**1. Computing Models**

(HPC/HTC, Data services,...): <https://box.in2p3.fr/index.php/s/FszQfLjcxpXnWSq>

Ranked most important topic by survey participants

**Keywords:**

* HTC
* HPC
* Heterogeneous resources
* Virtualisation / container
* DOMA/ESCAPE/XDC
* Data management
* Network
* Cost reduction - impact on computing models

**Key Questions:**

* Shall IN2P3 invest in the development of new tools to run on heterogeneous hardware (HTC/HPC) ?
* The current basis of IN2P3’s computing model is the grid (HTC) and we can make good use of HPC.  What is the best way to efficiently combine GENCI and other HPC resources with CC-IN2P3 HTC resources?
* The use of containers helps the portability and re-usability of codes on heterogeneous systems. Part of this work is done in the ComputeOps project. Is a better coordination between this group and IN2P3’s experiments in general necessary?
* The near future (e.g. HL-LHC) with the unprecedented growth of data will require and bring disruptive technologies. How can we keep provide a production system of high quality while integrating new technologies?
* Shall IN2P3 include commercial clouds in the workflows? What would be the impact?
* People currently involved in DOMA are mostly the site administrators themselves, thus having only a little time to devote to the important question of the data management and its cost. Shall we re-think this and dedicate people to this task ?
* How can we strengthen the collaboration between DOMA and ESCAPE projects ?

**Supporting Text:**

The computing paradigm is clearly changing now all over the world with the dual strong rise of both Heterogeneous hardware (HTC/HPC mixing). We clearly need to adapt our use cases to this new paradigm, so that we can benefit from the developments done by the outside world community.

As things get more complex we should aim at strengthening our best tools (that satisfy our use cases), so that we can effectively run efficiently on heterogeneous hardware  (on different sites) and make good use of the deep learning techniques that are in fast development.

IN2P3 should not be just a user but contribute to the developments of these tools to have a deeper understanding and adapt them.

We should continue to put a large effort to develop and improve our computing models in order to make the best possible use of High-Throughput Computing (HTC) and High-Performance Computing (HPC) infrastructures which will continue to be the basis of our data processing / analysis in the future. Understanding what is the best way to efficiently integrate GENCI HPC with CC-IN2P3 HTC resources is particularly important.

IN2P3 has already some experience in the usage of HPC resources, e.g. through ATLAS, but expertise needs to be developed for a broader range of scientific activities. Massively parallel computation has always been outside its scope, it has no production capable system and very little knowledge about super-computer administration is present within IN2P3. Nevertheless, the number of researchers needing that kind of resources and expertise is slowly increasing in IN2P3. Some even develop software for HPC in the frame of collaboration outside of IN2P3 and without any technical support from it.

The importance of HPC in the future challenges IN2P3 wants to address should be discussed and the lack of both equipment and expertise in this domain should be filled quickly if it becomes clear that it might play an important role in the next ten years.

Portability and reusability of codes and algorithms on HPC systems can be achieved through the use of containers. These "light virtual machines" allow encapsulating applications within its environment in Linux processes. Containers have been recently rediscovered due to their abilities to provide both multi-infrastructure environments for developers and system administrators and reproducibility due to image building.

CNRS/IN2P3 has several other R&D projects related to software and computing. DecaLog aims to promote the development of software solutions portable to heterogeneous hardware, collaborative work with laboratories and departments of computing and reinforce collaborative work with physicists. Two projects exist, ComputeOps to study the use of containers for high performance computing (HPC) and Reprises to study reproducible computing.

**Computing models - time scale**

What are the future models of computing in large experiments?

The ATLAS computing model for the Run-3 will be similar to the one used for the Run-2 with noticeable improvements in the analysis workflow and in the software. It will be reviewed in the next few years in order to adapt to the running conditions of HL-LHC. In the meantime, an ATLAS strategy document for the HL-LHC is expected in 2020. A WLCG Technical Design Report is expected for mid-2022.

A large R&D on data storage and access is ongoing (DOMA) from which emerged the proposal of Data Lakes in which data would be stored on much fewer sites and remotely accessed (possibly via caches) by analysis and/or simulation centers.

Such a model would have a very strong impact on the LCG-France sites, and in particular on the Tier 2s. Some sites actively participate in the DOMA R&D in the goal of becoming a storage site in the context of a Data Lake. Most sites have expressed that they would see little advantage to become diskless" as they would still have to manage storage for local LHC users and other experiments.

The HL-LHC data and processing model would consolidate storage resources in a smaller set of larger data centers, from O(100) we have today, to O(10); one large data center could be geographically distributed in several physical locations connected by fast enough network. It will leverage processing resources at a much larger number of heterogeneous facilities, some of which might host the data and enable the capability to process data remotely and/or cache the data in volatile storage.

On the data management side, the major step in the amount of data to be managed will require new developments to limit the complexity of the system and improve the data access experience from physicists. This R&D has started through the DOMA and ESCAPE projects and will continue over the next 5 years (and beyond) through the evaluation of solutions with testbed exposed to the HEP community and beyond (astro, nuclear reactor safety).

The DOMA-FR collaboration, which is supported by the IN2P3 institute, is the French component of the international collaboration Data Organization, Management and Access (DOMA). This collaboration aims to define the technical, operational and functional characteristics of computerized data storage systems for scientific data around 2025

Many experiments (notably HL-LHC but also LSST, CTA, etc.) have clearly identified the issues and are building their future on the ability of the DOMA project to propose solutions. It is the interest of IN2P3 to support these efforts, in order to be able to operate the service at our sites (CC-IN2P3 and IN2P3 sites / Engineers) but also make it "usable" by our experiments (Computing Model / Physicists - Physicists).

A global optimisation of these different aspects [data transfer and placement systems, along with catalogues, protocols, data explicitly staged-in or cached, remote access to the data, datasets] is necessary for the HL-LHC phase. This is done as part of the Data Organisation, Management and Access (DOMA) project, DOMA-FR consisting in the French effort.

**Cost reduction / impact on computing models**

Future challenges for High Energy Physics in the domain of software and computing are not simply an extrapolation of the challenges faced today. The needs of the HEP programme in the high-luminosity era far exceed those that can be met by simply making incremental changes to today’s code and scaling up computing facilities within the anticipated budget.

At the same time, the limitation in single core CPU performance is making the landscape of computing hardware far more diverse and challenging to exploit, whilst offering huge performance boosts for suitable code. Exploiting parallelism and other new techniques, such as modern machine learning, offer great promise, but will require substantial work from the community to adapt to our problems. Current directions include the utilization of GPGPUs, FPGAs and other state of the art computing technologies in order to lower the cost of processing, and optimized models and infrastructures for a cheaper storage solution.

Currently, the amount of resources needed in order to produce and store the required Monte Carlo (MC) samples represents about 70% of the computing budget. Understanding the amount of MC needed at HL-LHC will be critical to understand the cost, and reduce those needs, again without affecting the physics reach of the experiments. Storage is the main driver of the WLCG hardware cost. The experiment data model is in continuous evolution and will likely be revisited in preparation for HL-LHC.

It will be important to understand which formats will play a role at the start of Run-3 in 2021 and during its evolution trying to optimize storage needs and cost.

The cost of the infrastructure can be reduced by exploiting adapted quality of services from different storage technologies and therefore a study of the role of tactical storage and of different archival technologies should be made.

Thus, the scientific reach of the experiment could be limited by the rapidity with which data could be accessed and digested by the computing resources. Changes of computing technology and the increase of the volume of data to be treated will require new computing models, compatible with budget constraintsrestrictions. Currently, hardware costs are dominated by disk storage, closely followed by CPU, followed by tape and networking.

The WLCG storage and computing resource capacity is expected to grow by roughly a factor 4 in the next 10 years, if we assume flat funding and extrapolate the current growth. The computing challenge in HL-LHC will thus consist in delivering the compute and storage capacity for an affordable cost. To meet this challenge we need to understand the relationship between the performance of the WLCG system, its cost and the implications in delivering the service to the experiments. Members of ATLAS France are working in close relation with members of LCG France and DOMA France to evolve the infrastructure accordingly.

**2. Artificial Intelligence, Machine and Deep Learning**

Ranked second by survey participants; <https://box.in2p3.fr/index.php/s/FszQfLjcxpXnWSq>

**Keywords:**

* AI
* Competences / Skills

**Key Questions:**

* What is the potential of AI (and its limitation) in the different scientific domains of IN2P3?
* Can expertise on AI be transferred efficiently from one scientific domain to another within IN2P3?
* Can expertise on AI be transferred in other domains outside IN2P3 (science or industry)?
* Which of our data have a value for research in Machine Learning?
* How to build collaborations with Computer Scientists on AI?
* Do we need to create an AI platform on IN2P3 level, and how should this be constructed?
* Do we need to have access to the potential of big HPC machines and how?
* What is the need for training on AI techniques within IN2P3 and for which targets (Master students, PhD students, researchers, etc) ? What courses, workshop, schools are available or should be created to meet this need ?

**Supporting Text:**

Machine Learning and Deep Learning techniques have demonstrated their potential in many scientific fields. Physics being a historical provider and user of large amounts of data, there are many topics for which such techniques can be an interesting and efficient way to analyze data and produce new scientific results. They also are able to leverage the computing capabilities of power efficient hardware such as GPUs. Finally, the next generation of students will have to been trained to use and develop Machine and Deep Learning. Investing in this research domain would thus be a good move for the next years.

A national effort will be put in France on AI in the coming decade.

The problematic of AI in IN2P3 shall be reviewed in this context not only addressing its utilization in algorithm development but also in view of identifying the new competences in informatics that we need to acquire to use AI for our science.

AI is also gaining importance in data centers for daily operations (event detection/prediction).  More communication of the opportunities that AI could provide in IN2P3’s scientific context would be important.

Machine Learning (ML) is a rapidly evolving approach to characterize and describe data with the potential to radically change how data is reduced and analysed. Some applications will allow ameliorating the obtained physics results. Others will allow using in a more efficient way computing and storage resources, increasing the physics reach of the experiment. The democratization and the development of these techniques imply the use of new tools containing libraries and tools (e.g scikit), as well as access to high performance computing machines.

In the field of astroparticles, the increase of the performance of the detectors generates data flow more and more important. Today, in2p3 has to learn HPC techniques and deep learning to be able to actually analyze this data. Otherwise, the improvement of the experiments will have no interest.

Some efforts have been made to advance Machine Learning in a structured effort at IN2P3 level. Lots of tools can be shared in principle, but building a community beyond the boundaries of each project is difficult in practice. This document is a very brief review of the relevant topics, not an overview of all ML activity at IN2P3. Many details can be found in the Machine Learning in High Energy Physics Community White Paper arXiv 1807.02876.

**AI for data analysis and trigger**

The role of AI is growing in the whole pipeline from raw data to scientific results, not mentioning the now classic use of Boosted Decision Trees for classification:

* in the reconstruction/identification of analysis objects (e.g. clustering signals) whether at LHC, or future ILC, experiments, on telescopes, gravitational wave detectors, in extensive air shower detectors or gamma camera for medical application or supernovae
* in the final analysis  yielding the final measurements with limited systematics or in the search for new physics.

One particular difficulty at IN2P3 is the need to quantify precisely (and reduce) the impact of training models on (usually) simulated data and apply them on real data. This is the generic issue of transfer learning. One example is event generator reweighting.

In the case of trigger, part of the analysis pipeline need to be run with the additional constraint of speed. AI can be used as well, also in fast hardware implementations.

One important remark is that we are currently developing ML algorithms to analyse data from detectors designed often decades ago. We should envisage the design of a new generation of detectors designed specifically to be analysed with ML techniques.

One specificity at IN2P3 is that the data to be analysed are rarely images with square pixels, but rather more complex and specific, so that off-the-shelf algorithms like Convolutionnal Neural Network rarely apply.

**Surrogate models**

At IN2P3 we use very sophisticated and precise models (e.g N(NN)LO LHC event generator, Geant4, nuclear scenarii simulation). The price of their complexity is the large amount of resources they require (for example at least half of the WLCG grid is devoted to running simulations). AI provides different techniques to build NN based models which, once trained, promise several orders of magnitude speed-up over the traditional models, which are still needed to provide the training data. Examples are : Generative Adversarial Models for calorimeter shower simulation, or NN for electro nuclear scenarii or cross section calculation (in phenomenology).

**QC and Monitoring**

AI can also be used to monitor data taking in large detectors and spot abnormal behavior, as well in computing infrastructures.

**AI for particle accelerators**

Particle accelerators are large complex machines. AI/ML can be used there to help diagnostic, to speed up simulations, or even to help automatic tuning.

**Computing resources**

Machine Learning development can usually start on a laptop with open source libraries (Keras+Tensorflow, PyTorch). However, after further developments, several hours or days (even more) per training is needed, to be multiplied by the architecture trials and Hyper Parameter optimisations. IN2P3 members have access to the CC-IN2P3 GPU clusters (with 40 GPU) as well as various local resources (university or through collaboration with computer science laboratories). Using these resources is not plug and play, and exchange of experience and expertise building in using these resources efficiently (in particular at CC IN2P3) should be stream lined.

**Training and lectures**

Employing efficiently Machine Learning techniques requires having skills on the usage of modern ML librairies as well as having solid knowledge of the underlying statistitical concepts. IN2P3 personel and students interested in these topics would certainly benefit from dedicated lectures or workshops on ML. Training courses are proposed by several universities, generally at a Master degree level, and some are open to staff for continuous training. In addition several schools and workshops are organised on a regular basis by IN2P3 researchers and engineers. To encourage access to these training courses it would be beneficial to identify and list the existing ones within a catalog,  and to have them included in the "plan de formation" of CNRS. Uncovered needs should trigger specific training actions.

**Physics ML Competitions**

One way to advertise the specific IN2P3 problems to a larger community (in particular Computer Scientists) is to organise competitions where a dataset is released with an objective function. Two recent examples with a strong IN2P3 involvement:

* The Tracking Machine Learning Challenge, on Kaggle and Codalab, where 3D points from simulated HL-LHC collision in an all-silicon tracker are released, and the participants are tasked to associate the points into tracks.
* The PLASTICC challenge, where light curves from a variety of supernova models are released, and the participants are tasked to classify them

**AI for Gravitational wave physics**

As in many fields, machine learning and AI are likely to be transformative in astroparticle physics.

This naturally includes gravitational wave astrophysics where a number of very promising results have been recently obtained in the area of noise background rejection and source parameter estimation. The COST Action CA17137 - A network for Gravitational Waves, Geophysics and Machine Learning that gathers GW and data scientists provides a good model to accelerate the dissemination of those techniques to the production pipelines. PhD co-tutorship is also particularly efficient. IN2P3 should consider offering fellowships specifically for thesis co-tutored by two domain-science and data-science experts.

# 3. Overall infrastructure needs / development

Ranked 3rd by survey participants; <https://box.in2p3.fr/index.php/s/EXPB6t7S6wPxmcB>

### Keywords:

* IN2P3 computing platforms
* Tier system for HEP
* Global IN2P3 model
* Large cloud deployment / industry standard
* EOSC/EGI
* Need for HPC resources
* Network
* Cost and funding

### Key Questions:

* Do we need a stronger investment into HPC resources at IN2P3?
* What is the role of the different computing platforms within IN2P3 in the future?
* Is the Tier system still the right answer to the HEP computing needs?
* How to best follow up on / implement the change of nature of the Tier-2s? Is it a national question or a question that each site see for itself?
* Should LCG-France evolve to also coordinate or advise on analysis resources?
* Do we need a global IN2P3 model regarding infrastructures and available person-power for computing?
* Is the deployment of large scientific cloud a way to better mutualise services?
* A number of experiments (big compute, storage and network consumers) will share the same infrastructure as the one of the LHC experiments. Is a new coordination between these users necessary?

### Supporting Text:

Astroparticle physics seems to rely stronger on HPC resources than e.g. the HEP activities within IN2P3. Should this be considered in future investment, i.e. should a significant amount of HPC infrastructure be provided through the CC-IN2P3?

Computing platforms: As an example, VirtualData has been the basis for developing and implementing several shared services, in addition or in replacement to the laboratory-specific services.

The associated scientific cloud, based on the open-source middleware OpenStack, started as an evolution of the previous cloud effort as one of the H2020 StratusLab project partners, and it is currently made of 3500 cores open to *Université Paris Sud* users and a few external users with specific partnership. The data centre facility and the cloud became the DataCenter@UPSud mésocentre in 2017 and in the last year, this platform has been labeled as a research platform by both CNRS/IN2P3 and CNRS/INSU institutes. The cloud already gives advanced access to several scientific services such as Apache Spark and JupyterHub. Some of the services have been developed as part of an ongoing R&D for scientific experiments, and helped projects to reach production state.

Historically, WLCG and EGI have very strong connections. In practice, a large part of the infrastructure is common or co-hosted. At the national level, the roles and cooperation between LCG-France and France-Grille should be clarified.

Large scientific clouds deployment has been rising for several years. Thanks to their intrinsic flexibility, clouds are often the basis for deploying advanced data processing services and they are becoming strong assets of datacenters. In our communities, they are targeting scientific applications with specific needs such as big data processing or interactive data mining. In addition to providing numerous services, cloud computing e.g. relies on sharing of resources to achieve economies of scale, enhances collaborations between research groups, and allows small research groups to access large computing resources at a lower cost. The evolution of the LCG-France Tier 2 sites is also driven by their ability to participate in local funding projects (CPER, FEDER, LABEX, . . . ) and we note that in several instances such projects are funded on a broader scope implying non grid specific technical solutions, like cloud computing or industry adopted storage solutions.

### Network and cost

The evolution of the LCG-France sites in the context of the way data is accessed (remotely or locally, the amount of data moved) has a strong impact on the network evolution to cope with our needs. It is therefore crucial to keep a close collaboration with RENATER and to consolidate the current work on the global infrastructure cost (disk vs. tape vs. CPU vs. network vs. performance) making sure we evolve towards a sustainable national infrastructure.

### Costs /financial support

The role of the Tier in the IN2P3 landscape and their long term funding by the institute is beneficial to secure funds and help the individual labs to promote their sites getting additional funds from projects or at the regional level for example.

Over the period 2013-2022, LCG-France has been organised under the conditions of an agreement between the funding agencies (IN2P3 only for the second period), sites (except GRIF-IRFU for the second period) and the LHC experiments. It is felt that such an agreement is very beneficial in both securing funds and defining rules and duties. LCG-France would like to see a similar agreement on a similar basis be negotiated for the period that leads to the HL-LHC startup and beyond. Such an agreement should include funding at least at the current level. Tier 2 and Tier 3 are also funded from local projects at laboratory, university/school or regional levels. It is important to maintain an organisation that facilitates and possibly attracts such funding projects. In the same spirit, there are Horizon2020 funded projects that are related to WLCG (either promoting R&D useful to us or promoting solutions from our community for use in other communities) that can help support the effort in our sites.

Since there is great tension in being able to fund the resources needed in future experiments for example for HL-LHC, it is important to continue the current effort in France to participate in the costing of facilities and experiment workflow as part of DOMA-FR.

It is well known that simple extrapolation of the current models and technologies fall short by a factor 5-10 of the HL-LHC needs considering a flat budget. Intensive tape usage (``tape carousel”) could reduce the need for data on disk. On the other hand the economic future of this technology is uncertain.

### Tier platform

LCG-France is mostly funding and coordinating the use of pledged resources with some exception like the Analysis Facility (AF) at CC-IN2P3. For analysis, physicists are relying on the AF or local resources at their institutions (on- or off-grid) or at CERN. The demand for resources is growing and evolving (e.g. for GPU, see below). Should LCG-France evolve to also coordinate or advise on analysis resources?

GPU platforms

The demand for GPU resource is steadily increasing and LCG-France as an infrastructure provider should investigate solutions, in particular understand if access should be part of the distributed computing model (e.g. GPU/grid queues) or in dedicated off-grid T3 facilities.

The usage can grow but user will be faced with the inevitable heterogeneity of the resources. Indeed, it is documented that performances depend not only on the hardware but also on firrmware, drivers and host OS type and settings. These arguments would favor dedicated off -grid facilities like the GPU farm at CC-IN2P3 unless there is a strong push from the LHC experiments towards the grid solution.

#### *Shared infrastructure*

Large non-LHC HEP experiments have recently started (e.g. Belle II) or will start relatively soon (e.g. DUNE). Most use the same compute and storage centers together with the LHC experiments. As a consequence, in a submission to the European Strategy, the chair and deputy of WLCG proposed a new organisation. It was proposed that the operational part of the current WLCG organisation (GDB, monitoring, accounting and working groups) would be opened to non-LHC experiments. That would allow a better coordination of infrastructures and a re-use of computing tools developed by the LHC community. If validated, such a change of WLCG organisation raises the question of the validity of the current LCG-France organisation.

In the past few years, there have been concerns within the LCG-France community about other HEP and astro/cosmo experiments and their computing models. Even within IN2P3, they vary greatly in their computing needs and organisation (Belle II, CTA, DUNE, Euclid, LSST,...). Even if a large computing center like CC-IN2P3 can cope with the diversity of needs, Tier 2 administrators would not easily be able to support very different tools and computing requirements.

This would suggest that a loose coordination of experiment computing experts and R&D effort representatives would be beneficial at IN2P3 level (and possibly at a higher level as well). This would include also the aspect of powerful network connections.

#### *Special requirements by Gravitational Wave Physics*

The future of astroparticle physics includes a range of large-data science experiments, such as CTA, LSST and the third generation of gravitational-wave (GW) observatories (Einstein Telescope in Europe). The amount of data storage and computing that those experiments will need is orders of magnitude larger than what is dealt with today. It is clear that a common international data management and computing infrastructure capable of supporting all of these would be mutually beneficial to the experiments, the computing facilities, and the funding agencies. The highluminosity LHC experiments face similarly-daunting computational scaling problems over the same time frame. There are thus obvious synergies to exploit in this area.

The demand for data analysis computing for third-generation gravitational wave detectors (our focus here) will be driven by the high number of detections (up to hundreds per day compared to few per week today) and a expanded search parameter space (0.1 solar masses to 1000+ solar masses) due to the larger observational frequency bandwidth going down to few Hertz. GW signals will last hours to days in the detector (as opposed to seconds/minutes currently). These longer-duration signals will strain physical memory abilities and lead to a steep increase in the computing demand.

In a recent review [1], it is estimated that ~300 million CPU core-hour per year are required to complete the science associated with the current GW detectors (second generation). More than 3 orders of magnitude more CPU and RAM will be required for the third generation barring an algorithmic breakthrough in the way the analysis is performed. Assuming a conservative 10 %/year computing performance/cost improvement at constant cost during the next decade (following projections of WLCG), it is almost certain that the current approach will not scale.

This calls for significant changes in the computing infrastructure (going from a centralized system around few large to mid-scale computing centers to a more distributed system) and software infrastructure (optimization of the methods, codes and workflows; hardware portability and compatibility to heterogeneous platforms).

The review [1] provides recommendations for the funding agencies to anticipate the necessary future changing by encouraging more domain-science and computer-science interactions, and by creating discussion forums for exchange know-how between the HEP and astroparticle and GW communities that will face similar challenges.

[1] Bird, I et al., [Gravitational-Wave Data Analysis Computing Challenges in the 3G Era](https://gwic.ligo.org/3Gsubcomm/documents/Gravitational-Wave_Data_Analysis_Computing_Challenges_in_the_3G_Era-July2019.pdf), 3G

subcommittee reports, 2019. https://gwic.ligo.org/3Gsubcomm

# 4. Accelerators (GPUs, FPGAs, ...)

Ranked 4th by survey participants; <https://box.in2p3.fr/index.php/s/N4J3MmPRa5DbZG2>

### Keywords:

* Usage

### Key Questions:

* How to estimate the investment needed to use accelerators, and whether this investment (mainly in RH) is justified

### Supporting Text:

Accelerators are an excellent example of a direction that physics communities are taking, in terms of the limitations we are facing using standard architectures and the need to involve different technologies. They are a good example of a diversity problem we have with such devices: there are few standards yet, and applications may have to be written differently for different devices. That is a real challenge in an international context, especially for experiments running on the computing grid.

A possible two-fold strategy for IN2P3 concerning accelerators might include: a) engage efforts and extend expertise in programming algorithms with parallel processors to keep up with the raising needs in computing power, and b) select and procure suitable devices by interacting with research communities.

One potential application of extremely accelerated simulations, which is not yet accessible but could become a viable option for studying the impact of various parameters on reconstruction performance once simulation is significantly faster, is to make simulation-based reconstructions, that is instead of computing the maximum likelihood of the data in respect to a parametrized hypothesis containing a series of assumptions, we could compute the same maximum likelihood of the data in relation to a hypothesis generated on the fly by simulating the full detector, and in which we could have all detailed parametrization of the detector as desired, even in cases where creating the parametrization of the likelihood would have been non trivial.

The usage of GPU in scientific computing has become significant in recent years. GPU provide large computing power for tasks that involve moderate amount of data flow, little branching and simple, essentially parallel, algorithms. As a consequence, GPU acceleration of full detector simulation or traditional event reconstruction remains modest.

On one hand, GPU are used in event building as part of fast reconstruction at trigger level and in part of the analysis workflow (Machine Learning in particular).

#### Accelerators for gravitational wave physics

Gravitational-wave astronomy is one of the fields that can largely benefit from GPU computing.

LIGO and Virgo have carefully investigated the use of parallel GPU and MIC architectures for its most compute-intensive searches; CUDA GPUs have been the most successful and cost-effective, and were deployed at scale for the first time in science run O3 (2019). GPU allow a faster generation of spectrograms and noise classification, and thus to reduce the latency to release gravitational-wave alerts. Long-term software development costs may be higher for rapidlyevolving parallel hardware platforms (GPU, MIC, AVX512, etc.) than for traditional CPUs, given that parallel programming interfaces are less stable targets. Single-threaded CPU codes have worked on new hardware with minimal modifications for 30+ years. Data analysis on more distributed, non-dedicated computing grid/cloud platforms will require ongoing computing infrastructure investment.

# 5. Big Data (non-structured data)

Ranked 5th by survey participants; https://box.in2p3.fr/index.php/s/Wtg7CpG6kCK4dwZ

### Keywords:

* BigData
* Internet of Things (IoT)

### Key Questions:

* How to estimate the investment needed to use big data tools, and whether this investment (mainly in RH) is justified

### Supporting Text:

The HEP community has been involved since a long time in the data analysis of massive amount of data, and in a way was a precursor in that field. Others such as GAFA have been developing some Big Data tools that could be potentially beneficial for our scientific community as well. It should also be an opportunity for the HEP community to evolve. Even though HEP collaborations have a long-standing experience in dealing with the so-called data deluge, the work that is being achieved in other scientific communities or elsewhere is significant and certainly will be in the future. Hence, we need to be open to other experts in the Big Data and data management field.

We need to improve our knowledge of these technologies, for example for log and monitoring data management (HADOOP, ElasticSearch, ...).

The Internet of Things (IoT) is not yet identified as a topic of interest for IN2P3 computing and data experts. However, IoT technologies open new perspectives for large-scale ground experiments requiring the remote control of multiple sensors.

Wherever the installation of a local WiFi is not feasible, for instance in remote locations of interest for Astroparticle Physics, low-power wide area LoRaWAN network technology provides secure, low-power and long-range communication. Its wireless signals can reach across large distances (up to several kilometers), delivering tiny packets of data to and from multiple low-power node devices.

IN2P3 could lead the deployment of large-scale sensor networks for scientific research. Only

IN2P3 within CNRS has the skills and expertise to master all the steps from the sensors to the cloud where the data are made available to the scientific community. For instance, we experiment at LPC Clermont-Ferrand the use of software technologies (for instance the

ELASTIC software stack used to collect and analyze logs from IT servers) that are well mastered by IN2P3 laboratories to collect sensor data in a data lake for the observation of agroecosystems.

A new model of data analysis, developed outside of HEP, maintains the concept of sequential ntuple reduction but mixes interactivity with batch processing, such as Spark or TensorFlow, in particular with implicit parallelism. The use of such new analysis models is not yet developed in the IN2P3 scientific community.

#### Open Data

Open data would deserve a specific section in this chapter on Big Data. The political context at the French national [1] and European [2] level is rapidly evolving with major decision that will change the way we (IN2P3) do science in the next decade. Soon, there will be a legal obligation to publicly-funded research to release the data to the public.

Globally speaking, the current funding scheme of major experiments does include a provision for opening the data, while this activity includes many tasks such as data preparation, curation, documentation and release that have a significant cost and should be anticipated in the budget. The role of data curator is currently missing in the type of careers that can be considered at IN2P3.

[1] Plan national pour la science ouverte, 2018.

http://cache.media.enseignementsuprecherche.gouv.fr/file/Actus/67/2/PLAN\_NATIONAL\_SCIENCE\_OUVERTE\_978672.pdf

[2] Open science policy platform, 2019. https://ec.europa.eu/research/openscience/index.cfm?pg=home

# 6. Emerging Technologies (FPGAs, Quantum Computing, ...)

Ranked 6th by survey participants; <https://box.in2p3.fr/index.php/s/2DpQBpkG7jeFnnS>

### Keywords:

* FPGA
* Quantum Computing: potential to accelerate algorithms, simulations, ...
* Quantum Computing: risk investment with difficult to evaluate benefit
* Quantum Computing: R&D
* Software

### Key Questions:

* What is the right level of engagement in emerging technologies? Shall IN2P3 follow the evolution, or take active part in its development?
* How shall we organise ourselves to better explore new technologies (mutualisation within the laboratories)?
* How can we convert a test platform into a production service?

### Supporting Text:

As could already be witnessed when CPU clock frequencies reached their peak around 2005, simple strategies for scaling up overall computing hardware performance are reaching their limits. As a result, hardware manufacturer’s main strategy for increasing computing power has shifted towards introducing a growing diversity of specialized architectures (GPUs, FPGAs, TPUs...) that are either tailored to a specific problem domain (such as neural network training & inference), difficult to program efficiently, or both. Significant effort will be needed on both the software and computing side to support this new wave of computational hardware. For example, porting of the large base of legacy software to programming models that can efficiently support current (multi-core, wide-vector CPUs) and emerging (GPUs, FPGAs, ...) hardware architectures.

With the arrival of the first quantum computer demonstrators such as those developed by IBM or Google, this field is experiencing a new boom on a global scale. The programming on these prototypes of quantum computers remains today a challenge in particular concerning adapted algorithms and control of the quantum noise inherent to these computers. Nevertheless, in the next decade, these computers as well as the new associated algorithms may hypothetically provide computing power hitherto unequaled in terms of speed and storage.

Qubit-based architectures are particularly well suited to calculate fermion networks, thus is it is likely that the advent of quantum computers will be a major breakthrough to simulate systems of interacting fermions, such as atomic nuclei or network-described quarks. Thus, quantum computers will be a priori a breakthrough solution for the theoretical physics of IN2P3 themes. Programming on quantum computers requires to rethink completely the usual algorithms. In perspective of this arrival, algorithms associated with the simulation of N-body problems can already be developed and validated on existing quantum computer simulators. At the international level, the United States have already shown their willingness to develop this field in particle physics, and also within CERN OpenLab a working group on this topic has been established.

The integration of new resources like *quantum computing* look promising but will require a major effort to integrate these resources transparently in the LHC software infrastructure. Quantum computing thus remains an R&D topic with probably little production use in the period covered by the prospective but it remains important to follow what happens there. In later stages, an important aspect is how test and development platforms can be converted to be ready to be integrated into the production workflows.

The impact and need to follow emerging technologies in their first years of development must be reserved to a very small group of people able to alert the community as soon as close to the market introduction. On longer term, technologies related to quantum computing hold the promise of substantially speeding up computationally expensive tasks. While significant developments are being made in the field of quantum computing, today’s hardware has not yet reached the level at which it could be put into production within our community. It remains difficult to foresee when more stable hardware - capable of providing concrete benefits for the HEP community - will be available. Given both the potential and the uncertainty surrounding quantum computing, it is important to explore what these new technologies could bring to our field.

# 7. Other

<https://box.in2p3.fr/index.php/s/ZTdsGTksHeDicx7>

Ranked least important by survey participants (although a large number of contributions finally ended up in this category)

### Keywords:

* Quality Assurance
* Software Quality
* Software developments (including tools for end users)
* Data life cycle and data curation
* Computing science and data policies
* Competences and skills
* EU projects
* Recruiting, training, careers

### Key Questions:

* How we can improve our impact on decision processes on the national and international, and how can we engage colleagues in this process?
* What would be the best channel to coordinate efforts between IN2P3 and different EU projects?
* How to better collaborate with theorists on software development of the generator codes?
* Shall we invest in an institute-wide effort on s/w quality, in its general sense?
* Do we need to identify the required and missing competences and expertise on IN2P3 level in order to be able to respond to the needs in the upcoming years?
* Is the improvement of the data life cycle a key subject for IN2P3?
* How can we stimulate publications and how can we change the factors that are important for the career?

### Supporting Text:

### Software Quality

Software quality is becoming a real requirement for many experiments. Traceability, provenance, continuous integration, coding organisation, and fine production control are in the heart of new requirements, as software is now as important (or even more sometimes for Big Data Factories) as hardware. The quality of the software comes also from the quality of the programming that has an impact of the running cost of any experiment, but also on the environmental impact of Computing Centres. A bad code that runs 2 times longer than a good code leads to the production of more CO2 via CC-IN2P3’s energy consumption and heat release. The side effect of such bad code is also the reduction of the CPU nodes availability, and thus it is a cost for the community.

While data analysis in CERN has already been promoted to a high level, many applications suffer from a lack of design, which can be related to a lack of communication between IT engineers & physicists/other engineers.

An institute-wide effort on s/w quality, in its general sense, is now needed as it impacts infrastructures and Big Data management.

Another aspect of software quality is the control of floating point computing precision, mix-precision computing, validation and reproducibility of numerical results.

### Software development

A solution to dramatically reduce the computation time is to make the most out of the capabilities of modern CPUs. To do so, software must use vectorisation and make the most of data prefetching techniques.

Cache prefetching is a technique used by computer processors to boost execution performance by fetching instructions or data from their original storage in slower memory to a faster local memory before it is actually needed. So prefetching data and then accessing it from caches is usually many orders of magnitude faster than accessing it directly from main memory. This is automatically done by the CPU, but if one does not provide a correct memory pattern in the software, the program will not take advantage of it.
 CPU manufacturers developed SIMD instructions based on vector registers to improve CPU use. The main advantage of SIMD is that processing multiple data elements at the same time, with a single instruction, can dramatically improve performance.
 Compilers try to vectorize code if asked for it. Again, if one does not provide a correct memory pattern in the software, the program will not take advantage of it.

To improve the computing needs, one could rely more on event filtering and reweighting, or to improve the parallelism and concurrency of the generator code, allowing the software to exploit at best modern hardware and facilities. A strong collaboration between theorists and software experts is needed.

Currently, the French sites are operated mostly using grid specific solutions. We are not involved in the development of these solutions, thus IN2P3 has little impact on their evolution. We should discuss as a community if there are projects were we would like to have a stronger involvement.

France has a larger involvement in higher level grid middleware or experiment software developments, to cite a few: DIRAC, the Operation Portal, AMI. . . There are other tools like Rucio (data management), which seems to gain a wide adoption and has adopted an open development model so speci c contributions could be planned. More broadly, the DOMA R&D activities have lots of opportunities for contributions, even beyond IN2P3 (e.g. INRIA or other computing researchers).

Several areas have been identified where R&D is necessary to exploit the full power of the enormous datasets that will be collected. In addition to a general effort to reengineer software, there are needs to understand what kinds of algorithms are best suited to what kinds of hardware architectures. This includes GPGPUs, FPGAs and possibly custom ASIC chips. The reconstruction code is expected to become more complex, and more difficult to write; this can be partially mitigated by experiment frameworks and modern software technologies. It is an area where collaboration with the computer science community is required. Closer interactions between ATLAS France members and technical projects such as DecaLog should be pursued.

In general terms, the effort software development is spread half among many physicists involved in analysis software and half among few physicists and engineers in core software and applications.

#### *Engagement/collaboration with EU project*

The expertise acquired by the technical teams working on the LHC experiments, as well as some tools developed initially for the LHC experiments, have helped to include successfully other Virtual Organization, in particular in astroparticle, inside the same grid infrastructure. The dissemination towards future experiments started also recently and is discussed within the European Science Cluster of Astronomy & Particle physics ESFRI research infrastructures (ESCAPE) project.

The work within IN2P3 in the computing and data science domain is impacted by national and European policies and decision-making. IN2P3 has been playing an active role in this context, by participating in national and international committees. A more structured approach to influencing decision processes on these levels might be necessary in the future in order to defend our interest. A crucial point here is the fact that a very small number of colleagues at IN2P3 is willing to engage in this kind of activities. Thus, a question for the future is, how we can improve our impact on decision processes on the national and international, and how can we engage colleagues in this process.

#### *Data curation/life cycle*

Scientific data (both real data and simulations) have been growing and will grow at a very high rate. Scientific collaborations will work at the Exabyte scale before the end of the next decade. This is challenging both from the scientific point of view, the financial point of view and the technical point of view.

The improvement of the data life cycle should be a key subject for IN2P3 in the next decade, in order for the end users to be able to access efficiently their data (data management and auditing). Other scientific communities will have be able to reuse them (data curation, Open Data and archival).

#### *Recruting, training, careers*

Recognition of computing work, in particular for R&D, has resulted in a number of talks and publications. Still such involvement is complicated because of the implication of the teams in operation of the sites. Recognition of software work as a key part of science has resulted in a number of journals where developers can publish their work. Journal publication also disseminates information to the wider community in a permanent way and is the most established mechanism for academic recognition. Publication in such journals provides proper peer review, beyond that provided in conference papers, so it is valuable for recognition as well as for dissemination. However, this practice is not widespread enough in the community and needs further encouragement.

In addition, computing and software activities rely on a small number of people, which are technicians and engineers as well as physicists. Experiments and institutions should take care to recognise computing work and ensure careers able to keep experienced persons in the field and attract talented young people.

As tools become more complex, physicists, as well as software and computing engineers, must be continuously retrained in order to utilize them effectively. It is therefore imperative that training opportunities and resources are made available.

The ATLAS-France group proposes a review of existing work and R&D developments on software and computing activities within the ATLAS collaboration, with emphasize on the French contribution. Over the next decade, there will almost certainly be disruptive changes that cannot be planned for, and we must remain agile enough to adapt to these.

In LCG-France we have a network of skilled collaborators with long experience in grid computing. In between nowadays and the early HL-LHC operation, a number of Tier 2 site administrators might retire. From a recent survey, operations at most sites will continue if recruitment at a reasonable level is foreseen, with enough overlap time too allow passing on the knowledge and experience.

As alluded above, there is tendency of shifting technologies to those already widely used in the computing or academic world, this will have to be followed closely to define the profile of skills we need to operate Tier 2 sites during HL-LHC operations.

Finally, there is a general need for training programs to bring researchers up to date in the domain of scientific computing, in particular in the domains of concurrent programming and artificial intelligence. This activity should encompass different levels of trainees, from undergraduates, to young researchers up to senior physicists and software and computing engineers.