PROCEEDINGS OF SCIENCE

The Future of Software and Computing for HEP

Pushing the Boundaries of the Possible

Elizabeth Sexton-Kennedy[1]
Fermi National Accelerator Laboratory
Batavia, IL, USA
E-mail: sexton@fnal.gov

The nature of computing is changing. Driven by the solid state physics of CPU technology, industry is moving to multi-core systems with less memory, power, and memory bandwidth per core. The result is that the pleasantly parallel HEP event processing paradigm has to adjust to this new reality. New techniques such as machine learning and algorithms capable of exploiting vector processors, will be needed. Advances in instrumentation enabling fine-granularity high-precision measurements, have driven a data revolution. In this paper I will capture the needs, and where we are in addressing these data and compute challenges.

The 39th International Conference on High Energy Physics (ICHEP2018)
4-11 July, 2018
Seoul, Korea

[1] Speaker

1.Introduction

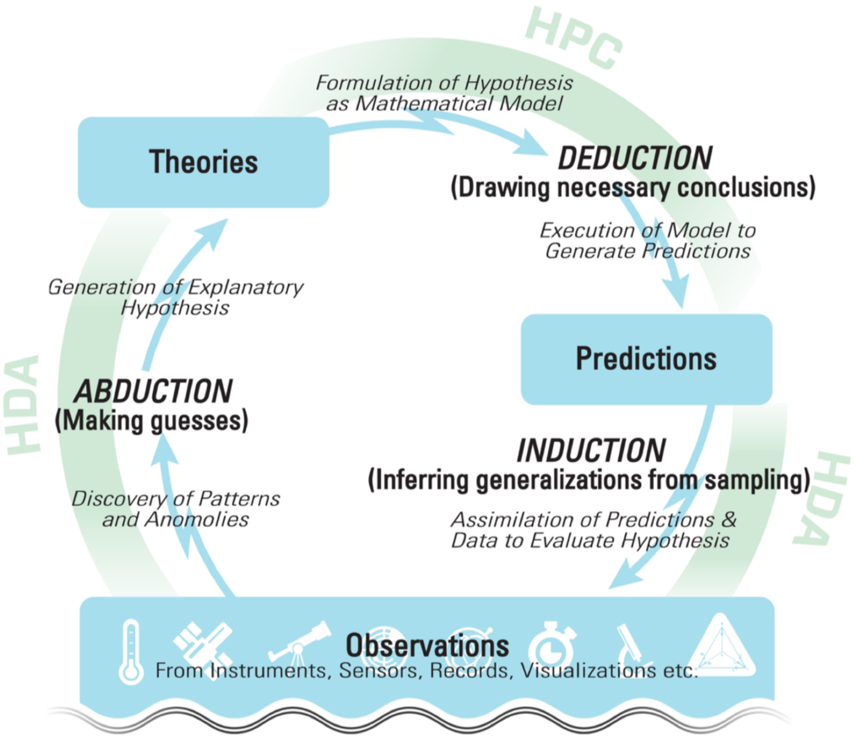
HEP experiments have long been at the leading edge of data movement and storage technologies. While it was challenging to think that we needed a Petabyte store of data to do analysis of Tevatron data in 2005, the experiment did reach that scale at about that time. No other field needed that. Fast forward 13 years later and many sciences and commercial service providers need 100 times this scale. In addition, in 2005 the data could be made available at a central facility like Fermilab. Today, it must be stored at data centers around the world. The number of participants and their growing data and compute requirements now make this impractical to do in any other way.

Computing must become a community activity driven by a goal to extract as much science as possible from computing facility investments made at laboratories and universities around the world. Thankfully, technological advances in networking and storage have evolved to meet the challenges created by the large scale instruments and scientific programs of the current era.

2. A Data Centric Vision for the Future

HL-LHC, SKA, DUNE, LIGO, LSST are or will all be data intensive science experiments. While we know their computing challenges are equally large, others, outside of HEP are planning to build exascale compute. An exascale computer is one capable of at least one thousand petaflops, or 10 to the 18th floating point operations per second. The US, China, Japan, and European HPC communities have plans to reach exascale computing by 2023. Building exascale data facilities is a challenge the experimental science community will have to drive for itself.

There are two visions currently competing for resources. One is to concentrate computing power into highly power-efficient exascale facilities. The other is to use a large number of more conventional facilities connected by high performance networking, all feeding from a data ocean. I do not believe this is an either or proposition for HEP; we will need both. As discussed in a white paper authored by an international group of high performance computing (HPC) experts, [1] a combination of HPC and high data analytics (what we call high throughput computing, HTC) are needed in the different phases of uncovering scientific insight. Combining HPC and HTC applications and methods in large-scale workflows that orchestrate simulations or incorporate them into the stages of large-scale analysis pipelines for data generated by simulations, experiments or observations is what we will need to do as we drive for a more precise understanding of the standard model.


HPC
Formulation of Hypothesis as Mathematical Model
Theories
DEDUCTION
(Drawing necessary conclusions)
Execution of Model to Generate Predictions
Generation of Explanatory Hypothesis
Predictions
HDA
ABDUCTION
(Making guesses)
INDUCTION
(Inferring generalizations from sampling)
HDA
Discovery of Patterns and Anomolies
Assimilation of Predictions & Data to Evaluate Hypothesis
Observations
From Instruments, Sensors, Records, Visualizations etc.

2.1 A Sampling of Data Needs

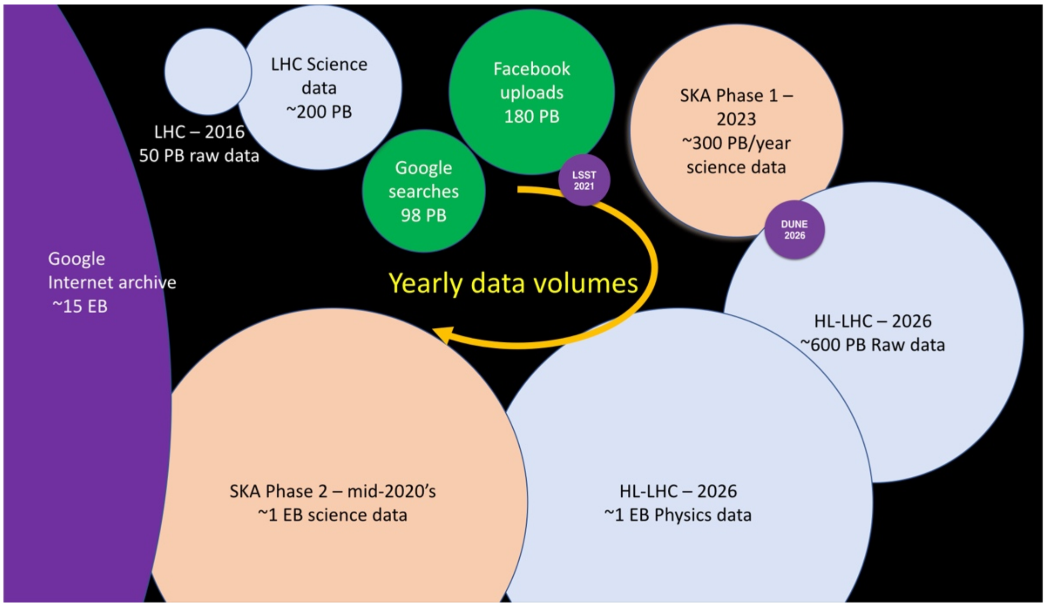
International science requires international data movement and storage. The LHC computing grid is both an example and a foundation on which to build an exascale data facility. This will be an expensive enterprise meaning it will have to be shared with all of the international data intensive sciences. This trend has already started with both DUNE and Bell II using LHC networks and co-located computing facilities.

In order to get a sense of the scale of needs, I'd like to sample from a number of experiments in particle and astro-particle physics planned to be in operation in the mid-2020s. CMS and ATLAS project that in order to support the number of existing collaborators with data formats similar to the ones in use today they will need 5 exabytes of disk storage each. Work is ongoing to reduce the size of data analysis formats while maintaining their usefulness for physics analysis. However, no matter what, the raw data from the HL-LHC will need to be stored and archived. In 2026 that is projected to require 600PB per year.

One of the major neutrino experiments operating on that time scale is DUNE. DUNE has the capability of generating an impossibly large amount of data. If all channels are completely read out (no zero suppression), at the full rate of the DAQ bandwidth limit, continuously over a year, DUNE would collect 150exabytes / year. Fortunately that level of detail for the full detector is only need for recording super nova events. Suppression of 39Ar decay, cold electronics noise, space charge effects, and argon impurities in these new liquid Ar TPC detectors, need to be studied and understood well enough to allow for effective zero suppression strategies in the readout. The target for DUNE is to produce 30PB per year.

LSST will conduct a deep survey with a frequency that results in taking repeat images of every part of the sky every few nights in multiple bands for ten years. They plan on collecting

50PB per year. SKA is a software telescope operating in complementary bans of the electromagnetic spectrum. It plans to collect 300PB per year.



The above figure summarizes these data needs projections, and compares them to the Google and Facebook yearly data volumes in 2018. Providing software and computing capable of extracting science from these unprecedentedly large data sets is a large challenge for the data science community. It would not be possible to do this with the technologies we have in use today. Therefor a focused R&D plan investigating and incorporating new technologies, is needed in order to effectively carry out improvements, over the next few years, to meet this challenge.

3. A Community Challenge and Response

In January of 2017 the Hep Software Foundation organized a kickoff workshop to seed the creation of a software and computing plan for the coming decade. A series of following workshops were used to get community buy-in, and authorship participation in, a white paper document [2]. It was released in December of 2017. Inspired by the P5 process [3] and guided by its goals, the scientific software and computing community white paper (CWP) provides a roadmap to extend commonality to a broader set of software. It is a 70 page document containing 13 topical sections summarizing R&D in a variety of technical areas for HEP Software and Computing. Almost all major domains of HEP Software and Computing are covered including one section on Training and Careers. With 310 signatories from 124 HEP-related institutions, it represents a broadly representative community view.

As presented at the conference, I will detail a subset of these 13 topics that are of most interest to the ICHEP audience. These are the areas that require HEP domain specific knowledge to effectively contribute to.

3.1 Simulation

Simulating our detectors today consumes huge computing resources in both the energy and intensity frontiers. Atlas uses 300,000 cores, CMS 200,000 cores continuously mostly for the creation of simulation datasets.

Simulation is an area in which we already have an engine and related community tools, the Geant family of tools, and Geometry modelers. The simulation chapter of the CWP makes the case that the best strategy is one in which Geant4 remains the workhorse that advances in GeantV can be back ported to, including an option to use the vectorized transport engine if that R&D demonstrates significant advantage over the default engine. G4 needs continued investment in physics and technical performance, that is continually validated by the experiments. This can only be done with an evolutionary approach. Therefor the main R&D topics identified covers (also where applicable I have added references to talks at ICHEP2018 where people have presented early results on these topics):

1. In adapting to new computing architectures can a vectorized transport engine actually work in a realistic prototype? How painful would evolution be (re-integration into Geant4)? Are there other strategies for improving technical performance on emerging computing architectures e.g. Single Instruction Multiple Data (SIMD) vectorization, Non-uniform Memory Access (NUMA) hierarchies, and offloading to accelerators like, Graphic Processing Unit (GPU), Field Programmable Gate Array (FPGA), and Tensor Processing Unit (TPU)?
2. Will experiments adopt common solutions for detector geometry descriptions that service the needs of simulation and experiment reconstruction, as was the goal of the DD4hep project?
3. Will machine learning play a new role in the domain of fast simulation? Can we develop a common toolkit for tuning and validation of fast simulation? [4] [5]
4. Review the physics models assumptions, approximations and limitations of the simulation engine. Can the validity of these models be evolved to achieve higher precision, and extended up to FCC energies? [6]
5. Can we share techniques for background modeling, including contributions of multiple hard interactions overlapping the event of interest (data overlay, ML)
6. Can we explore opportunities for code sharing among experiments when sharing the same experiments are sharing readout electronics? Can we re-engineer digitization algorithms to improve performance by means of vectorization and sub-system parallelization techniques?

The CWP brought a more consistent view and work-plan among the different projects and experiments.

3.2 Software Triggers and Reconstruction

Whether it is reconstructing the simulated data or the detector data, reconstruction will be the most expensive processing step for many experiments operating 10 years from now. The reason for this change is that the complexity of the events is increasing or the highly granular nature of the newly designed detectors is increasing or both.

Moving offline software reconstruction into trigger farms is already a key part of the program for LHCb and ALICE in Run 3. ‘Real time analysis’ increases signal rates and can make computing more efficient in terms of storage and CPU. The main R&D topics identified by this working group are the following:

1. Can we control charged particle tracking resource consumption and maintaining current phyics performance?
2. Do current algorithms' physics output hold up at pile-up of 200 (or 1000)?
3. Can tracking maintain low pT sensitivity within budget?
4. Detector design itself be a big impact [7]. How much will be gained in necessary speed improvements by optimizing tracker layouts, and the use of timing detectors?
5. Can we improve our validation techniques by using modern continuous integration, multiple architectures (which introduce sources of non-bit-for-bit reproducibility which are stochastically insignificant) validation within reasonable turnaround times?
6. Is it possible to use common reconstruction toolkits such as ACTS [8], TrickTrack [9], and Matriplex [10] and adapt them to experiment specificities in a later step? This would allow for the sharing of rare expertise across the field if possible.

In addition to the above, just as with the simulation, the current generation of reconstruction software will need to be adapted to the newly available architectures listed in bullet one of the previous section.

3.3 Machine Learning

Neural networks and Boosted Decision Trees have been used in HEP for a long time, in for example, particle identification algorithms. In this conference, there was an entire session dedicated to presenting results of exploratory R&D aimed at expanding the use of deep neural networks (DNNs) in our field. These ten talks covered uses in data quality monitoring, simulation, and reconstruction applications, in addition to the traditional analysis applications.

DNNs are very good at dealing with noisy data and huge parameter spaces. This has driven research into the use of this tool for these applications:

1. Speeding up computationally intensive pieces of our workflows
2. Enhancing physics reach with better classification than our current techniques
3. Improving data compression by learning and retaining only salient features
4. Anomaly detection for detector and computing operations

Machine Intelligence has become a big industry, and commercial companies have developed open source tools that are accelerating HEP adoption trends. There is a Python software ecosystem developing around this high risk but high reword research. If successful, this technology will change the way we do software and computing in the field.

3.4 Quantum Computing

Almost 40 years ago, Richard Feynman was one of the originators of the idea of building a quantum computer. He said, “Nature isn’t classical … and if you want to make a simulation of Nature, you’d better make it quantum mechanical, and by golly it’s a wonderful problem, because it doesn’t look so easy.” Our field has made great use of perturbation theory, but as we require higher and higher orders of these calculations, they become computationally expensive. As experiments become more precise, the theory must also. However it is an unanswered question as to how much we need to spend on classical computers to attain adequate descriptions and predications of quantum field theories. Lattice quantum chromodynamics is an alternative approach to perturbative methods; however it is computationally expensive as well.

Over the next 10 years there will be intense R&D into quantum devices to enable general purpose quantum computing which promise the capability of easily solving problems that are hard to solve on classical computers. In the meantime there is a field of work that tries to take advantage of the mid-scale noisy sensors available today. This is the area of quantum algorithm development which has shown applicability to the problems of optimization, and machine learning in addition to quantum simulation. An example of such work is the development of a digital quantum computation of fermion boson interacting systems which are accurate and simple enough to use on near term hardware available on cloud platforms today. A first application was to simulate polarons; electron dressed by phonons [8].

4. Summary and Outlook

Stephen Hawking once said, "Intelligence is the ability to adapt to change". The only thing certain about computing in the next decade is that it will change. Our software needs to be modernized to fully benefit from new computing architectures, and new approaches like ML need to be adapted to our problems.

As seen by the large participation in the detector track of ICHEP18, there is a lot of innovation and intellectual effort going into detector design. Computing needs to be seen as the extension of these efforts necessary to extract the science, and is deserving of equal effort, side by side with the design of the detectors. Especially as detector designs have a large influence on computational complexity.

The data and compute challenges of the next decade are large, even daunting. In order to satisfy the scientific needs of our community, we will need to build unprecedented scientific facilities and capabilities. The scientific harvest that is possible with this new era of big data science, and exascale computing is extremely compelling.

References

[1] BDEC community, "*Big Data and Extreme-Scale Computing: Pathways to Convergence",* https://www.exascale.org/bdec/

[2] HEP Software Foundation, *A Roadmap for HEP Software and Computing R&D for the 2020s*, https://arxiv.org/pdf/1712.06982.pdf

[3] Report of the Particle Physics Project Prioritization Panel, *Building for Discovery; Strategic Plan for U.S. Particle Physics in the Global Context*, https://www.usparticlephysics.org/wp-content/uploads/2018/03/FINAL_P5_Report_053014.pdf

[4] *Fast calorimeter simulation in LHCb,* https://indico.cern.ch/event/686555/contributions/2976594/attachments/1680991/2701012/ICHEP_CaloFastSim_180705.pdf

[5] *New approaches using machine learning for fast shower simulation in ATLAS,* https://indico.cern.ch/event/686555/contributions/2976595/attachments/1681126/2700887/ATLAS_Fast_Shower_ICHEP_2018.pdf

[6] *Geant4 Detector Simulations for Future HEP Experiments*, https://indico.cern.ch/event/686555/contributions/2976611/attachments/1681082/2700810/FHARIRI_ICHEP_06.07.2018.pdf

[7] https://pos.sissa.it/archive/conferences/282/194/ICHEP2016_194.pdf

[8] A. Macridin, P. Spentzouris, J. Amundson, R. Harnik, *Electron-Phonon Systems on a Universal Quantum Computer*, Phys. Rev. Lett. 121, 110504 (2018) *10.1103/PhysRevLett.121.110504*