The Importance of Research-Industry Collaborations on Emerging Technologies towards Exascale Computing

A contribution from CERN openlab to the European Strategy for Particle Physics

Executive Summary

High Energy Physics is facing a shortage in computing power that has been evaluated between one and two orders of magnitude in the next ten years. To preserve the physics discovery potential and make the best use of the data that will be collected at great cost, it will be essential to explore emerging Information and Communication Technologies (ICT) in close collaboration with leading industry partners. CERN openlab has adopted a unique approach to foster this collaboration by defining and coordinating innovative projects between industry, HEP users and other research entities in hardware, software, algorithms, and education. We propose that the European Strategy for Particle Physics should consider the CERN openlab collaborative R&D model and support the establishment of collaborative programmes as efficient tools in the development of the future computing strategies for HEP.

On behalf of CERN openlab – IT Department, CERN

Alberto Di Meglio – Head (alberto.di.meglio@cern.ch)
Maria Girone – Chief Technical Officer (maria.girone@cern.ch)
Federico Carminati – Chied Innovation Officer (federico.carminati@cern.ch)

Motivation

The large increase in luminosity of the LHC with the High-Luminosity upgrade will greatly increase the amount of data to be analysed as well as their complexity. Whatever will the balance be between the online data filtering and the offline data processing and analytics, the availability of advanced computing solutions to deal with this data increase will be essential to the achievement of the physics goals of the new machine. As far as offline processing alone is concerned, current estimates, corrected for the foreseeable technology evolution, predict a shortage of a factor 10 by 2026. While the uncertainties affecting this factor are very large, history tells us that computing needs of future experiments are usually underestimated.

The evolution of ICTs is accelerating, and so is the offer of a large spectrum of solutions both in hardware and software. In hardware we are witnessing the development of all sort of accelerators and, more recently, even special purpose processor (e.g. for deep learning), with Quantum Computing being still a big unknown. Data storage is also in full evolution, with an evolving balance between spinning disk to solid state devices. In software we are in the middle of a new resurgence of the AI / DNN paradigm, and while these techniques are already well established for whole classes of problems, we are just beginning to assess their full field of application, for instance in areas such as detector simulation or data analytics.

The efficient exploitation of these new technologies will require important investments and possibly substantial changes in the computing models. Suboptimal choices will be very costly particularly at the scale of the HL-LHC computing infrastructures.

It is therefore necessary to conduct early assessment of emerging technologies in close collaboration with leading providers, in order to both understand their relevance for our field and also, possibly, to inform their development and evolution.

The CERN openlab Model

CERN openlab was created in 2001 as an explicit new way to promote close collaboration between research and industry to the benefit of both. The lightweight and efficient collaboration model with ICT industry developed by CERN openlab has allowed experimenting with computing technologies “at the bleeding edge” with substantial support from the best experts in the industry. Today CERN openlab is working on a number of project with leading technology providers and as part of the ongoing major HEP initiatives like the HEP Software Foundation (HSF), looking at promising ideas in machine learning, new computing platforms, and advanced software engineering.

The CERN openlab projects, largely self-supporting, have allowed the HEP community to perform in-depth and efficient assessments of emerging technologies much before they hit the market. In the past eighteen years, hundreds of senior experts and talented young scientists and engineers have found fertile ground to share ideas, produce innovation and put their collective intuition and expertise to the service of advancement of technology and and scientific research.

We advocate that such a model and the rich communication channels that have been put in place can be further exploited and become a reference for the type of collaborative efforts that will be needed in the future to support the increasing need of the HEP community and other sciences.

Future Perspectives

Emerging Themes

The purpose of this section is to enumerate a list of technologies that CERN openlab is already exploring or planning in close collaboration with industry and the user community. This list is neither complete nor exhaustive, but it provides an idea of the breath of the exploration needed, which CERN openlab is largely covering with its present and planned activities.

GP-GPU

Far from being a completely new technology, GP-GPUs have nevertheless only a niche role in HEP computing. Given the extreme performance enhancement that they can offer on specific algorithms and problems, both online and offline, and the sustained evolution trend of their capabilities, it will be important to continue their evaluation, both in terms of performance and under the aspect of their integration in the experiments' computing models.

Neuromorphic Computing

This is a new area of research that could provide substantial speedups for a specific class of problems and that is essentially unexplored for HEP applications. While this technology is in its infancy, early evaluation of the first hardware solutions and algorithmic evaluation with emulators would be already possible today. The applicability of such devices, possibly custom-made, for online applications could also lead to revolutionary applications for data filtering and triggering.

FPGA

FPGAs have been used since many years in online computing. However, the pace of evolution of both the hardware and the corresponding software stack is such that a continuous evaluation

of their capabilities is important not to miss the opportunities they offer for very fast data treatment. Moreover, their usage for Machine Learning training and inference and for Quantum Computing emulation opens new perspectives of application.

Quantum Technologies (Computing, Sensing, Networking)

Recent software and hardware developments have spurred a renewed interest in this technology. The success of QC will ultimately be decided by the technological feasibility of machines with many relatively stable qubits. While we cannot predict when and if this will happen, the advent of usable QC machines could provide exceptional performance improvements for given classes of problems, but also a substantial revolution in the experiment computing model. This justifies an investment in the evaluation of the possible impact of QC on our computing model. A substantial part of the work can be done with QC emulators. In other disciplines, quantum inspired algorithms developed on QC emulators have shown to be valid alternatives to classical algorithms.

Large Memories

The availability of very large memories could potentially offer substantial performance enhancements and simplify the computing models and change the I/O infrastructure. Now we do not have a clear idea of the impact of these technologies and therefore we should investigate their usage together with the online and offline software projects of the experiments.

Deep Learning

While Neural Networks have a long-established place in HEP software for data classification, the advent of Deep Learning has completely changed the landscape of the applicability of DNN to HEP research. While tools and theory are still in full development, and new applications are evaluated in many fields, from simulation via generative networks to advance data analytics, it is important for HEP to track these developments and to continue to explore their applications to various field of HEP computing, both online and offline.

Evolution of Storage Strategy Towards an Efficient Exascale QoS

The trend in the last decade has been an evolution to more transparent access to the data through large scale disk caches and limited access to data through data federations. Tape and cold storage technologies have largely been relegated to archiving and disaster recovery. The massive increase in expected data volumes will make the model of nearly all data online impossible to sustain with the expected technology improvements. Going towards Exascale storage, additional qualities of service will be needed. Active tape, cold disk, traditional disk, archival low power SSD, and high performance SSD will need to be deployed as a more granular QOS hierarchy.

DNA Data Storage

Data storage needs are evolving at an impressive rate. The standard storage needs towards HL-LHC will increase by close to two orders of magnitude. If the increasing importance of reproducibility and long-term preservation is fully acknowledged and supported, the limits of existing technology and funding will be rapidly reached. Completely new ideas have to be investigated. Recent approaches like DNA Data Storage have made giant leaps in just a few years and are predicted to be cost-competitive with tape storage within the next ten years.

HPC integration

To achieve processing speeds much beyond what is available through general purpose processors, industry and the HPC community are developing and new architectures and new strategies to oversee limitations of conventional systems, focusing on new architectures and high-end specialized accelerators, to achieve the required performance. Current studies show that heterogeneous computing may allow deploying different types of processing components within a single workflow, performing specific tasks on the best suited architectures. The capabilities offered by HPC centres will provide the right set-up.

Education and Training

One of the major challenges that the HEP community will have to face in the coming years is the increasing skills gap in compute and data sciences compared to future needs. CERN openlab has recognized this need early on and runs dedicated education and training activities in collaboration with industry and academic institutes. We therefore support and endorse the creation of new innovative education mechanisms where the close collaboration between academia, industry and research can help forming a new generation of scientists and engineers. The creation of a more permanent “advanced institute for compute and data science” at CERN advocated by many prominent members of the community would provide an excellent way to address this challenge over time.

Extensions to Other Fields

An important part of the work carried out by CERN openlab is enhancing collaborations and exploit synergies with sciences beyond HEP. A robust program of work has been established in Life Science and Medical Research, Astrophysics and Earth Sciences, and generally within communities sharing similar challenges to HEP. The collaboration across different scientific disciplines to address common problems of scaling, performance, and sustainability of large-scale computing infrastructures is fundamental. We therefore recommend supporting and integrating international initiatives to build common platforms for all sciences able to support the specific requirements of the HEP community, but with a concerted and collaborative approach.