Software and computing for ATLAS

Authors : C. Biscarat, S. Crépé-Renaudin, F. Derue, L. Duflot, E. Fede, A. Formica, S. Jézéquel, E. Le Guirriec, J-P. Meyer, L. Poggioli, E. Vamvakopoulos, *Computing ATLAS France group*

As part of the national foresight exercise in nuclear physics, particle and astroparticle physics for the period 2020-2030 [1], members of the Computing ATLAS France (CAF) group [2] have written this document as a contribution to the working group GT-09 « Computing, algorithms and data ».

1. Computational challenges

The Run-3 of the Large Hadron Collider (LHC) will take place from 2021 to 2023 with proton-proton collisions at a center-of-mass energy of 14 TeV [3]. It will deliver 300 fb^{-1} of data, about twice that recorded during the Run-1 and Run-2.

The high-luminosity period of the HL-LHC (HL-LHC) will start in 2026 until 2035. Computing systems will have to cope with greatly increased data samples, data acquisition and processing rates. The machine will deliver about 3000 fb^{-1} of data, 30 times that recorded during the Run-1 and Run-2. The collision events will gain in complexity, as for every bunch crossing about 200 low-energy collisions are expected to accompany the hard process selected by the experiments' trigger system, to be compared with an average of 35 during Run-2. On top of that, ATLAS and CMS experiments are planning to expand their trigger rate from the current 1 kHz up to 10 kHz possibly.

The needs for computing resources generally scale faster than linearly, due to the combinatorial behavior of the most time-consuming algorithms like tracking and clustering. Storage scales roughly linearly with event complexity and trigger rate. Figures 1 and 2 show the estimated needed CPU and storage resources for the years 2018 to 2032 for data and simulation processing [4]. The solid lines show the amount of CPU (storage) resources expected to be available, if a flat funding scenario is assumed, which implies an increase of 20% (15%) per year, based on the current technology trends. If the expected resources for the Run-3 are of the same order as the needs, for the HL-LHC the needs exceed by a factor 2-3 the expected CPU resources and by a factor 4-5 the expected storage resources. Different scenarios, implying changes in the software could allow to reduce the computing needs. For the storage, the problem persists, even in the scenario of a drastic reduction of replica files.

2. Software & Computing working groups and collaborations

Software and computing is discussed in the ATLAS France collaboration since 2002 in the CAF working group. Its role is to ensure that computing infrastructure and software activities provided by France to the ATLAS collaboration and to the physicists of the French laboratories meet the needs to pursue their research, in particular at the CC-IN2P3 [5].

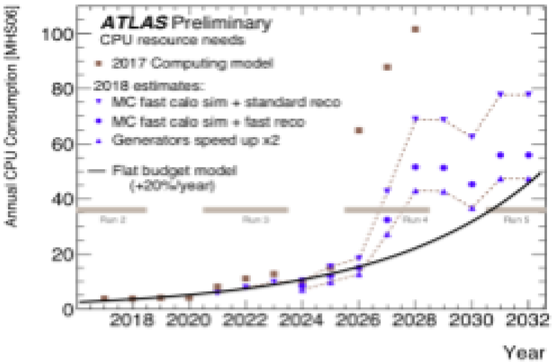

Figure *1: Estimated CPU resources (in MHS06) needed for the years 2018 to 2032 for both data and simulation processing [4].*

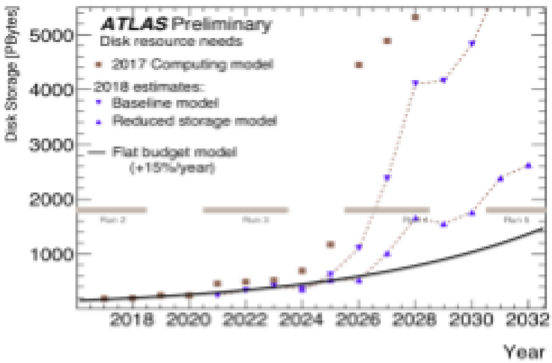

Figure 2 *Estimated total disk resources (in PBytes) needed for the years 2018 to 2032 for both data and simulation processing* [4].

The Worldwide LHC Computing Grid (WLCG [6]) is a global infrastructure whose mission is to provide computing resources to store, distribute and analyse the data generated by the LHC, making the data equally available to all partners, regardless of their physical location. The LCG France collaboration [7][8] is taking care of the development and the operation of the infrastructure in France, the Tier-1 located at CC-IN2P3 and the Tier-2s located in each laboratory. Members of ATLAS France, including members of CAF, are taking part to LCG France. A sub-project of WLCG is the Data Organisation, Management and Access (DOMA [9]) project; the DOMA-FR [10] being its French contribution. It is not specific to the ATLAS collaboration and is open to experiments outside LHC. Several members of LCG-FR sites and members of ATLAS France are taking part in this effort.

Developments in France on grid computing outside the LHC experiments, and also cloud computing, are followed by the France-Grilles and Cloud collaboration [11]. Most of French Tier-2s taking part to WLCG, 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 [12]) project.

CNRS/IN2P3 has several other R&D projects related to software and computing. DecaLog [13] 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 [14] to study the use of containers for high performance computing (HPC) and Reprises [15] to study reproducible computing. Another R&D project is CompStat [16] about Machine Learning activities.

The software and computing challenges and solutions currently considered, that ATLAS and other LHC experiments will face, are described in documents written by WLCG [17] and the HEP Software Foundation [18][19]. Many parts in the following are extracted from these documents.

3. Computing model

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.

The trigger system of the experiment will be upgraded for the HL-LHC period to select a wider variety of events and a larger event rate. Typically, the experiment target systems where the output of the trigger system is increased by an order of magnitude over the current capability.
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. ATLAS produces O(100) DAOD (derived AOD) formats from AODs through centralised analysis trains. It will be important to understand which formats will play a role at the start of Run-3 and during its evolution. The kind of information stored in each format will also be reviewed, trying to optimize storage needs and cost.
The experiment will review data retention policies based on the experience of previous runs and the evolution of the data model. 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.
Finally, the HL-LHC data processing model will also need to be reconsidered. The analysis model plays in particular an important role and the experiment will understand how much of today’s chaotic activity can be centralized and scheduled in planned and organised workflows.

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 restrictions. Currently, hardware costs are dominated by disk storage, closely followed by CPU, followed by tape and networking.

4. Experiment software

The ATLAS software environment, Athena [20], has been built to treat events in a serialized way. Today performances of the software are not sufficient to comply with available CPU. Several factors can explain this, going from the building of the code to the I/O performance. This lead to multi-core (AthenaMP) then multi-threaded (AthenaMT) implementation. With some reengineering of a part of the codes, a factor 2 gain of the global performance is expected for the HL-LHC period as seen in Figure 1.

Particle generators require an increasing level of precision, from Leading Order (LO) in LHC Run-1 to Next-to-Next-to-Leading Order (NNLO) in HL-LHC, which implies an increasing CPU time per event. 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.

Detector simulation and digitization activities are the main consumer of CPU. It is particularly well suited for parallel processing on new concurrent computing architectures. Geant4 is the common software package for detector simulation and improvements at this level will be beneficial to all experiments. Moreover, a larger fraction of HL-LHC simulation could be done through fast simulation. Beside parameterized detector simulation, various solutions should be evaluated and prototyped at all levels of the simulation chain.

Shift to newer computing architectures is equally important for software triggers and event reconstruction code. The CPU needed for event reconstruction tends indeed to be dominated by charge particle reconstruction (tracking), especially when the number of collisions per bunch crossing is high and an efficient reconstruction of low transverse momentum particles is required. The increasing granularity of detectors, and the addition of timing information, which seems mandatory to cope with the extreme pileup conditions at the HL-LHC, will require new kinds of reconstruction algorithms that are sufficiently fast for use in real-time. The development of efficient techniques to reconstruct and identify physics objects in a complex environment has started. The implication of physicists both in the detector upgrade program and in the software would benefit to this effort.

Several areas have thus 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.

5. Data analysis

Methods for analyzing the data have been developed over the years to produce physics results during Run-1and Run-2. The most common approach to analyze data is through a campaign of data reduction and refinement, ultimately producing data frames (flat ntuples) and histograms used to make plots and tables from which physics inference can be made. 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 [21] or TensorFlow [22], in particular with implicit parallelism. The use of such new analysis models is not yet developed in the ATLAS France community.

6. Machine Learning

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 to ameliorate the obtained physics results. Others will allow to use in a more efficient way computing and storage resources, increasing the physics reach of the experiment. With the advent of more powerful hardware, particularly GPUs and ML dedicated processors, as well as more performant ML algorithms, the ML toolset will be used to develop application software that could potentially replace the most computationally expensive parts of pattern recognition algorithms and parameter extraction algorithms for characterising reconstructed objects.

Members of ATLAS France are already involved in the use of ML for simulation or combined reconstruction of the detectors (e.g triggering, fast simulation of electromagnetic showers, tracking…), performances studies (e.g hyperscan parameters) and for final analyses. The democratisation 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. The French groups are already using GPUs located in their laboratories, universities or at CC-IN2P3.
This implies an effort of training, and training offers, to use these algorithms, implement existing code and use high performance computing machines. Closer interactions with the CompStat project, for example, should be pursued.

7. Data organisation, management and access

Data organisation is essentially how data is structured as it is written. Most data is written in files, in ROOT format. In the past, the key challenge for data management was the use of a distributed computing in the form of the grid. The experiments developed dedicated data transfer and placement systems, along with catalogues, to move data between computing centres. Originally, computing models were rather static: data was placed at sites, and the relevant compute jobs were sent to the right locations. Since LHC startup, this model has been made more flexible to limit non-optimal pre-placement and to take into account data popularity taking advantage of the good network between sites.
For data access, historically, various protocols have been used for direct reads (rfio, dcap, xrootd, etc.) where jobs are reading data explicitly staged-in or cached by the compute resource used or the site it belongs to. With direct

access, applications may use alternative protocols to those used by data transfers between sites. In addition, LHC experiments have been increasingly using remote access to the data, without any stage-in operations, using the possibilities offered by protocols such as xrootd or http.
Concerning the data granularity, the data is split into datasets, as defined by physics selections and use cases, consisting of a set of individual files. While individual files in datasets can be processed in parallel, the files themselves are usually processed as a whole.

A global optimisation of these different aspects 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.

8. Computing and storage infrastructure

Various computing resources are used : WLCG, cloud (academic or private), batch clusters and interactive machines, HPC (High Performance Computer) machines in large computing centres (CERN, CC-IN2P3 ...), in universities or laboratories.

The distributed computing system of the ATLAS experiment is built around two main components: the Production and Distributed Analysis system (PanDA [23]) and the data management system Rucio [24]. It manages the computing resources to process this data at the Tier-0 at CERN, re-processes it once per year at the Tier-1 and Tier-2 WLCG grid sites and runs continuous MC simulation and reconstruction. In addition continuous distributed analysis from several hundred ATLAS users is executed. The resources used are the Tier-0 at CERN and Tier-1/2/3 Grid sites world-wide and opportunistic resources at HPC sites, cloud computing providers and volunteer computing resources.

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 make evolve accordingly the infrastructure.
Networking will play a central role in HL-LHC. WLCG should continue engaging with Funding Agencies and NRENs ensuring enough capacity is made available and the LHC traffic does not get segregated below a critical level. The R&D should contribute defining what that critical level will be in the mid 2020s. 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.

A cost model need to be build which can be used to understand the impact of different future strategies quantitatively. The model should take into account the cost of the hardware, infrastructure and operations and provide a quantitative assessment for any proposed change in terms of computing model, workflow model, data placement, data access and data processing strategy, offline software evolution. Some members of LCG-FR French sites participate to these working groups building such a cost model usable for all LHC experiments.

CC-IN2P3 plays a particular role for computing and storage resources for ATLAS France. Besides its contribution to WLCG as one of the leading Tier-1, it offers also analysis facility through its interactive machines, batch cluster, HPC and GPU clusters. The availability of storage resources at CC-IN2P3 is also crucial for the ATLAS France collaboration, for home directories, semi-permanent storage as well as disk on grid and tapes. The evolution of CC-IN2P3 to keep its leading role in the coming years is of crucial importance.

Each laboratory offers to its members some local resources. Most of laboratories have interactive or batch clusters which represent several hundred of CPUs, shared with other experiments. The level of semi-permanent storage or disk on grid was found to be of several hundred of TB. The availability of these resources is of high importance for the final analyses.

The grid is the main infrastructure used by the ATLAS collaboration for both production and analysis because it is up to now the best tool to fulfill our needs. However, as mentioned previously, ATLAS experiment is able to use other kinds of infrastructure like cloud or HPC centers. As of today, these possibilities have not been exploited in France except for tests or local and short collaborations. These potential resources should not be neglected in the future and creating or strengthening the links with other operators may lead to fruitful collaborations.

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.

9. Conditions data and Metadata

Conditions data is defined as the non-event data required by data-processing software to correctly simulate, digitise or reconstruct the raw detector event data. The non-event data discussed here consists mainly of detector calibration and alignment information, with some additional data describing the detector configuration, the machine parameters, as well as information from the detector control system. Conditions data is used by event processing applications running on a very large distributed computing infrastructure, resulting in tens of thousands of jobs that may try to access the conditions data at the same time, and leading to a very significant rate of reading (typically O(10) kHz). ATLAS and CMS are working on a common next-generation conditions database. ATLAS is migrating, in preparation of the Run-3, current implementations based on COOL to the proposed REST-based approach.

The EventIndex is the complete catalogue of all ATLAS events, real and simulated, keeping the references to all permanent files that contain a given event in any processing stage. It provides the means to select and access event data in the ATLAS distributed storage system, and provides support for completeness and consistency checks and trigger and offline selection overlap studies. The EventIndex employs various data handling technologies like Hadoop and Oracle databases, and is integrated with other systems of the ATLAS distributed computing infrastructure. The main data store has worked well during Run-2 but new solutions are explored for the future.

The ATLAS France collaboration is the main contributor to the ATLAS Metadata Interface (AMI), a generic framework for metadata catalogues. This tool is now being used by non-LHC experiments.

10. Data and software preservation

Each of the LHC experiments has adopted a data access and/or data preservation policy [25] with public access to some subset of the data in a highly reduced data format for the purposes of outreach and education. ATLAS data associated with journal publications is made available and strives to make available additional material that allows reuse and re-interpretations of the data in the context of new theoretical models. ATLAS is exploring how to provide the capability for reinterpretation of searches in the future via a service in which the original internal analysis code (including full detector simulation and reconstruction) is preserved, as opposed to the re-coding approach with object-efficiency calibrations used by external reinterpretation toolkits.

11. Training, staffing and careers

The field of high energy physics has consistently exploited the latest innovations in computational tools and technologies for processing data. The software tool set of the particle physicist is ever-growing, and the problem set increasingly complex. These tools and R&D activities in ATLAS, as well as the ones done in related projects such as WLCG, are at the cutting edge of technology. As these 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.

Among the different existing training opportunities, the ones available from IN2P3 and IRFU are listed. A collection of online tutorials (mostly written in

French) are hosted on the Gitlab IN2P3 server [26]. The annual IN2P3/IRFU Computing Days [27] offer an opportunity to refresh and extend the collection of tutorials every year. The IN2P3 organizes every two years a School of Statistics [28] that together with training on classical statistical methods focuses on new techniques and tools including machine learning ones. Some training programmes exist also for PhD students as part of their doctoral schools, mostly geared towards machine learning techniques or the use of GPUs.

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 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 technicians and engineers as well as physicists. Experiments and institutions should take care to recognise computing work and ensure careers able to keep expertised persons in the field and attract talented young people.

12. Conclusions and prospects

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.

ATLAS France collaboration, in particular the CAF, is working in very close relation with LCG France collaboration since about 15 years to make available the needed infrastructure for WLCG in France and to make it evolve while keeping it in production. The infrastructure at CC-IN2P3 and in the Tier-2 sites have evolved, while keeping the budget and human resource constant. Technical teams, in charge of the Tier-2 sites, are working and exchange experience through dedicated forums, in particular the LCG-FR technical one, also sharing with other LHC experiments. All the infrastructure is also

monitored by different shift teams for production and distributed analysis, including also physicists from ATLAS France.
The effort of the ATLAS France collaboration in terms of computing includes engineers and physicists involved in the operation and development of the Tier-1 and Tier-2 facilities. This implication allows the good operation level of French contribution to the ATLAS computing resources. The effort in terms of software is spread half among many physicists involved in analysis software and half among few physicists and engineers in core software and applications.

The level of implication of the various engineers and laboratories for the long term, as well as possible needs for recruitment, is listed inside the LCG France collaboration. We know that in any future scenario development effort will be constrained, so it is vital that successful R&D projects provide sustainable infrastructure and software for the future. Other R&D projects on software and computing exist in France. For some of them, such as WLCG or DOMA, strong links already exist with the ATLAS France collaboration. On the other hand, links with R&D projects on the use containers or heterogeneous hardware are much weaker. Links with machine learning R&D projects are not yet very strong.

Finally, there is a need for training programmes 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.

This document 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.

References

[1] Prospectives nationales 2020-2030, https://prospectives2020.in2p3.fr
[2] Computing ATLAS France working group https://collaborationatlasfrance.web.cern.ch/content/caf
[3] LHC/HL-LHC Plan https://project-hl-lhc-industry.web.cern.ch/content/project-schedule
[4] ATLAS computing public plots, https://twiki.cern.ch/twiki/bin/view/AtlasPublic/ComputingandSoftwarePublicResults
[5] CC-IN2P3, https://cc.in2p3.fr/
[6] WLCG project http://wlcg.web.cern.ch/
[7] LCG France project, https://edms.in2p3.fr/document/I-033398/1
[8] Protocole d’accord LCG France 2019-2022 https://edms.in2p3.fr/document/I-033924/3
[9] Data Organisation, Management and Access, DOMA project https://twiki.cern.ch/twiki/bin/view/LCG/DomaActivities
[10] DOMA-FR project

https://indico.in2p3.fr/event/18092/sessions/11087
[11] France grilles et cloud, www.france-grilles.fr
[12] ESCAPE project, https://projectescape.eu/
[13] Master project CNRS/IN2P3 DecaLog https://gitlab.in2p3.fr/CodeursIntensifs/DecaLog/wikis/home
[14] ComputeOps, sub-project of DecaLog https://gitlab.in2p3.fr/CodeursIntensifs/DecaLog/wikis/ComputeOps
[15] Reprises, sub-project of DecaLog https://gitlab.in2p3.fr/CodeursIntensifs/DecaLog/wikis/Reprises
[16] Master project CNRS/IN2P3 CompStat https://indico.in2p3.fr/event/16883/contributions/58826/attachments/46459/58070/tr180129_davidRousseau_IN2P3_IN2P3ML.pdf
[17] WLCG Strategy towards HL-LHC, CERN-LHCC-2018-019 ; LHCC-G-169; https://cds.cern.ch/record/2621698
[18] The HEP Software Foundation Community White Paper https://hepsoftwarefoundation.org/organization/cwp.html
[19] A Roadmap for HEP Software and Computing R&D for the 2020s, The HEP Software Foundation, Albrecht, J. et al., Comput Softw Big Sci (2019) 3, 7, https://doi.org/10.1007/s41781-018-0018-8
[20] ATLAS Collaboration, ATLAS Computing Technical Design Report, ATLAS-TDR-017 (2005), See also http://atlas-computing.web.cern.ch/atlas-computing/pack-ages/athenaCore/athenaCore.php
[21] Spark, https://spark.apache.org/
[22] Martin Abadi et al. “TensorFlow: A System for Large-Scale MachineLearning.” In:OSDI. Vol. 16. 2016, pp. 265–283
[23] PanDa system, T. Maeno et al. [ATLAS Collaboration], J. Phys. Conf. Ser.331(2011) 072024
[24] Rucio system, V. Garonne et al. [ATLAS Collaboration], J. Phys. Conf. Ser.898(2017) no.6, 062019
[25] CERN Open Data Portal, http://opendata.cern.ch/
[26] gitlab IN2P3, https://gitlab.in2p3.fr/
[27] IN2P3/IRFU computing days, https://indico.in2p3.fr/event/17206/
[27] IN2P3 School of Statistics, http://sos.in2p3.fr