same query algorithms of very different complexity, called in an unpredictable manner and unfriendly to compiler optimizations. These studies also indicate the need to simplify the geometry models being used, highly reducing or eliminating unnecessary abstractions.

Performance optimization is particularly important for common geometry navigation tasks such as collision detection of the simulated particle trajectories with the geometry setup, and relocation after crossing volume boundaries. Navigation helpers are using techniques such as *voxelization* [28] or bounding volume hierarchies (BVH) [29] to achieve logarithmic complexity in setups having several millions objects.

Adopting efficient optimization strategies will be more relevant for the more complex detectors of the future.

The same problem of collision detection is addressed by graphics systems, in particular, ray-tracing (RT) engines such as NVIDIA OptiX [30] that make use of dedicated hardware acceleration. Adapting HEP detector simulations to use such engine was implemented in the Opticks library [31], and demonstrated speedups of more than two orders of magnitude compared to CPU-based Geant4 simulations of optical photon transport in large liquid-Argon detectors. This required adapting the complete optical photon simulation workflow to GPUs, but also simplifying and transforming the geometry description to match OptiX requirements. Generalizing this approach for future HEP detector simulation would require a major re-engineering effort, in particular for the geometry description. How exactly RT technology evolves will likely have a big impact on the solutions adopted for detector geometry modeling. As the use of RT acceleration proliferates in the gaming industry, APIs supported by dedicated languages and libraries will most probably be made publicly available. Combined with larger on-chip caches, future low-latency graphics chips may allow externalizing geometry as an accelerated service for simulation. Such service could become an important booster, but would be conditional to the simplification of the geometry description and added support for batched multi-track workflows.

Evolution in computing technology will most probably present game-changing opportunities to improve simulation software, as described in Section 7). For example, tensor cores [32] provide a large density of Flops, although at a cost in precision. Geometry step calculations cannot make use of half-precision floating point (*FP16*) directly because rounding errors would become too important and affect both physics precision and transportation over large distances in the detector. Some optimizations may however be delegated to a *FP16*-based navigation system using ML inference to, for instance, prioritize candidate searches. Single-precision *FP32*-based geometry distance computation should be given more weight in the context of the evolution of reduced-precision accelerator-based hardware, because the option to reduce precision fulfils physics requirements in most cases. Furthermore, it would provide a significant performance boost due to a smaller number of memory operations for such architectures. Recent studies report performance gains as large as 40% for certain GPU-based simulation workflows [26], making R&D in this area a good investment, as long as memory operations remain the dominant bottleneck, even if chips evolve to provide higher Flops at *FP32* precision or better. The precision reduction option is, however, not suitable for e.g., micron-thin sensors, where the propagation rounding errors become comparable to the sensor thickness. Addressing this will require supporting different precision settings depending on the detector region.

4 PHYSICS PROCESSES AND MODELS

As mentioned in Sec. 1, Geant4 has emerged as the primary tool to model particle physics detectors. Geant4 offers a comprehensive list of physics models [33] combined with the continuous deployment of new features and improved functionality, as well as rigorous code verification and physics validation against experimental data.

4.1 Current status

During the first two periods of data taking in 2010-2018, the LHC experiments produced, reconstructed, stored, transferred, and analyzed tens of billions of simulated events. The physics quality of these Geant4-based Monte Carlo samples produced at unprecedented speed was one of the critical elements enabling these experiments to deliver physics measurements with greater precision and faster than in previous hadron colliders [7, 34]. Future accelerator programs will, however, require the implementation of additional physics processes and continuous improvements to the accuracy of existing ones. A quick review of the current status of physics models in Geant4 will precede a discussion of future needs.

Physics in Geant4 are subdivided over several domains, the most relevant for HEP being particle decay, electromagnetic (EM) interactions, hadronic processes, and optical photon transport. The precision of the modeling has to be such that it does not become a limiting factor to the potential offered by detector technology. EM physics interactions of $e^-/e^+/\gamma$ with the detector material, producing EM showers in calorimeters, consume a large fraction of the computing resources at the LHC experiments. Reproducing the response, resolution, and shower shape at a level of a few per mille requires modeling particle showers down to keV levels, which contain a large number of low-energy secondary particles that need to be produced and transported through magnetic fields. This level of accuracy is required in order to distinguish EM particles from hadronic jets, and to efficiently identify overlapping showers. Highly accurate models for energy deposition in thin calorimeter layers are also essential for reconstruction of charged particles and muons. Simulation of tracking devices requires accurate modeling of multiple scattering and backscattering at low and high energy, coupled with very low energy delta electrons. Geant4 delivers on all these requirements by modeling EM processes for all particle types in the 1 keV to 100 TeV energy range. The accuracy of Geant4 EM showers is verified by the CMS [35] and ATLAS [36] experiments.

Geant4 models physics processes for leptons, long-lived hadrons, and hadronic resonances. Simulation of particle decay follows recent PDG data. The decay of b-, c-quark hadrons and τ-leptons is outsourced to external physics generators via predefined decay mechanisms.

Simulation of optical photon production and transport is also provided by Geant4. The main accuracy limitation arises from the large compute time required to model the large number of photons and the many reflections that may occur in within the detector. Various methods to speed-up optical photon transport are available, depending on tolerance to physics approximations.

Hadron-nuclear interaction physics models are needed to simulate hadronic jets in calorimeters, hadronic processes in thin layers of tracking devices, and for simulating shower leakage to the muon chambers. Geant4 hadronic physics is based on theory models and tuned on thin target data [3]. This approach guarantees a more reliable predictive power than that offered by parametric models for a wide range of materials, particle types and energy ranges for which data measurements are not available. Parameter tuning and model extensions are necessary to describe all particle interactions at all energies [2]. Geant4 has adopted the approach of combining several models that fit the data best in different energy ranges, since it is unrealistic to expect that one single model would do the job over the full kinematic range of interest. This is done by providing several sets of predefined ”physics lists”, which are combinations of EM and hadronic processes and models. Experiments need to identify the most suitable for their own physics program by performing the necessary physics validation studies and possible applying calibration corrections [36, 37, 35, 38].

4.2 Future needs

The large data volumes to be collected by the HL-LHC experiments will enable experiments to reduce statistical uncertainties, therefore demanding more accurate simulation to help reduce systematic uncertainties in background estimation and calibration procedures. The next generation of HEP detectors to be commissioned at the LHC and designed to operate in future lepton and hadron colliders will have finer granularity and incorporate novel materials, requiring simulation physics models with improved accuracy and precision, as well as a broader kinematic coverage. Materials and magnetic fields will also need to be described in more detail to keep systematic uncertainties small. Moreover, new technologies [39, 40, 41] will allow detectors to sample particle showers with a high time resolution of the order of tens of picoseconds, which will need to be matched in simulation. Consequently, the simulation community has launched an ambitious R&D effort to upgrade physics models to improve accuracy and speed, re-implementing them from the grounds up when necessary (e.g., GeantV [25], Adept [26], Celeritas [42]). Special attention will be needed to extend accurate physics simulation to the $O(100) TeV$ domain, including new processes and models required to support the future collider programs.

Achieving an optimal balance between accuracy and software performance will be particularly challenging in the the case of EM physics, given that the corresponding software module is one of the largest consumers of compute power [34]. Reviews of EM physics model assumptions, approximations and limitations, including those for hadrons and ions will be needed to support the HL-LHC and Future Collider (FCC) programs. The Geant4 description of multiple scattering [43] of charged particles provides predictions in good agreement with data collected at the LHC. Nevertheless, the higher spatial resolution in new detectors, [44, 39, 45, 46] may require even higher accuracy to reproduce measured track and vertex resolutions. Excellent modeling of single-particle scattering and backscattering across Geant4 volume boundaries for low energy electrons are critical for accurate descriptions of shower shapes in calorimeters, such as CMS's high granularity hadronic calorimeter. At the very high energies present at the FCC, nuclear size effects must be taken into account, and elastic scattering models must be extended to include nuclear form factors in the highest energy range. The description of form factors may affect EM processes at high energies in such a way that it affects shower shapes and high energy muons. A theoretical description of the Landau-Pomeranchuk-Migdal (LPM) effect, significant at high energy, is included in the models describing the bremsstrahlung and pair-production processes in Geant4. For the latter, introducing LPM leads to differences in cross sections at very high energies that will need to be understood when data become available. A relativistic pair-production model is essential for simulation accuracy at the FCC. Rare EM processes like γ conversion to muon and hadron pairs also becomes important at very high energies and will have to be added. This is also essential to properly model beam background effects in the collision region of a Higgs Factory. In the cases of the FCC and dark matter search experiments, the description of pair production will need to be extended to include the emission of a nearby orbital electron (triplet production) and to take into account nuclear recoil effects. Finally, γ radiative corrections in EM physics may effects significantly the accuracy of measurements at Higgs factories and will need to be added to the models. All these rare processes must be added to the simulation to improve the accuracy in the tails of the physics distributions, where backgrounds become important. These corrections must be included such that they are invoked only as needed, thus not increasing the computing cost of EM modeling. At the FCC collision energy, the closeness of tracking devices to the interaction points will also require widening the range of physics models of short lived particles. This will be particularly important for high-precision heavy flavor measurements, as non-negligible fractions of beauty and charm hadrons will survive long enough to intercept beam pipes and the first detector layers. Describing the interaction of such particles with matter may already be required at the HL-LHC program because of a reduction of the distance between the trackers and the interaction point [44, 39]. A review of how detector simulation interfaces to dedicated decay generators during particle transport may be necessary.

In hadronic interactions, more than one model is needed to describe QCD physics processes accurately over the whole energy range. Typically, a hadronic interaction is initiated when a high energy hadron collides with a nucleon in the nucleus of a given material. This is followed by the propagation of the secondary particles produced through the nucleus, the subsequent de-excitation of the remnant nucleus and particle evaporation, until the nucleus reaches the ground state. Different sets of models map naturally to these phases depending on the initial energy of the collision: a parton string model for energy above few GeV, an intra-nuclear cascade model below that threshold. Pre-compound and de-excitation models are used to simulate the last steps in the evolution of the interaction. A reliable description of showers in hadronic calorimeters requires accurate descriptions of all these processes.

Geant4 offers two main physics lists to describe hadronic physics in high energy collider experiments. The main difference between the two consists in the choice of the model describing the initiating quark-parton phase mentioned above, either a quark-gluon string model, or a Fritiof model [3]. Having more than one model allows to estimate the systematic uncertainties arising from the approximations they use. Unfortunately, neither of them is accurate enough to describe the hadronic interactions at multi-TeV energies occurring at the FCC. New processes will need to be implemented in the hadronic physics simulation suite to address this higher energy domain, taking inspiration from those available in the EPOS generator [47], used by the cosmic ray and heavy ion physics communities.

Another element essential for the simulation of hadronic physics is precise calculations of interaction cross-sections. At the highest energies, Geant4 uses a general approach based on the Glauber theory [48], while at lower energies cross sections are evaluated from tables obtained from the Particle Data Group [49]. This approach profits from the latest thin-target experiment measurements and provide cross-sections for any type of projectile particle. The precision of cross-section calculations for different types of particles will need to be improved as more particle types become relevant to particle flow reconstruction in granular calorimeters.

A correct description of particle multiplicity within hadronic showers is also needed to model the physics performance of highly granular calorimeters (e.g. CMS [50]), and is also essential to simulate high-precision tracking devices (e.g. LHCb spectrometer). The parameters describing hadronic models must be tuned to describe all available thin target test beam data, and the models expanded to provide coverage to as many beam particles and target nuclei as possible. For flavor physics, it is important to take into account the differences in hadronic cross-sections between particle and anti-particle projectiles.

5 BEAM BACKGROUNDS AND PILEUP

The main categories of beam backgrounds at ee colliders are machine and luminosity induced [51]. The former is due to accelerator operation and includes Synchrotron Radiation (SR) and beam gas interactions. The latter arises from the interaction of the two beams close to the interaction point of the experiment.

The SR that may affect the detector comes from the bending and focusing magnets closest to it. While detectors will be shielded, a significant fraction of photons may still scatter in the interaction region and be detected. This is expected to be one of the dominant sources of backgrounds in the FCC-ee detector [52]. Beam gas effects are a result of collisions between the beam and residual hydrogen, oxygen and carbon gasses in the beam pipe inside the interaction region.

The luminosity induced background is generated from the electromagnetic force between the two approaching bunches, which leads to the production of hard bremstrahlung photons. These may interact with the beam and an effect similar to e^+e^- pair creation can occur, or they scatter with each other which can result in hadrons, and potentially jets, in the detector. Stray electrons due to scattering between beam particles in the same bunch can also hit the detector.

The main background at pp colliders are the large number of inelastic proton–proton collisions that occur simultaneously with the hard-scatter process, collectively known as pileup. This usually results in a number of soft jets coinciding with the collision. The number of interactions per crossing at the future colliders is expected to exceed one thousand, compared to no more than 200 at the end of the HL-LHC era. An additional source of luminosity induced background is the cavern background. Neutrons may propagate through the experimental cavern for a few seconds before they are thermalized, thus producing a neutron-photon gas. This gas produces a constant background, consisting of low-energy electrons and protons from spallation.

Machine induced backgrounds at pp colliders are similar to the ee ones [36]. Besides the beam gas, the beam halo is a background resulting from interactions between the beam and upstream accelerator elements. In general, pile-up dominates over the beam gas and beam halo.

Muon colliders are special in that the accelerated particles are not stable. The main source of beam backgrounds are decays of primary muons and the interaction of their decay products with the collider and detector components [53]. Compared to ee colliders this represents an additional source of background resulting in a large number of low momentum particles that may not be stopped by shielding end enter the interaction region of the detector. Additionally, this type of background needs to be simulated with higher precision outside of the interaction region.

An important consideration is the detector response and readout time compared to the time between collisions, which is often longer. In-time and out-of-time pile-up should be considered separately. In-time pileup are additional collisions that coincide with the hard-scatter one, while out-of-time pile-up are collisions from different bunch crossings than the hard-scatter one, but affect the readout implicitly.

5.1 Bottlenecks in computational performance

The biggest bottleneck in the time it takes to model pileup in a pp collider is the number of interactions per bunch crossing. As seen in black circles in Fig. 1, the CPU time requirement has a very steep dependence on this parameter, which needs to match data-taking conditions. The second issue can be the slow response time of the detectors, requiring a large number of out-of-time bunch crossings to be simulated. This can be solved by only simulating the detectors when needed, as not all have the same sensitive time range. Improvements in detector technologies that will be used in future experiments may make these times small enough not to cause a significant overhead.

Traditionally each in-time or out-of-time interaction is sampled individually and taken into account at the digitisation step, when detector digital responses are emulated. Experiments pre-sample pile-up events and reuse them between different samples to reduce computational time [54, 55]. While the pre-sampling itself still has the same CPU limitations, using those pileup events barely depends on the amount of pileup (red circles in Fig. 1), but could cause larger stress on storage. Thresholds to analogue signals are applied at digitization to reduce the amount of saved digits significantly, at the cost of reduced precision when two digital channels are merged. Thus pre-sampling thresholds need to be tuned for each individual detector, and computing resources can only be saved by reusing pre-sampled events, where a compromise between CPU savings and increased storage needs to be made in a way that maintains optimal physics performance.

Another option to fully avoid the CPU bottleneck of pileup pre-sampling is to use pileup events from data. The main bottlenecks here are non-constant detector conditions and alignment. Re-initializing the simulated geometry adds overheads which may be mitigated by averaging conditions over long periods. However, this solution will come at the cost of reproducing data less precisely. Furthermore detector readout only provides digital information above some thresholds which are usually tuned for primary collisions and thus relatively high. This reduces precision when merging the information with the simulated hard-scatter event.

While other types of background are much lower at pp colliders and their simulation can usually be skipped, this is not the case for ee colliders. Some of those backgrounds, e.g. beam gas effects, synchrotron radiation and intra-beam scattering, happen outside the detector cavern. They are simulated by the accelerator team as they also affect beam operations. To avoid re-simulating the same type of background, the simulation can be shared with the experiment as a list of particles that enter the interaction region [56], though this is still a large number of low-momentum particles to simulate. Experiments thus also use randomly-triggered collision events for the background estimation, while also being affected by the threshold effects.

5.2 Optimal strategy for future colliders

During the development stage of the future experiments, detailed simulation of all types of beam backgrounds is of utmost importance. Simulation provides estimates of the physics impact of backgrounds and helps to optimize the detector design to minimize them as much as possible [57]. Some backgrounds can be parametrized or even completely neglected. One such example is that of cavern background neutrons at hadron colliders. In most cases their contribution is orders of magnitude smaller than that of pileup, although outer muon chambers would require a detailed description, if high precision is required. As low momentum neutron simulation is very slow, it can be performed only once and used to derive parametrized detector responses, which can then be injected at the digitization stage.

As discussed earlier in this section, separate simulation of beam backgrounds and pre-digitization saves computing resources and has a negligible impact on physics performance when reused randomly between samples. With the increased background rates expected in future colliders, iterative mixing and merging of background contributions will become an essential technique. Detector readout thresholds must be set sufficiently low to allow merging of digital signals multiple times with negligible degradation of accuracy. This would allow iterative pileup pre-sampling, where multiple events with a low number of interactions could be merged to give an event with a high number of interactions. It would also allow to merge different types of backgrounds that would be prepared independently. Furthermore, a special set of lower background thresholds could be setup in the actual detector to enable the use of real data events as background sources. The latter would yield a reduced performance degradation as compared to current detectors.

Most of all the beam background simulation strategy depends on physics accuracy requirements. As mentioned in Sec. 1, current experiments are moving towards a more frequent use of fast simulation methods, either based on parametrized detector responses or on machine learning technologies. The latter could be used to choose the precision of the simulation algorithm depending on the event properties, or to fully generate the background on the fly. Regardless of the choice of the strategy used to simulate large volumes of physics samples, a detailed modeling as provided by full simulation will always be needed, if nothing else to derive and tune the faster methods.

6 ELECTRONIC SIGNAL MODELING

The ambitious physics program at future accelerator-based experiments requires detectors which can perform very accurate measurements and handle high occupancy at the same time. To achieve these goals, it is of paramount importance to collect as much information from each individual detector channel as possible, including the three spatial coordinates, time and energy.

For simplicity, this section focuses on two main classes of detectors that pose the most challenges from a computational point of view: tracking detectors and calorimeters. Those are the ones that usually employ the largest number of electronic readout channels, thus their behavior needs to be simulated in detail.

New generation calorimeters are designed as tracking devices as well as providers of energy deposition information in the form of the five-dimensional measurement referred to in the first paragraph. These extended capabilities beyond traditional calorimetric observables present challenges to the simulation effort, since modeling must achieve accurate descriptions of all these observables simultaneously. Additionally, calorimeters will often operate in a high-occupancy environment in which sensor and electronics performance degrade fast as a consequence of radiation damage.

The digitization step of simulation takes as input the Geant4-generated analogue signals from the detector. The first step of the digitization process accumulates this input and groups it for individual read out elements. This is done in a number of time slots which define the integration time for the detector. Beyond this step, modeling is highly detector dependent. It is driven by detailed descriptions of readout electronics including the noise component, cross-talk, and the readout logic which involves the shaping of the signal and the digitization of the pulse. Finally, a digit is recorded when the signal is above a predefined threshold.

6.1 Tracking detectors

Various types of tracking detectors are currently employed in HEP experiments at colliders [49], with the most widely used being silicon, gaseous (RPC, MDT, Micromegas, etc), transition-radiation, and scintillation detectors. Of these, silicon-based detectors are among the most challenging and computationally expensive to simulate, given the large number of channels and observables involved.

Silicon detectors give rise to electron-hole pairs which are collected with a certain efficiency, amplified, digitized, and recorded. When biased by a voltage difference, the response of the sensor to the passage of ionizing particles is characterized by its charge collection efficiency (CCE) and its leakage current (I_{leak}). As the sensors are operated well above its full depletion voltage, the CCE is expected to be high. The current digitization models for silicon detectors are mostly based on a bottom-up approach, where the overall energy deposit is used to generate multiple electron-hole pairs that are then propagated through a detailed simulation of the electric field and used to compute the expected signal generated at the electrodes. Several models are employed for how the overall deposited energy is split. They range from simple models performing an equal-splitting along the expected trajectory to more complex models [58], each giving different increasing levels of accuracy at the price of being computationally more expensive.

Exposure to radiation induces displacements in the lattice and ionization damage, liberating charge carriers. These effects contribute to a reduction of the CCE and and increase in the I_{leak}. The increase in instantaneous luminosity projected at the HL-LHC collider challenged experiments to implement simulation models able to predict the reduced CCE expected in the presence of radiation damage. A detailed simulation of the electric field is used with more refined models describing the probability of charge-trapping and reduced CCE [59, 60, 61]. Those models tend to be heavy on computing resources, prompting parametric simulation approaches to be developed as well.

Detectors designs for future colliders differ substantially depending on the type of environment they will have to withstand. Detectors at moderate to high-energy e^+e^- colliders will see a clean event and moderate rates of radiation. For such detectors, a detailed simulation strategy is crucial for high precision physics measurements; however, the demand for large simulated samples makes a hybrid approach including parametrizations most likely. Silicon-based tracking detectors are also the technology of choice at muon colliders. The radiation environment within this machine poses unique challenges due to the high level of beam-induced backgrounds (BIBs). Real-time selection of what measurements are most likely to come from the interaction point rather than from BIBs is likely to rely on detailed shape analyses of the neighboring pixels that give signals as well as possible correlation across closely-spaced layers [62]. A hybrid approach will likely be needed, consisting of a detailed simulation of the detector layers where the raw signal multiplicity is the highest and needs to be reduced, together with a fast simulation approach for the rest of the tracking detector. For detectors at future hadron colliders, the extreme radiation environment near the interaction point will make it mandatory to implement radiation damage effects in the simulation. For this, a parametrized approach would also be the most realistic path to keep computational costs under control.

6.2 Calorimeters

Traditional calorimeters utilize photons generated through the process of scintillation, Cerenkov radiation, or transition radiation to measure particle energy depositions. Photons are detected by a photo-transducer where the photons first give rise to electrons and then go through successive steps of amplification and digitization. Modeling photon transport to the photo-transducers is CPU intensive and traditionally implemented as a parametrization tuned to predictions obtained from a specialized simulation package [31, 63]. Nowadays, simulation of optical photons is offloaded to GPUs to mitigate computing costs, taking advantage of the high levels of parallelism achievable for electromagnetically interacting particles' transport. The photon transmission coefficient is affected by radiation damage due to formation of color centers in the medium, thus an assumption is made on the distribution of color centers in the medium. The light output, $L(d)$, after receiving a radiation dose d, is described by an exponential function that depends on the dose:

$$L(d) \quad = \quad L_0 \exp(\mu \cdot d) \ ,$$

where the parameter μ is a property of the material and depends on the dose rate. The radiation damage parametrizations are typically calibrated from data coming out of a monitoring system. The radiation dose and the neutron fluence (flux over time) are estimated using an independent simulation of the detector setup.

The next step in the simulation chain for calorimeters is the treatment of the photo transducer, the most commonly used type being silicon photo-multipliers. These devices also suffer time-dependent effects related to the radiation exposure: decrease of photo-statistics (fewer photons reaching the device) and increase of the noise coming from dark currents. The noise increases with the square root of the fluence, which in turn is proportional to the sensor's area. Signal simulation in silicon photo-multipliers involves: emulation of photo-statistics using a Poisson distribution, description of the distribution of the photo electrons according to pulse shape, adjustmentment of the signal arrival time, as well as the modeling of the dark current (thermal emission of photo-electrons), the cross-talk among the channels induced in the neighbors of the fired pixels, the pixel recovery time after being fired, and the saturation effect for large signals when several photo-electrons fall on the same pixel. An exponential function describes accurately the re-charge of the pixel as a function of time, while cross-talk can be modeled using a branching Poisson process. The Borel distribution [64, 65] analytically computes the probability of neighboring cells to fire.

Finally, the simulation of the readout electronics includes: the readout gain, adjusted to get an acceptable signal to noise ratio throughout the life time of the detector); the electronics noise, with contributions from the leakage current in the detector, the resistors shunting the input to the readout chip, and the implementation of the so-called common mode-subtraction; and the ADC pulse shape, which decides the fraction of charge leaked to the neighboring bunches. Zero suppression is also modeled, keeping only the digits which cross a threshold in the time bunch corresponding to sample of interest.

In future colliders, simulation of silicon-based calorimeters will face similar challenges than those described in the previous section for tracking devices. Parametrizations of time consuming photon transport may be replaced with detailed modeling and processed on computing devices with hardware accelerators. Radiation damage will be more pronounced in high-background environments such as high-energy hadron colliders and muon colliders, introducing a time-dependent component all through the signal simulation chain which will need to be measured from data and modeled in detail.

7 COMPUTING

Non traditional, heterogeneous architectures, such as GPUs, have recently begun to dominate the design of new High Performance Computing centers, and are also showing increasing prevalence in data centers and cloud computing resources. Transitioning HEP software to run on modern system is proving to be a slow and challenging process, as described in Sec. 7.3. However, in the timescale of future colliders, this evolution in the computing landscape offers tremendous opportunity to HEP experiments. The predicted increase in compute power, the capability to offload different tasks to specialized hardware in hybrid systems, the option to run inference as a service in remote locations in the context of a machine learning approach, open the field of HEP simulation to a world where simulation data could grow severalfold in size, while preserving or improving physics models and detector descriptions.

7.1 Projection of hardware architecture evolution

For example, the U.S. Department of Energy (USDOE) will be standing up three new GPU-accelerated, exascale platforms in 2023–2024 at the Oak Ridge Leadership Computing Facility (OLCF [66]), Argonne Leadership Computing Facility (ALCF [67]), and Lawrence Livermore National Laboratory. Additionally, the National Energy Research Scientific Computing Center (NERSC [68]) is deploying an NVIDIA-based GPU system for basic scientific research. Figure 2 shows peak performance in Flops for machines deployed at the OLCF between 2012 and 2023. In addition to the projected $\sim 55\times$ increase in computing performance from 2012 to 2022, the percent of peak provided by GPUs has increased from $\sim 91\%$ to greater than 98% over that period. This situation is reflected in computing centers around the world such as Piz Daint in Swizterland [69], Leonardo in Italy [70], and Karolina in Czechia [71] that heavily use NVIDIA GPUs, LUMI in Finland [72] that will use AMD GPUs, and MareNostrum 4 in Spain [73] that uses both NVIDIA and AMD GPUs. Japan's Fugaku [74], the current leader of the Top 500 supercomputers list [75], has a novel architecture with very wide registers that behave very much like a GPU. We see similar heterogeneous computing center designs in smaller institutional clusters, and grid computing sites. Thus, in order to take advantage of the massive increases in computing capability provided at the HPC centers, optimizing existing and future simulation codes for GPUs is essential. The other HPCs at the head of the current Top500 List which do not explicitly use GPUs, such as Fugaku, have hybrid architectures that have very wide vector processors that offer much the same functionality as traditional GPUs.

The primary driver of this evolution is the power requirements driving high-performance computing. Figure 3 shows power consumption for OLCF machines from 2012 to 2022. Here, we see that for a $3\times$ increase in total power consumption there is a 17 fold increase in Flops per MW.

It is difficult to predict the exact nature of the hardware landscape beyond 5 years or so, but undoubtedly we will see evolutionary changes of current hardware rather than revolutionary ones - a failed product can now cost billions of dollars due to design and fabrication costs. Core counts will continue to go up, as transistor feature sizes decrease, with increasing use of multi-chip and 3D stacked solutions needed to avoid overly large silicon sizes. It is also likely that vendors will devote larger sections of silicon to specialized functions, such as we see with Tensor and Ray Tracing cores in current GPUs. FPGA and ASIC vendors are now offering specialized component layouts for domain specific applications, and this level of customization will likely increase. We are also beginning to see the combination of multiple different types of cores, such as high and low power CPUs and FPGAs into the same silicon die or chiplet array, leading to more integrated heterogeneous architectures with faster communication channels between the various components and much quicker offload speeds.

7.2 Description of heterogeneous architectures

Heterogeneous architectures such as GPUs and FPGAs are fundamentally different from traditional CPU architectures. CPUs typically possess a small number of complicated cores that excel at branch prediction and instruction prefetching. They have multiple levels of large, fast caches, and typically have very low access latencies. GPUs, on the other hand, have a very large number of simple cores (hundreds of thousands for modern GPUs), that do not handle branch mis-predictions gracefully. GPU cores that are grouped in a block must operate in lockstep, all processing the same instruction. Branch mis-predictions and thread divergence will cause a stall, greatly decreasing throughput. GPUs often have much more silicon devoted to lower and mixed precision operations than they do for double precision calculations, which are heavily used in High Energy Physics. GPUs are optimized for Single Instruction Multiple Data (SIMD) style of operations, where sequential threads or cores access sequential memory locations - randomized memory access causes significant performance degradation. Finally, GPUs have very high access latencies compared to CPUs - it can take tens of microseconds to offload a kernel from a host to a GPU. The combination of massive parallelism, memory access patterns, and high latencies of GPUs require a fundamentally different programming model than that of CPUs.

7.3 Challenges for software developers

All of the GPU manufacturers support programming only with their own software stack. NVIDIA uses CUDA, AMD promotes HIP, and Intel employs oneAPI. Other heterogeneous architectures such as FPGAs also use unique programming languages such as Verilog and HLS. The vast majority of current HEP software is written in C++, and supported by physicists who are usually not professional developers. Typical HEP workflows encompass millions of lines of code, with hundreds to thousands of kernels, none of which dominate the computation. In order to target the current diverse range of GPUs and FPGAs, we would have to rewrite a very large fraction of the HEP software stack in multiple languages. Given the limited available workforce, and the extremely challenging nature of validating code that executes differently on multiple architectures, experiments would have to make very difficult choices as to which hardware they could target, ignoring large amounts of available computing power. Fortunately, we have seen a number of portability solutions start to emerge recently, such as Kokkos, Raja, Alpaka, and SYCL, which are able to target more than one hardware backend (see Figure 4). Furthermore, hardware vendors have seen the benefits of cross platform compatibility, and are working to develop standards which they are trying to incorporate into the C++ standard. Ideally, a single language or API that could target both CPUs and all available heterogeneous architectures would be the preferred solution.

Currently, mapping computational physics and data codes to GPU architectures requires significant effort and profiling. Most HEP code bases are not easily vectorizable or parallelizable, and many HEP applications are characterized by random memory access patterns. They tend to follow sequential paradigms, with many conditional branch points, which make them challenging to adapt to GPUs. Even tasks such as particle transport, which in high luminosity environments such as the HL-LHC seemingly offer very high levels of parallelism, are in fact very difficult to run efficiently on GPUs due to rapid thread divergence cause by non-homogeneous geometrical and magnetic field constraints.

One avenue that offers some hope for easier adoption of GPUs is the use of Machine Learning (ML) techniques to solve physics problems. We are seeing increasing acceptance of ML algorithms for pattern recognition and feature discrimination tasks in HEP, as well as for more novel tasks such as generative models for energy depositions in calorimeter simulations. ML backends for all GPU and other heterogeneous architectures already exist, and are often supported directly by the hardware manufacturers, which greatly eases the burden for HEP developers.

8 SOFTWARE TOOLKITS

The evolution of simulation software toolkits will depend greatly on the hardware, whose evolution on the timescale of 10 years is uncertain as discussed in Section 7. Today's leading toolkit, the Geant4 toolkit [3] used by most large experiments' detector simulation, and also the particle transport tools FLUKA [5] and MARS15 [6] used in the assessment of radiation effects, are large, complex, and have evolved over thirty years of CPU-centric computation.

8.1 Computing hardware accelerator usage

Whether current simulation toolkits can be adapted to profit adequately from a variety of computing hardware accelerators, principally GPUs, or whether new accelerator-centric codes can be created and then interfaced into existing toolkits is a key research question. The profitability of the conversion also involves the effort required for the development of the production level code, and the cost to create GPU-capable applications. The latter is under active exploration.

The research into GPU usage is inspired by efforts in related particle transport applications in HEP and other fields. As discussed in Secs. 3 and 6, the Opticks project [31] offloads simulation of optical photons to NVidia GPUs and demonstrates methods to deal with complex specialised geometries on these devices, specifically ones that have many repetitive structures. MPEXS, a CUDA-based application for medical physics [76] using Geant4-derived physics models, also demonstrated efficient use of GPU resources for regular 'voxelised' geometries. However, the general problem of modeling a large range of energies for particles combined with the full complexities of modern detector geometries has not been tackled yet. Solving these general problems is the domain of two ongoing R&D efforts, the Celeritas project [42] and the AdePT prototype [26]. Both are starting by creating CUDA-based proof-of-concept implementations of electromagnetic physics, and particularly showering, in complex detector geometries on GPUs. Key goals of the projects include identifying and solving major performance bottlenecks, and providing a first template for efficiently extracting energy deposits, track passage data, and similar user-defined data. Initially, both are targeting the simulation of electron, photon, and positron showers in complex geometrical structures currently described by deep hierarchies with many repetitions of volumes at different levels. They have identified the need for a geometry modeller adapted for GPUs and accelerators, and sufficiently capable to handle these complex structures (see Sec. 3). They are in the process of defining and developing solutions for such a geometry modeller.

The limitations of the bandwidth and latency for communication between the CPU and accelerator are an important constraint in the utilization of GPUs and other accelerators for particle transport simulation, and for the overall application. Minimising the amount of data exchanged, such as input particles and output hits, between the CPU and accelerator, is an important design constraint for GPU-based particle transport. The types of detectors for which it is suitable may depend on this. The contention for this resource may also constrain the overall application which integrates the particle transport and showering with event generation, generation of signal, and further reconstruction.

Existing prototypes such as AdePT and Celeritas strongly focus on keeping computation inside the accelerator, and moving back to the CPU only the absolute minimum of data and work. When only a selected region of a geometry is accelerated, a particle which escapes that region must be returned - as must particle tracks which undergo (rare) interactions not currently simulated in GPU code, e.g. photo-nuclear interactions. Of course the largest and critical data transferred out of the accelerator are the experiment hit records (or processed signal sum values) and other user information such as truth information.

Early phase exploration of the potential of FPGAs for particle transport is being conducted for medical physics simulation [77]. Yet the challenges involved appear more daunting, due to the need to compile a complex tool into hardware. It seems likely that this approach would be investigated only after implementations are built using 'simpler' building blocks on GPUs. Potentially these will profit from leveraging implementations created for portable programming frameworks.

Based on current trends, except situations where ultimate performance is required for time critical applications, we expect the established vendor-specific libraries (CUDA, Hip, DPC++) to be slowly supplanted by the emerging portable programming paradigms (Kokkos, Alpaka, SYCL), and within a few years a convergence to be established on standard-supported languages and libraries such as C++'s standard library std::par execution policy. With the importance of portability between hardware of different vendors, it is critical to identify and invest in cross-vendor solutions, and potentially paradigms that can be used to investigate alternative hardware platforms, as mentioned above for FPGAs.

8.2 Opportunities for Parallelism

We expect applications and future toolkits will need to expose multiple levels of parallelism in order to manage resources and to coordinate with other computations, such as reconstruction and event generation. Such levels could entail parallel processing of different events as well as parallel processing of multiple algorithms or even multiple particles within an event. A detector simulation toolkit cannot assume that it controls all resources, but must cooperate with other ongoing tasks in the experiment application. At this point, it is unclear how to accomplish this cooperation efficiently.

Seeking to obtain massive parallelism of thousands or tens of thousands of active particles is challenging to develop in detector simulation. The GeantV project [25] explored the potential of SIMD-CPU based parallelism by marshalling similar work ('event-based' in the parlance of neutron simulation), e.g. waiting till many particles entered a particular volume before propagating the particles through that volume. The project's conclusion was that the speedup potential was modest - between 1.2 and 2.0.

It seems clear that the ability to execute many concurrent, independent kernels on recent GPUs is of crucial interest to HEP, as it avoids the need for very fine grained parallelism at the thread level, which was the goal of the GeantV project. Given the difficulty of taking advantage of the full available parallelism of modern GPUs by a single kernel, being able to execute many kernels doing different tasks will be invaluable.

8.3 Parametrized Simulation

In parallel with the need for a full, detailed simulation capability to meet the physics requirements of the future colliders, the focus is growing on developing techniques that replace the most CPU-intensive components of the simulation with faster methods (so called "fast simulation" techniques), while maintaining an adequate physics accuracy. This category includes optimization/biasing techniques that aim at tuning parameters concerning simulation constituents such as geometry or physics models and which are strictly experiment specific, as well as the possibility of parametrizing part of the simulation (i.e. electromagnetic shower development in calorimeters), by combining different machine learning techniques. R&D efforts are ongoing in all the major LHC experiments to apply cutting-edge techniques in generative modelling with deep learning approaches, e.g., GANs, VAEs and normalizing flows, targeting the description of electromagnetic showers.

We expect the bulk production of Monte Carlo simulation data to be performed with a combination of detailed and parametrized simulation techniques. To this end, enabling the possibility to combine fast and full simulation tools in a flexible way is of crucial importance. Along these lines, we expect Geant4 to evolve coherently by providing tools allowing integration of ML techniques with an efficient and smooth interleaving of different types of simulation.

8.4 Future of Geant4

Due to its versatility, the large number of physics modeling options, and the investment of many experiments including the LHC experiments, we expect an evolved Geant4 to be a key component of detector simulation for both the ongoing and the near future experiments well into the 2030s. Over the next decade, we expect Geant4's capabilities to evolve to include options for parameterized simulation using machine learning, and acceleration for specific configurations (geometry, particles and interactions) on selected hardware, both of which should significantly increase simulation throughput. These enhanced capabilities will however come with significant constraints, due to the effort required to adapt user code to the accelerator/heterogeneous computing paradigm. Furthermore, there is a need to demonstrate that substantial speedup or throughput improvements can be obtained before such an investment in adaptation of user applications can be undertaken. Full utilization of accelerators may not be required as offloading some work to accelerators should free up CPU cores to do additional work at the same time thereby improving throughput. In addition, some HPC sites may require applications to make some use of GPUs in order to run at the site. Therefore, some minimum GPU utilization by simulation may make it possible for experiments to run on such HPC resources thereby reducing the total time it takes to do large scale simulation workflows.

9 APPLICATIONS OF HEP TOOLS TO MEDICAL PHYSICS AND OTHER FIELDS

After the initial developments of Monte Carlo (MC) methods for the Manhattan project, the tools became available to the wider research community after declassification in the 1950's. One of the early adapters of MC methods were physicists in radiation therapy. Researchers were eager to predict the dose in patients more accurately as well as designing and simulating detectors for quality assurance and radiation protection. The simulations were done mainly using in-house developed codes, with some low energy codes modeling photons up to 20 MeV developed or transferred from basic physics applications [78, 33]. Use of MC tools from the HEP domain mainly started with heavy charged particle therapy, first using protons and Helium ions and later employing heavier ions such as Carbon ions. Early research here was also done with in-house codes mostly studying scattering in inhomogeneous media. In the early 1990's more and more high-energy physicists entered the field of medical physics and brought their expertise and codes with them. Thus started the use of general-purpose MC codes in radiation therapy that were initially developed and designed for high energy physics applications, such as Geant4 and Fluka. Fruitful collaborations were also established with the space physics field, with HEP-developed toolkits applied to particle detector design as well as the similar areas of dosimetry and radiation damage [79].

9.1 Beam line design and shielding calculations

Beam line design and shielding calculations are done prior to installing a treatment device. These applications of MC are no different to the HEP use case except for the beam energies studied. Beam line transport would be done by the machine manufacturers and is often based on specialized codes such as, for instance, Beam Delivery Simulation (BDSIM) [33]. On the other hand, shielding calculations aim at a conservative estimate with limited required accuracy and would use mostly analytical methods.

Shielding calculations are also critical in both manned and unmanned space missions to determine the radiation environment for humans [80] and instrumentation, as well as detector backgrounds [81].

9.2 Detector design studies

Nuclear and HEP physics hardware developments are frequently finding applications in radiation therapy and space missions due to similar requirements concerning sensors and real-time data processing. Detectors are less complex compared to HEP but the components used in simulations are very similar. Differences are in the particles of interest as well as the energy region of interest. As in HEP, MC simulations are a powerful tool to optimize detectors and treatment devices [82, 83]. In fact, for radiation therapy or diagnostic imaging, MC are not only being employed by researchers but also by vendors to optimize their equipment.

9.3 Dose calculation

Predicting the dose in patients is arguably the most important task in radiation therapy and has therefore been the most active MC topic [84]. It has similar importance in space physics for predicting dose rates for astronauts and in materials/electronics [80, 85].

Despite its accuracy, MC dose calculation has not found widespread use in treatment planning in medicine. However, vendors of commercial planning systems have now developed very fast Monte Carlo codes for treatment planning where millions of histories in thick targets need to be simulated in minutes or seconds in a very complex geometry, i.e. the patient as imaged with CT [86]. Therefore, these specialized codes have replaced multi-purpose MC codes that are often less efficient. Multi-purpose codes are however being used as a gold standard for measurements that are not feasible in humans. In addition, they are often used to commission treatment planning and delivery workflows. As we are dealing with biological samples such as patients, scoring functionality often goes beyond about what is typically used in HEP such as scoring phase spaces on irregular shaped surfaces or dealing with time-dependent geometries.

9.4 Diagnostic medical imaging

MC has long been used in the design of imaging systems such as positron emission tomography (PET) or computed tomography (CT) [87]. Like in therapy, HEP codes are being applied either directly or tailored to imaging applications, i.e. for low energy applications [88]. Time of flight as well as optical simulations are done using MC. In recent years MC is more and more used to also understand interactions in patients.

As radiation therapy is pursuing image-guided therapy, imaging devices are also incorporated in treatment machines resulting in problems that are being studied using MC such as the interaction between magnetic resonance imaging (MRI) and radiation therapy, either conventional (photon based) or magnetically scanned proton treatments.

9.5 Simulation requirements for non-HEP applications

9.5.1 Physics models and data for energy ranges of interest

Medical and many space applications typically fall not under high-energy but low-energy physics. HEP tools might therefore not simulate some effects accurately or their standard settings are not applicable for low energies and have to be adjusted and potentially even separately validated [89]. Measurements of fragmentation cross-sections and attenuation curves are needed for MC applications in clinical environments. Most cross sections and codes are indeed not very accurate for applications outside HEP because materials and energy regions of interest are very different. In fact, cross sections needed for medical physics applications go mostly back to experiments done in the 1970's and are no longer of interest to the basic physics community. For instance, considerable uncertainties in nuclear interaction cross sections in biological targets are particularly apparent in the simulation of isotope productions [90]. Furthermore, the interest of high-energy physics is mainly in thin targets whereas medical physics needs accurate representations of thick target physics to determine energy loss in patients or devices including Coulomb scattering and nuclear halo. For positively charged particles, the range in tissue materials needs to be predicted with mm accuracy and 2 accuracy in energy deposition at mm volumes. Novel approaches to verify treatment rely on detecting secondary gamma radiation outside of the patient requiring accurate nuclear excitation cross sections in the MeV region.

9.5.2 Computational efficiency (variance reduction)

In the future we may see two types of MC tools in medical physics, i.e. high-efficiency MC algorithms focusing solely on dose calculation for treatment planning and multi-purpose codes from high energy physics for research and development. The latter can and will be used more and more to replace difficult or cumbersome experiments such as detector design studies for dosimetry and imaging. Nevertheless, thick target simulations are often time consuming and variance reduction techniques have been developed in medical physics [91] that may also be applicable for high-energy physics applications, as discussed in Section 8, with cross-fertilization of the two fields.

9.6 Future role of MC tools outside of HEP

The main application of high-energy physics tools to other domains will continue to be in detector design, quality assurance and dose calculation. Furthermore, not only researchers in medical and space physics but also manufacturers of therapy and detector equipment are employing MC methods to develop new devices. Whilst these fields may not in general have the extreme requirements on performance and throughput as the future experiments discussed in Section 2, the improvements necessary here for HEP will benefit other user communities. By delivering higher accuracy physics with a smaller computational resource for a given sample size, a commensurate reduction in the costs to research time, money, and environmental impact will be possible.

It is important that collaborations between the many communities utilizing simulation codes are maintained to ensure sharing of requirements and methodologies to mutual benefit. Medical physics increasingly overlaps with radiation biology, where research promises a higher clinical impact than pure physics studies, a paradigm shift that became apparent in the last decade. Monte Carlo codes will thus be applied also in the field of radiation biology and radiation biochemistry [92]. Multiple efforts have already started, most notably the extensions of Geant4 (Geant4-DNA) and TOPAS (TOPAS-nBio) [93, 94]. These extensions require codes to evolve particularly when it comes to physics in small nanometer volumes and computational efficiency when using very small step sizes, which may have commonalities with the geometry developments discussed in Section 3. Figure 5 shows an an example of the geometries of typical size and complexity of molecular structures that are targeted by these simulations. The toolkit/API design of codes such as Geant4 have been critical in allowing such extensions, as well as allowing development of a wide range of applications for generic use cases [88, 95, 96, 97]. It is vitally important that HEP MC codes continue to use this software architecture to allow such innovation and extension. With simulation geometries, energy regions, materials, particle tracking and scoring that may be very different from HEP applications, continued exchange of ideas from other user communities will be invaluable in maintaining and developing HEP simulation codes.

10 SUMMARY AND CONCLUSIONS

Detector simulation codes such as Geant4, FLUKA, and MARS have played a central part in the development and operation of the current generation of HEP experiments and in the analysis and interpretation of their physics results. This critical role will continue as physicists design and plan the next generation of collider facilities to operate during the mid-21^{st}. These experiments, like their predecessors, will push the boundaries of accelerator and detector technology to explore and improve our knowledge of fundamental physics. While simulation codes have already been significantly upgraded through the LHC era to take full advantage of technologies including multi-core CPUs and machine learning, further evolution will be needed for this software to run on future computing architectures and deliver the large and accurate data samples demanded by future collider programs.

The primary challenges for detector simulation posed by future accelerators and detector designs are driven by the increased beam luminosities and energies combined with the high granularity (in space and time) of the proposed detectors. Higher luminosity naturally means that simulations will need to deliver larger sample sizes to reduce statistical uncertainties in, for example, background estimations, driving an overall need to increase performance and hence throughput. Corresponding increases in the accuracy and precision of models for electromagnetic and hadronic physics processes will thus be required to reduce systematic uncertainties, and to extend their domain of validity to cover higher beam energies and novel materials. Beam backgrounds will also increase in line with luminosity, and are a especially important area to model during the design phase of experiments to optimize physics and instrumental backgrounds therefore improving the precision of physics measurements and extending the reach of new particle searches. Higher granularity detector systems will challenge current codes for describing their geometries with the increased number of volumes, as well as propagating particles over large distances while retaining precision of their intersections with small or thin detector elements. R&D programs are already underway to explore directions for evolving this critical area of simulation. They are exploring techniques and hardware used in the computer graphics industry for ray-tracing and Computer-Aided Design (CAD), a particularly promising direction of research. Both high luminosity and detector granularity impact the final simulation step of digitization. The increased number of detector readout channels generates a higher computational load, especially for bottom-up models of signal creation, while the more intense radiation environment will require time-dependent effects measured from data to be modelled.

None of these components of the overall simulation toolkit exist in isolation. For example, the accuracy of energy depositions in a fine grained tracking calorimeter will be dependent on the interplay between the physics models and navigation of particles through the geometry elements under the influence of a magnetic field. Balancing physics accuracy against computing performance will be an important aspect for experimentalists and simulation code developers to consider. It is clear that employing a hybrid of full and fast parametrized or ML-based techniques is a realistic strategy for simulating detectors. Fast simulation may well find application in a broader range of cases than at the present time, either as a full generative step, or to optimize inputs to, or choice of, full Monte Carlo algorithms. Complete, high throughput, ”full” simulation workflows will nonetheless be required to develop, validate, and tune ”fast” methods, as well as to retrain or optimize them in response to changes in experiment conditions or physics program.

While the debate here is driven by the requirements of future HEP collider programs, simulation software evolves in the context of changes in a broader landscape of developments in hardware and software for High Performance Computing in academia and industry. The ever rapid pace of technology development limits predictions of how this may impact HEP over the next five years, let alone the 2040-2050 timescale for experiments in future collider facilities, but even the current evolutionary trends in GPU, FPGA and other new architectures offer many exciting opportunities for greater computational power at lower monetary and environmental cost. Equally, a significant challenge for HEP simulation will be in evolving existing interfaces and algorithms to effectively utilize this diverse range of emerging architectures. Software portability tools to assist targeting multiple hardware backends are developing rapidly, and experience in their use is building within the HEP community. HEP-originated simulation codes have permeated to other fields requiring modeling of radiation transport, especially in medical, bio-, and space physics. The collaborations established through this wide range of use cases have lead to many mutually beneficial developments and impact in both research and industry, and this can be expected, and should be encouraged, to continue. Though there are differences in energy ranges and detector complexity, increased physics accuracy and computational efficiency and throughput will be to the benefit of all. Furthermore, new or novel commonalities may be found, for example in modeling and navigating complex geometries whether that be a future collider detector or a DNA molecule.

Predicting the future for any technological or scientific endeavour can only offer a blurred snapshot of reality, but it is clear at least that the HEP community will continue to require accurate and computationally efficient detector simulation codes to develop and utilize its next generation of facilities. Developing software that meets these requirements presents a major, yet exciting, challenge that will foster collaboration across fundamental physics, high performance computing and computer science, medical, bio- and space physics, both in academia and industry. It is this depth and breadth of expertise across domains that will support and drive innovation in HEP simulation, making this human resource the most important to nurture and grow to enable the realization of HEP physics programs at future colliders during the second half of the 21^{st} century.

11 ADDITIONAL REQUIREMENTS

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

FUNDING

Details of all funding sources should be provided, including grant numbers if applicable. Please ensure to add all necessary funding information, as after publication this is no longer possible.

Ben Morgan is supported by STFC.

V. Daniel Elvira, Christopher Jones, and Kevin Pedro are supported by Fermilab, operated by Fermi Research Alliance, LLC under Con- tract No. DE-AC02-07CH11359 with the U.S. Department of Energy.

Harald Paganetti is in part funded by National Institute of Health: U24 CA215123 (The TOPAS Tool for Particle Simulation, a Monte Carlo Simulation Tool for Physics, Biology and Clinical Research)

Simone Pagan Griso is supported by the U.S. Department of Energy, Office of Science under contract DE-AC02- 05CH11231.

Thomas Evans is supported by the Exascale Computing Project (ECP), project number 17-SC-20-SC. Work for this paper was supported by Oak Ridge National Laboratory (ORNL), which is managed and operated by UT-Battelle, LLC, for the U.S. Department of Energy (DOE) under Contract No. DEAC05-00OR22725.

Charles Leggett is supported by the DOE HEP Center for Computational Excellence at Lawrence Berkeley National Laboratory under B&R KA2401045 with the US Department of Energy.

ACKNOWLEDGMENTS

This is a short text to acknowledge the contributions of specific colleagues, institutions, or agencies that aided the efforts of the authors.

REFERENCES

[1] Agostinelli S, Allison J, Amako K, Apostolakis J, Araujo H, Arce P, et al. Geant4—a simulation toolkit. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **506** (2003) 250–303. doi:https://doi.org/10.1016/S0168-9002(03)01368-8.

[2] Allison J, Amako K, Apostolakis J, Araujo H, Arce Dubois P, Asai M, et al. Geant4 developments and applications. *IEEE Transactions on Nuclear Science* **53** (2006) 270–278. doi:10.1109/TNS.2006.869826.

[3] Allison J, Amako K, Apostolakis J, Arce P, Asai M, Aso T, et al. Recent developments in Geant4. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **835** (2016) 186–225. doi:https://doi.org/10.1016/j.nima.2016.06.125.

[4] Ferrari A, Sala PR, Fassò A, Ranft J. *FLUKA: A multi-particle transport code (program version 2005)*. CERN Yellow Reports: Monographs (Geneva: CERN) (2005). doi:10.5170/CERN-2005-010.

[5] Böhlen T, Cerutti F, Chin M, Fassò A, Ferrari A, Ortega P, et al. The FLUKA code: Developments and challenges for high energy and medical applications. *Nuclear Data Sheets* **120** (2014) 211 – 214. doi:10.1016/j.nds.2014.07.049.

[6] [Dataset] Mokhov N. MARS15, Version 00 (2016).

[7] Elvira VD. Impact of detector simulation in particle physics collider experiments. *Physics Reports* **695** (2017) 1–54. doi:https://doi.org/10.1016/j.physrep.2017.06.002.

[8] The HEP Software Foundation. A roadmap for HEP software and computing R&D for the 2020s. *Computing and Software for Big Science* **3** (2019) 7. doi:10.1007/s41781-018-0018-8.

[9] Aad G, et al. AtlFast3: the next generation of fast simulation in ATLAS. *Comput. Softw. Big Sci.* **6** (2022) 7. doi:10.1007/s41781-021-00079-7.

[10] Tang J, et al. Concept for a Future Super Proton-Proton Collider (2015).

[11] Lu M, Levin AM, Li C, Agapitos A, Li Q, Meng F, et al. The physics case for an electron-muon collider. *Adv. High Energy Phys.* **2021** (2021) 6693618. doi:10.1155/2021/6693618.

[12] Kaya U, Ketenoglu B, Sultansoy S, Zimmermann F. Luminosity and Physics Considerations on HL-LHC and HE-LHC based mu-p Colliders (2019).

[13] Acosta D, Li W. A muon–ion collider at BNL: The future QCD frontier and path to a new energy frontier of $\mu+\mu-$ colliders. *Nucl. Instrum. Meth. A* **1027** (2022) 166334. doi:10.1016/j.nima.2022.166334.

[14] Talman R. Scaling Behavior of Circular Colliders Dominated by Synchrotron Radiation. *Int. J. Mod. Phys. A* **30** (2015) 1544003. doi:10.1142/S0217751X15440030.

[15] Delahaye JP, Diemoz M, Long K, Mansoulié B, Pastrone N, Rivkin L, et al. Muon Colliders (2019).

[16] Bartosik N, et al. Full Detector Simulation with Unprecedented Background Occupancy at a Muon Collider. *Comput. Softw. Big Sci.* **5** (2021) 21. doi:10.1007/s41781-021-00067-x.

[17] D'Arcy R, Chappell J, Beinortaite J, Diederichs S, Boyle G, Foster B, et al. Recovery time of a plasma-wakefield accelerator. *Nature* **603** (2022) 58–62. doi:10.1038/s41586-021-04348-8.

[18] The Phase-2 Upgrade of the CMS Endcap Calorimeter (2017).

[19] Pedro K. Current and future performance of the CMS simulation. *EPJ Web Conf.* **214** (2019) 02036. doi:10.1051/epjconf/201921402036.

[20] Pedro K. Integration and Performance of New Technologies in the CMS Simulation. *EPJ Web Conf.* **245** (2020) 02020. doi:10.1051/epjconf/202024502020.

[21] Hoschek J, Lasser D. *Fundamentals of computer aided geometric design* (AK Peters, Ltd.) (1993).

[22] Brun R, Bruyant F, Maire M, McPherson AC, Zanarini P. GEANT3 (1987).

[23] Foley J. *Computer graphics : principles and practice* (Reading, Mass: Addison-Wesley) (1995).

[24] Stroud I. *Boundary representation modelling techniques* (Springer Science & Business Media) (2006).

[25] Amadio G, Ananya A, Apostolakis J, Bandieramonte M, Banerjee S, Bhattacharyya A, 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.

[26] [Dataset] AdePT code repository. `https://github.com/apt-sim/AdePT` (2021).

[27] Apostolakis J, et al. Towards a high performance geometry library for particle-detector simulations. *J. of Phys.: Conf. Ser.* **608** (2015) 012023. doi:10.1088/1742-6596/608/1/012023.

[28] Foley J. *Computer graphics : principles and practice* (Reading, Mass: Addison-Wesley) (1990).

[29] Larsson T, Akenine-Möller T. A dynamic bounding volume hierarchy for generalized collision detection. *Computers & Graphics* **30** (2006) 450–459.

[30] Parker SG, Bigler J, Dietrich A, Friedrich H, Hoberock J, Luebke D, et al. Optix: A general purpose ray tracing engine. *ACM Trans. Graph.* **29** (2010). doi:10.1145/1778765.1778803.

[31] Blyth SC. Opticks : GPU optical photon simulation for particle physics using NVIDIA® OptiX™. *Journal of Physics: Conference Series* **898** (2017) 042001. doi:10.1088/1742-6596/898/4/042001.

[32] Markidis S, Chien SWD, Laure E, Peng IB, Vetter JS. NVIDIA tensor core programmability, performance & precision. *CoRR* **abs/1803.04014** (2018).

[33] Nevay L, Boogert S, Snuverink J, Abramov A, Deacon L, Garcia-Morales H, et al. BDSIM: An accelerator tracking code with particle–matter interactions. *Computer Physics Communications* **252** (2020) 107200. doi:https://doi.org/10.1016/j.cpc.2020.107200.

[34] Apostolakis J, et al. HEP Software Foundation Community White Paper Working Group - Detector Simulation (2018).

[35] Sirunyan AM, et al. Measurements of Higgs boson production cross sections and couplings in the diphoton decay channel at $\sqrt{\mathrm{s}}$ = 13 TeV. *JHEP* **07** (2021) 027. doi:10.1007/JHEP07(2021)027.

[36] Aad G, et al. The ATLAS Simulation Infrastructure. *Eur. Phys. J. C* **70** (2010) 823–874. doi:10.1140/epjc/s10052-010-1429-9.

[37] Aad G, et al. Measurements of differential cross-sections in top-quark pair events with a high transverse momentum top quark and limits on beyond the Standard Model contributions to top-quark pair production with the ATLAS detector at $\sqrt{s} = 13$ TeV (2022).

[38] Sirunyan AM, et al. Performance of missing transverse momentum reconstruction in proton-proton collisions at $\sqrt{s} = 13$ TeV using the CMS detector. *JINST* **14** (2019) P07004. doi:10.1088/1748-0221/14/07/P07004.

[39] LHCb Collaboration CM. Framework TDR for the LHCb Upgrade II
- Opportunities in flavour physics, and beyond, in the HL-LHC era
. Tech. rep., CERN, Geneva (2021).

[40] Butler JN, Tabarelli de Fatis T. A MIP Timing Detector for the CMS Phase-2 Upgrade (2019).

[41] Technical Design Report: A High-Granularity Timing Detector for the ATLAS Phase-II Upgrade. Tech. rep., CERN, Geneva (2020).

[42] Johnson S, Tognini S, Canal P, Evans T, Jun S, Lima G, et al. Novel features and GPU performance analysis for EM particle transport in the Celeritas code. *EPJ Web of Conferences* **251** (2021). doi:10.1051/epjconf/202125103030.

[43] Howard A, Ivanchenko V, Novak M, Ribon A. Status of Geant4 simulations of calorimeters. *JINST* **15** (2020) C05073. doi:10.1088/1748-0221/15/05/C05073.

[44] Collaboration L. LHCb VELO Upgrade Technical Design Report. Tech. rep. (2013).

[45] Tumasyan A, et al. The Phase-2 Upgrade of the CMS Tracker (2017).

[46] Technical Design Report for the ATLAS Inner Tracker Pixel Detector. Tech. rep., CERN, Geneva (2017).

[47] Drescher H, Hladik M, Ostapchenko S, Pierog T, Werner K. Parton-based Gribov–Regge theory. *Physics Reports* **350** (2001) 93–289. doi:https://doi.org/10.1016/S0370-1573(00)00122-8.

[48] Grichine VM. Integral cross-sections of light nuclei in the Glauber-Gribov representation. *Nucl. Instrum. Meth. B* **427** (2018) 60–62. doi:10.1016/j.nimb.2018.04.044.

[49] Zyla PA, et al. Review of Particle Physics. *PTEP* **2020** (2020) 083C01. doi:10.1093/ptep/ptaa104. See the review on *Particle Detectors at Accelerators*.

[50] The Phase-2 Upgrade of the CMS Endcap Calorimeter (2017).

[51] Natochii A, Vahsen SE, Nakayama H, Ishibashi T, Terui S. Improved simulation of beam backgrounds and collimation at SuperKEKB. *Phys. Rev. Accel. Beams* **24** (2021) 081001. doi:10.1103/PhysRevAccelBeams.24.081001.

[52] Voutsinas G, Bacchetta N, Boscolo M, Janot P, Kolano A, Perez E, et al. Luminosity- and Beam- Induced Backgrounds for the FCC-ee Interaction Region Design. *8th International Particle Accelerator Conference* (2017). doi:10.18429/JACoW-IPAC2017-WEPIK004.

[53] Collamati F, Curatolo C, Lucchesi D, Mereghetti A, Mokhov N, Palmer M, et al. Advanced assessment of beam-induced background at a muon collider. *JINST* **16** (2021) P11009. doi:10.1088/1748-0221/16/11/P11009.

[54] Aad G, et al. Emulating the impact of additional proton-proton interactions in the ATLAS simulation by pre-sampling sets of inelastic Monte Carlo events. *Comput Softw Big Sci* **6** (2022). doi:10.1007/s41781-021-00062-2.

[55] Lange DJ, Hildreth M, Ivantchenko VN, Osborne I. Upgrades for the CMS simulation. *J. Phys. Conf. Ser.* **608** (2015) 012056. doi:10.1088/1742-6596/608/1/012056.

[56] Altmannshofer W, et al. The Belle II Physics Book. *PTEP* **2019** (2019) 123C01. doi:10.1093/ptep/ptz106. [Erratum: PTEP 2020, 029201 (2020)].

[57] Dannheim D, Sailer A. Beam-Induced Backgrounds in the CLIC Detectors (2012).

[58] Bichsel H. Straggling in Thin Silicon Detectors. *Rev. Mod. Phys.* **60** (1988) 663–699. doi:10.1103/RevModPhys.60.663.

[59] Aaboud M, et al. Modelling radiation damage to pixel sensors in the ATLAS detector. *JINST* **14** (2019) P06012. doi:10.1088/1748-0221/14/06/P06012.

[60] Chiochia V, Swartz M, Bortoletto D, Cremaldi L, Cucciarelli S, Dorokhov A, et al. Simulation of heavily irradiated silicon pixel sensors and comparison with test beam measurements. *IEEE Transactions on Nuclear Science* **52** (2005) 1067–1075. doi:10.1109/tns.2005.852748.

[61] Swartz M, Fehling D, Giurgiu G, Maksimovic P, Chiochia V. A new technique for the reconstruction, validation, and simulation of hits in the CMS pixel detector. *PoS* **VERTEX2007** (2007) 035. doi:10.22323/1.057.0035.

[62] Muon Collider Collaboration. Expected Detector Performance at the Muon Collider (????) to appear.

[63] Gentit FX. Litrani: a General Purpose Monte-Carlo Program Simulating Light Propagation In Isotropic or Anisotropic Media (2001).

[64] Vinogradov S. Analytical models of probability distribution and excess noise factor of Solid State Photomultiplier signals with crosstalk. *Nucl. Instrum. Meth. A* **695** (2012) 247–251. doi:10.1016/j.nima.2011.11.086.

[65] Gallego L, Rosado J, Blanco F, Arqueros F. Modeling crosstalk in silicon photomultipliers. *JINST* **8** (2013) P05010. doi:10.1088/1748-0221/8/05/P05010.

[66] Oak Ridge National Laboratory: Leadership Computing Facility (2022). https://www.olcf.ornl.gov.

[67] Argonne Leadership Computing Facility (2022). [Online] https://alcf.anl.gov.

[68] National Energy Research Scientific Computing Center (2022). [Online] https://www.nersc.gov.

[69] PIZ DAINT (2022). [Online] https://www.cscs.ch/computers/piz-daint.

[70] Our supercomputer davinci-1 (2021). [Online] https://www.leonardo.com/en/innovation-technology/davinci-1-supercomputer.

[71] KAROLINA (2020). [Online] https://www.it4i.cz/en/infrastructure/karolina.

[72] LUMI (2022). [Online] https://lumi-supercomputer.eu.

[73] Barcelona Supercomputing Center (2022). [Online] https://www.bsc.es/marenostrum/marenostrum.

[74] Supercomputer Fugaku (2021). [Online] https://www.fujitsu.com/global/about/innovation/fugaku.

[75] TOP500 LIST—NOVEMBER 2021 (2021). [Online] https://www.top500.org/lists/top500/list/2021/11/.

[76] Okada S, Murakami K, Incerti S, Amako K, Sasaki T. MPEXS-DNA, a new GPU-based Monte Carlo simulator for track structures and radiation chemistry at subcellular scale. *Medical Physics* **46** (2019) 1483–1500. doi:https://doi.org/10.1002/mp.13370.

[77] Khusainov B, Kerrigan EC, Constantinides GA. Automatic software and computing hardware co-design for predictive control. *CoRR* **abs/1710.08802** (2017).

[78] Rogers DWO. Fifty years of Monte Carlo simulations for medical physics. *Physics in Medicine and Biology* **51** (2006) R287–R301. doi:10.1088/0031-9155/51/13/R17.

[79] Antin G, Nieminen P, Daly E. Applications of GEANT4 for the ESA space programme (2006), 511 – 518.

[80] Chavy-Macdonald MA, Menicucci A, Santin G, Evans H, Jiggens P, Nieminen P, et al. High-accuracy simulations of the ISS radiation environment and applications to interplanetary manned missions. *IEEE Transactions on Nuclear Science* **60** (2013) 2427 – 2434. doi:10.1109/TNS.2013.2273375.

[81] Eraerds T, Antonelli V, Davis C, Hall D, Hetherington O, Holland A, et al. Enhanced simulations on the Athena/Wide Field Imager instrumental background. *Journal of Astronomical Telescopes, Instruments, and Systems* **7** (2021). doi:10.1117/1.JATIS.7.3.034001.

[82] Park H, Paganetti H, Schuemann J, Jia X, Min CH. Monte Carlo methods for device simulations in radiation therapy. *Physics in Medicine and Biology* **66** (2021). doi:10.1088/1361-6560/ac1d1f.

[83] Baumann KS, Kaupa S, Bach C, Engenhart-Cabillic R, Zink K. Monte Carlo calculation of perturbation correction factors for air-filled ionization chambers in clinical proton beams using TOPAS/GEANT. *Zeitschrift fur Medizinische Physik* **31** (2021) 175 – 191. doi:10.1016/j.zemedi.2020.08.004.

[84] Paganetti H. Range uncertainties in proton therapy and the role of Monte Carlo simulations. *International Journal of Radiation Oncology Biology Physics* **86** (2013) 396. doi:10.1016/j.ijrobp.2013.02.017.

[85] Matthiä D, Hassler D, de Wet W, Ehresmann B, Firan A, Flores-McLaughlin J, et al. The radiation environment on the surface of Mars - summary of model calculations and comparison to RAD data. *Life Sciences in Space Research* **14** (2017) 18 – 28. doi:10.1016/j.lssr.2017.06.003.

[86] Fracchiolla F, Engwall E, Janson M, Tamm F, Lorentini S, Fellin F, et al. Clinical validation of a GPU-based Monte Carlo dose engine of a commercial treatment planning system for pencil beam scanning proton therapy. *Physica Medica* **88** (2021) 226 – 234. doi:10.1016/j.ejmp.2021.07.012.

[87] Costa P, Nersissian D, Umisedo N, Gonzales A, Fernández-Varea J. A comprehensive Monte Carlo study of CT dose metrics proposed by the AAPM Reports 111 and 200. *Medical Physics* **49** (2022) 201 – 218. doi:10.1002/mp.15306.

[88] Sarrut D, Bała M, Bardi sM, Bert J, Chauvin M, Chatzipapas K, et al. Advanced Monte Carlo simulations of emission tomography imaging systems with GATE. *Physics in Medicine and Biology* **66** (2021). doi:10.1088/1361-6560/abf276.

[89] Jarlskog C, Paganetti H. Physics settings for using the Geant4 toolkit in proton therapy. *IEEE Transactions on Nuclear Science* **55** (2008) 1018 – 1025. doi:10.1109/TNS.2008.922816.

[90] España S, Zhu X, Daartz J, El Fakhri G, Bortfeld T, Paganetti H. The reliability of proton-nuclear interaction cross-section data to predict proton-induced PET images in proton therapy. *Physics in Medicine and Biology* **56** (2011) 2687 – 2698. doi:10.1088/0031-9155/56/9/003.

[91] Ramos-Méndez J, Schuemann J, Incerti S, Paganetti H, Schulte R, Faddegon B. Flagged uniform particle splitting for variance reduction in proton and carbon ion track-structure simulations. *Physics in Medicine and Biology* **62** (2017) 5908–5925. doi:10.1088/1361-6560/aa7831.

[92] Plante I. A review of simulation codes and approaches for radiation chemistry. *Physics in Medicine & Biology* **66** (2021) 03TR02. doi:10.1088/1361-6560/abbd19.

[93] Bernal MA, Bordage MC, Brown JMC, Davídková M, Delage E, El Bitar Z, et al. Track structure modeling in liquid water: A review of the Geant4-DNA very low energy extension of the Geant4 Monte Carlo simulation toolkit. *Physica Medica* **31** (2015) 861–874. doi:10.1016/j.ejmp.2015.10.087.

[94] Schuemann J, McNamara A, Ramos-Méndez J, Perl J, Held K, Paganetti H, et al. TOPAS-nBio: An extension to the TOPAS simulation toolkit for cellular and sub-cellular radiobiology. *Radiation Research* **191** (2019) 125 – 138. doi:10.1667/RR15226.1.
[95] Faddegon B, Ramos-Méndez J, Schuemann J, McNamara A, Shin J, Perl J, et al. The TOPAS tool for particle simulation, a Monte Carlo simulation tool for physics, biology and clinical research. *Physica Medica* **72** (2020) 114–121. doi:10.1118/1.4758060.
[96] Santin G, Ivanchenko V, Evans H, Nieminen P, Daly E. GRAS: A general-purpose 3-d modular simulation tool for space environment effects analysis. *IEEE Transactions on Nuclear Science* **52** (2005) 2294 – 2299. doi:10.1109/TNS.2005.860749.
[97] Lei F, Truscott P, Dyer C, Quaghebeur B, Heynderickx D, Nieminen P, et al. MULASSIS: A Geant4-based multilayered shielding simulation tool. *IEEE Transactions on Nuclear Science* **49 I** (2002) 2788 – 2793. doi:10.1109/TNS.2002.805351.
[98] High-Luminosity Large Hadron Collider (HL-LHC): Technical Design Report V. 0.1 **4/2017** (2017). doi:10.23731/CYRM-2017-004.
[99] Cid Vidal X, et al. Report from Working Group 3: Beyond the Standard Model physics at the HL-LHC and HE-LHC. *CERN Yellow Rep. Monogr.* **7** (2019) 585–865. doi:10.23731/CYRM-2019-007.585.
[100] Abada A, et al. HE-LHC: The High-Energy Large Hadron Collider: Future Circular Collider Conceptual Design Report Volume 4. *Eur. Phys. J. ST* **228** (2019) 1109–1382. doi:10.1140/epjst/e2019-900088-6.
[101] Bordry F, Benedikt M, Brüning O, Jowett J, Rossi L, Schulte D, et al. Machine Parameters and Projected Luminosity Performance of Proposed Future Colliders at CERN (2018).
[102] Agostini P, et al. The Large Hadron-Electron Collider at the HL-LHC. *J. Phys. G* **48** (2021) 110501. doi:10.1088/1361-6471/abf3ba.
[103] Benedikt M, Blondel A, Janot P, Mangano M, Zimmermann F. Future Circular Colliders succeeding the LHC. *Nature Phys.* **16** (2020) 402–407. doi:10.1038/s41567-020-0856-2.
[104] Abada A, et al. FCC Physics Opportunities: Future Circular Collider Conceptual Design Report Volume 1. *Eur. Phys. J. C* **79** (2019) 474. doi:10.1140/epjc/s10052-019-6904-3.
[105] Abada A, et al. FCC-ee: The Lepton Collider: Future Circular Collider Conceptual Design Report Volume 2. *Eur. Phys. J. ST* **228** (2019) 261–623. doi:10.1140/epjst/e2019-900045-4.
[106] Abada A, et al. FCC-hh: The Hadron Collider: Future Circular Collider Conceptual Design Report Volume 3. *Eur. Phys. J. ST* **228** (2019) 755–1107. doi:10.1140/epjst/e2019-900087-0.
[107] CEPC Conceptual Design Report: Volume 1 - Accelerator (2018).
[108] Dong M, et al. CEPC Conceptual Design Report: Volume 2 - Physics & Detector (2018).
[109] Schulte D. The International Muon Collider Collaboration. *JACoW* **IPAC2021** (2021) THPAB017. doi:10.18429/JACoW-IPAC2021-THPAB017.
[110] Aarons G, et al. ILC Reference Design Report Volume 1 - Executive Summary (2007).
[111] Robson A, Burrows PN, Catalan Lasheras N, Linssen L, Petric M, Schulte D, et al. The Compact Linear e^+e^- Collider (CLIC): Accelerator and Detector (2018).
[112] The Compact Linear e^+e^- Collider (CLIC): Physics Potential (2018).
[113] Bai M, et al. C^3: A ”Cool” Route to the Higgs Boson and Beyond (2021).

TABLES

FIGURES

Collider	Particles	$\sqrt{s}$	Peak lumi. $[10^{34}\ \mathrm{cm}^{-2}\,\mathrm{s}^{-1}]$	Peak pileup	Total lumi. $[\mathrm{ab}^{-1}]$
HL-LHC [98, 99]	pp	14 TeV	7.5	200	3–4
HE-LHC [100, 99]	pp	27 TeV	16	500	15
LHeC [101, 102]	ep	1.3 TeV	0.5–2.4	0.1	1
HE-LHeC [101, 102]	ep	1.77 TeV	1.5	0.1	2
FCC-ee [103, 104, 105]	ee	88–365 GeV	1.5–230	0	1.5–150
FCC-hh [103, 104, 106]	pp	100 TeV	30	1000	20
FCC-eh [103, 104, 106]	ep	3.5 TeV	1.5	1	2
CEPC [107, 108]	ee	90–240 GeV	32–3	0	2.6–16
Muon Collider [109]	$\mu\mu$	3–14 TeV	1.8–40	*	1–20
ILC [110]	ee	250–500 GeV	2.7–3.6	0	1
CLIC [111, 112]	ee	0.38–3 TeV	1.5–6	0	1–5
CCC [113]	ee	250–550 GeV	1.3–2.4	0	2–4

Table 1. The parameters of various future accelerators. * Muon colliders face beam-induced backgrounds, which have different properties from pileup at ee or pp colliders.

Technology	Tracker	Calorimeter	Muon detector	PID
Solid state	Planar, 3D, MAPS, LGAD, CMOS	Si		LGAD
Gas	TPC, DC	RPC, MPGD	RPC, MPGD, DT, MWPC	TPC, DC, MRPC
Scintillator	SciFi, SiPM	Tiles, fibers, crystals	Panels	
Noble liquid		LAr		
Cherenkov		Quartz fibers		RICH, TOF, TOP, DIRC

Table 2. Summary of technologies and applications for future projects.

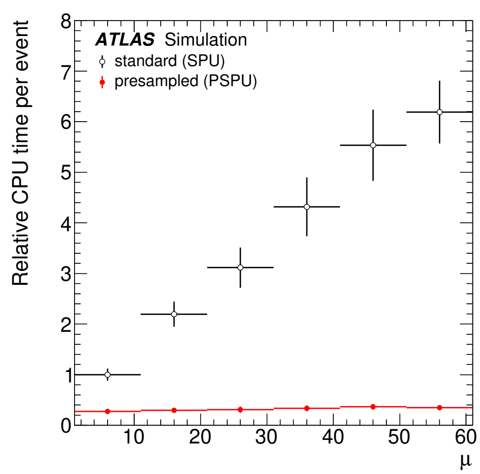


Figure 1. Comparison of the average CPU time per event in the standard ATLAS pileup digitization (black open circles) and the pre-sampled pileup digitization (red filled circles) as a function of the number of pp collisions per bunch crossing (μ). The CPU time is normalized to the time taken for the standard pileup for the lowest μ bin. Taken from Ref. [54].

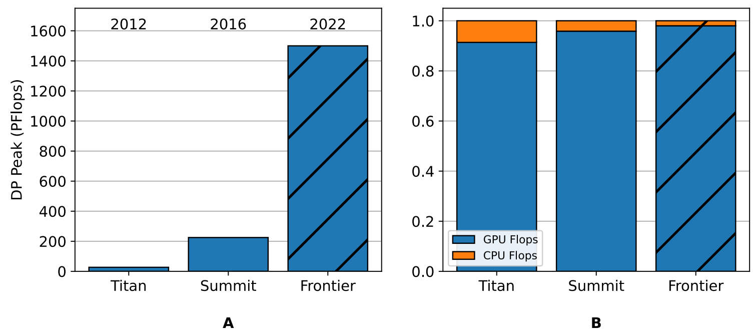


Figure 2. Peak performance in Flops (**A**) and fraction of Flops provided by GPU and CPU (**B**) for GPU-accelerated systems deployed at the OLCF. The peak performance for Frontier is projected.

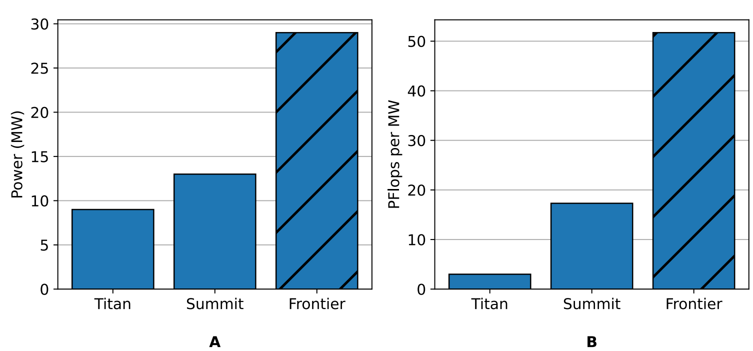
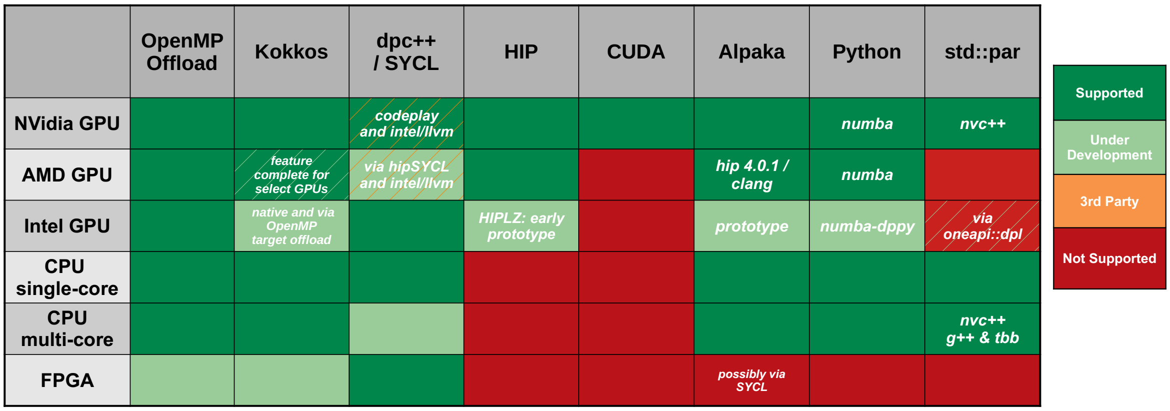


Figure 3. Power consumption (**A**) and Flops per MW (**B**) for GPU-accelerated systems deployed at the OLCF. The power requirements for Frontier are projected.



	OpenMP Offload	Kokkos	dpc++ / SYCL	HIP	CUDA	Alpaka	Python	std::par
NVidia GPU			codeplay and intel/llvm				numba	nvc++
AMD GPU		feature complete for select GPUs	via hipSYCL and intel/llvm			hip 4.0.1 / clang	numba	
Intel GPU		native and via OpenMP target offload		HIPLZ: early prototype		prototype	numba-dppy	via oneapi::dpl
CPU single-core								
CPU multi-core								nvc++ g++ & tbb
FPGA						possibly via SYCL		

Figure 4. Portability solutions for heterogeneous architectures.

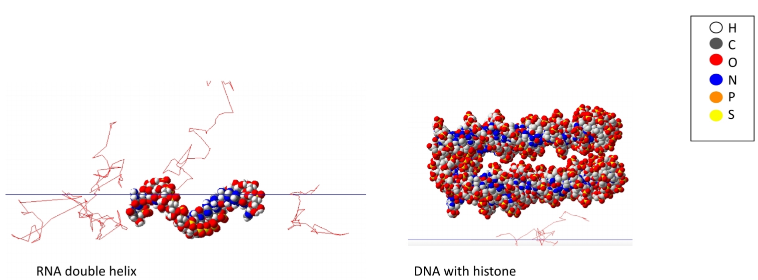


Figure 5. Molecules from the protein data bank read into TOPAS-nBio with a proton track (blue) and secondary electrons (red). Two nucleic acids are shown; an RNA strand and a nucleosome.