Rucio beyond ATLAS: experiences from Belle II, CMS, DUNE, EISCAT3D, LIGO/VIRGO, SKA, XENON

Mario Lassnig[1],[*], *Martin* Barisits[1], *Paul J* Laycock[2], *Cédric* Serfon[2], *Eric W* Vaandering[3], *Katy* Ellis[4], *Robert A* Illingworth[3], *Vincent* Garonne[5], *John* White[6], *James A* Clark[7], *Gabriele* Fronze[8], *Rohini* Joshi[9], *Ian* Johnson[4], and *Boris* Bauermeister[10]

[1]European Organization for Nuclear Research
[2]Brookhaven National Laboratory
[3]Fermi National Accelerator Laboratory
[4]Science and Technology Facilities Council
[5]University of Oslo
[6]Nordic e-Infrastructure Collaboration
[7]Georgia Institute of Technology
[8]Istituto Nazionale di Fisica Nucleare
[9]Square Kilometre Array Organisation
[10]Stockholm University

Abstract. For many scientific projects, data management is an increasingly complicated challenge. The number of data-intensive instruments generating unprecedented volumes of data is growing and their accompanying workflows are becoming more complex. Their storage and computing resources are heterogeneous and are distributed at numerous geographical locations belonging to different administrative domains and organisations. These locations do not necessarily coincide with the places where data is produced nor where data is stored, analysed by researchers, or archived for safe long-term storage. To fulfil these needs, the data management system Rucio has been developed to allow the high-energy physics experiment ATLAS at LHC to manage its large volumes of data in an efficient and scalable way. But ATLAS is not alone, and several diverse scientific projects have started evaluating, adopting, and adapting the Rucio system for their own needs. As the Rucio community has grown, many improvements have been introduced, customisations have been added, and many bugs have been fixed. Additionally, new dataflows have been investigated and operational experiences have been documented. In this article we collect and compare the common successes, pitfalls, and oddities that arose in the evaluation efforts of multiple diverse experiments, and compare them with the ATLAS experience. This includes the high-energy physics experiments Belle II and CMS, the neutrino experiment DUNE, the scattering radar experiment EISCAT3D, the gravitational wave observatories LIGO and VIRGO, the SKA radio telescope, and the dark matter search experiment XENON.

[*]E-Mail: Mario.Lassnig@cern.ch

1 Motivation

The original motivation for a common data management solution comes from the anticipation of the limited amount of data resources available in the mid-term future. Shared use of global research infrastructures will become the norm, especially once sciences at the scale of the High Luminosity Large Hadron Collider (HL-LHC)[1], DUNE[2], and SKA[3] are online at the same time. Competing data operation requests on a very limited set of storage and network resources will cause under-utilisation and create bottlenecks, most prominently seen in data centres that host multiple experiments already. The computational part of the infrastructure is usually well-covered, e.g., via common scheduling, interfaces, and specifications. However, data was always missing a common open-source solution to tackle these shared challenges. Ensuring more efficient use of available data resources across multiple experiments has become a strategic goal for many communities, where they can allocate storage and network based on science needs and not based on administrative domains. Dataflow policies will need to be orchestrated across experiments to reduce or eliminate bottlenecks, and eventually be able to dynamically support compute workflows with adaptive data allocations. Additionally, providing unified monitoring, reporting and analytics to data centres and funding agencies, eases the administrative load for many communities. Finally, this also opens up the potential for shared operations across experiments, which will reduce the human effort and enables smaller communities to benefit from orchestrated dataflows.

Rucio[4] manages location-aware data in a heterogeneous distributed environment, including creation, location, transfer, deletion, and annotation. Declarative orchestration of dataflows with both low-level and high-level policies is its main mode of operation. Rucio was principally developed to address the data requirements of the ATLAS Experiment[5], is open-source software licensed under Apache v2.0, and makes use of established open-source toolchains. Functionality-wise, Rucio provides a mature and modular scientific data management federation, including seamless integration of scientific and commercial storage and their network systems. Data is stored in files and can contain any potential payload. Facilities can be distributed at multiple locations belonging to different administrative domains. It was designed with more than a decade of operational experience in large-scale data management.

Rucio provides many features that can be enabled selectively, starting from a file and dataset catalogue. Transfers can be managed between different types of physical media, such as disk and tape, and between different types of facilities, such as grids, clouds, and HPCs. A collection of user-interface tools such as the Web-UI, CLI, and an API to discover, download, upload, transfer, and annotate data are provided. Extensive monitoring for all resources and dataflows is included. Support for caches and CDN workflows is provided via expressive policy engines using rules and subscriptions. Corrupted data can be automatically identified and recovered. Popularity based replication can be enabled to support users who repeatedly need to access the same data. Finally, Rucio can easily be connected with different Workflow Management Systems (WMS) and already supports PanDA[6], the ATLAS WMS. Integration with more flavours of WMS is underway, most importantly DIRAC[7].

In an effort to foster collaboration throughout a growing and diverse community, the Rucio team hosts annual Rucio Community Workshops, semi-annual Rucio Coding Camps, and weekly development meetings.

2 ATLAS

The ATLAS experiment is one of two large general-purpose particle physics detectors built at the Large Hadron Collider at CERN. At the current scale for ATLAS, Rucio is managing more than 1 billion files, more than 500 petabytes of data, with data operations at a 400

Hz interaction rate for more than 1000 users. The data is spread across 120 data centres, which includes 5 HPC centres, and also connects commercial clouds. Per year, more than 500 petabytes of data are both transferred and deleted, while more than 2.5 exabytes/year are uploaded and downloaded for users and jobs.

For High Luminosity LHC, an increase of at least one order of magnitude for data volume is expected. Rucio is a central component to tackle the HL-LHC data deluge and will make use of even smarter orchestration features for the dataflow. To help with potential new dataflows, easy integration of new systems, ideas, and components is mandatory and supported by Rucio. Several combined effort research and development activities have been launched, i.e., distributed storage (Data Lakes), smart caching and access, fine-grained data delivery services for analysis, and even better commercial cloud integration. One of the highlight activities started in 2019 has been the Data Carousel, which promptly transfers and processes a sliding window of small inputs onto faster buffer storage, such that only a small percentage of input date are available at a time with the bulk data resident on offline storage. The data carousel mode of operations requires tight integration of workflow and dataflow systems for more efficient use of high-latency storage such as tape, which required the implementation of new algorithms on multi-site scheduling for both writing and reading, as well as smart placement of data based on estimated access patterns.

3 Belle II

The Belle II experiment[8] is a particle physics experiment designed to study the properties of B mesons. Belle II is the successor to the Belle experiment at the SuperKEKB accelerator complex at KEK in Tsukuba. Belle II has 981 members across 118 institutes in 26 countries. The data requirements include 200 PB of raw data expected by the end of data-taking in 2024, with 2 replicas distributed over 6 sites. Physics data taking started in 2019.

Belle II's current distributed data management uses a bespoke design with adequate performance and supports up to 150,000 transfers/day. Some scalability issues in the system were addressed, but others are inherent to the design of the system, most importantly the lack of automation: this means that data distribution and deletion are done by experts at a very fine granularity. The Belle II team at Brookhaven National Laboratory (BNL) are evaluating Rucio as an alternative and all studies so far look promising. Most importantly, the performance on the PostgreSQL database at BNL shows capabilities beyond the Belle II requirements.

Integration of Rucio with the rest of the Belle II distributed computing system, based on DIRAC, is planned in two stages. In the first stage the current data management APIs are extended with an implementation that uses Rucio under the hood. This is mostly transparent to the rest of Belle II and allows both data management backends to work in parallel during the transition phase. However, this still relies on a legacy file catalogue, and does not take full advantage of Rucio and its functionalities, being limited to the currently used APIs by definition. Nevertheless this stage allows the BNL team to gain experience in a production environment of using the DIRAC WMS with Rucio.

The second stage integration leads to an eventual migration that will use Rucio as the master file catalogue, using a new DIRAC plugin to remove the dependency on the legacy file catalogue. Since almost all components in the Belle II computing system have to interact with the master file catalogue, the DIRAC file catalogue plugin must hide Rucio requirements from DIRAC itself as well as the Belle II users. This means that the plugin takes care of much of the work done in the original APIs and leads to a potentially simpler system.

4 CMS

The Compact Muon Solenoid (CMS) experiment[9] is one of two large general-purpose particle physics detectors built on the Large Hadron Collider at CERN. CMS has an equivalent data volume and rate as ATLAS, with many hundreds of petabytes on disk and tape, and on the order of 100 storage systems to integrate. The production file size is in the gigabyte range, and user file sizes are in the hundred megabyte range. Per day, CMS transfers 2 petabytes across 1 million user and production files. The current data management is done by two layers of in-house products, PheDEx[10] and Dynamo[11]. Each site must host a PheDEx agent to manage its own data, including tape, and this requires non-trivial effort at each of the sites. The transfer component is ageing and likely will not scale to HL-LHC requirements. The Dynamo layer makes requests to dynamically distribute and clean up data, based on experiment plans and popularity. There is no user data management and user data transfers are done with a thin layer over FTS[12].

CMS performed an evaluation of data management systems from early 2018 through summer 2018, and eventually selected Rucio. The plan is to have Rucio deployed and ready for LHC Run 3: the transition period will last from 2018-2020, and the CMS team has expressed excitement about participating in a sustainable community project. The production infrastructure is based on Docker, Kubernetes, Helm, and OpenStack, with the official Rucio Helm charts customised with minimal configuration changes for CMS. The zero-to-operating cluster timing, including dependencies, is in the order of tens of minutes, which allows fast and easy integration with CMS software and infrastructure; Rucio upgrades are nearly instantaneous. This also allows CMS to have its production and Rucio testbed on a shared set of resources. The development environment is thus identical regardless which of the various flavors of central clusters are used.

In 2019, a test distributed 1 million files between all CMS T1 and T2 sites. The critical factor for data management scalability is the number of files, not the actual volume of data to be moved. The entire successful test took 1.5 days, and was purely driven by dataset injection rate; it ran in parallel to regular experiment activity.

5 DUNE

The Deep Underground Neutrino Experiment (DUNE) is a neutrino experiment under construction, with a near detector at Fermilab and a far detector at the Sanford Underground Research Facility (SURF) that will observe neutrinos produced at Fermilab.

DUNE's data management challenges are unique because they have multiple geographically separated detectors asynchronously collecting data, at an expected rate of tens of petabytes per year. Opportunely, DUNE is also sensitive to supernovae, which potentially produce hundreds of terabytes over a 100 second period. It is a large collaboration that intends to store and process data at many sites worldwide, and the current ProtoDUNE prototype detector already recorded 6 petabytes of reconstructed data. The next test beam run for both single and dual phase prototypes is expected in the 2021-22 timeframe.

DUNE has a Rucio instance at Fermilab with a PostgreSQL backend, and has contributed several database extensions to Rucio. So far, more than 1 million files have been catalogued from ProtoDUNE, including raw and reconstructed data. Rucio is being used to distribute ProtoDUNE data from CERN and FNAL to other sites for analysts. The replication rules make this easy: making a rule for a dataset and site or group of sites eliminates operational overhead for DUNE. The current integration plan is to progressively replace the legacy data management system, and transition to a purely Rucio based solution. The main challenge is that DUNE intends to make heavy use of HPC resources, and the data management system

needs to integrate with many very heterogeneous supercomputing sites. This is in line with the global HEP move towards using more HPC resources. Additionally DUNE data will benefit from fine grained object store style access, however it is not clear how to combine this with the traditional file based approach. The DUNE community has expressed interest in contributing to these developments in the near future.

6 EISCAT3D

EISCAT3D[13] will be a radar system for the scientific study of the Earth's atmosphere and ionosphere. It will use a technique called incoherent scatter radar (ISR) to measure basic physical parameters of the ionospheric plasma and upper atmosphere near the Earth. This kind of system supports the study of phenomena such as the aurora borealis (northern lights) and noctilucent clouds. Using separate stations in Norway, Sweden, and Finland, based on phased array technology, EISCAT3D will be able to make three-dimensional measurements of the plasma densities and temperatures and the direction of motion of that plasma, among other atmospheric measurements.

It is thus a data intensive instrument that generates a large volume of data, and its researchers need to analyse the data and share their results. The main question is how the data replication can be automated and if and how it can by synchronised with third-party systems, such as data management tools and catalogues.

EISCAT3D ran an automatic replication exercise with data uploaded to the experiment's storage system, which is then dynamically registered using a notification mechanism into Rucio. This new *Panoptes* service detects new files and registers them into Rucio for replication and data sharing using standard Rucio data flow policies.

7 LIGO/VIRGO

The Laser Interferometer Gravitational-Wave (LIGO) Observatory[14], based in the US, is a large-scale observatory to detect cosmic gravitational waves and to develop gravitational-wave observations as an astronomical tool. Virgo[15] is the European equivalent interferometer, based in Italy at the European Gravitational Observatory (EGO).

LIGO and Virgo are building the International Gravitational Wave Network (IGWN), with a combined 20 terabytes of astrophysical strain data and 1 petabyte of raw data (environmental, instrumental monitors) per instrument per observing year. Near real-time online analyses are data streamed with Apache Kafka[16] to dedicated computing resources. A data management solution is needed for offline deep searches and parameter estimation, as well as support for dedicated and opportunistic resources, as well as archival data.

The IGWN archival data distribution was done using the LIGO Data Replicator (LDR), the legacy data distribution system using MySQL and Globus. Rucio now enhances the IGWN data management through a large choice of protocols, an accessible catalogue, comprehensive monitoring and support for detector data flows. This includes domain-specific daemons that register new dataframes in the Rucio catalogue and then create rules to trivially implement dataflow to the archives and resources. IGWN has stated that they will investigate many opportunities beyond this, as well as being happy to update to a modern, high-availability version of existing functionality.

The deployment for the collaboration is primarily done through OSG[17] and IceCube[18] personnel. Rucio services are deployed on the Nautilus hypercluster[19] with Rucio webservers, daemons and the PostgreSQL database running in Kubernetes. IGWN is currently using CERN FTS, but is interested in hosting their own. Rucio is now being used in

production for limited frame data replication to volunteering sites, and a transition away from LDR is expected over the coming months. Upcoming work includes integration of existing data discovery services and remote data access, e.g., HTCondor file transfers[20], enhanced database redundancy, and management of new data products, e.g., analysis pipeline data products. A mountable Rucio POSIX namespace is under development as a potential CVMFS[21] alternative for gravitational wave software distribution.

8 Square Kilometre Array

The Square Kilometre Array (SKA) is an intergovernmental radio telescope project to be built in Australia and South Africa. With receiving stations extending out to a distance of more than 3,000 kilometres (1,900 mi) from a concentrated central core, and a very large field-of-view (FOV), it will allow astronomers to create and study extremely sensitive images of the universe.

The SKA Regional Centres (SRC) will provide a platform for transparent data access, data distribution, post-processing, archive storage, and software development. Up to 1 PB/day will be ingested from each telescope, and made available for access and post-processing around the globe. SKA will thus need a way to manage data in a federated way across many physical sites transparent to the user.

SKA has begun evaluating Rucio for SRC data management. Data has been uploaded, replicated, and deleted from storage systems using parameterised replication rules and sustained data transfers have already been demonstrated from South Africa to the United Kingdom. A full mesh functional test has been put in place and is demonstrating connectivity. Tests were conducted using data from the LOFAR telescope, an SKA pathfinder instrument. Currently, the Elasticsearch/Logstash/Kibana (ELK) monitoring stack[22] is being set up up, and already 8M data operation events from more than a year of testing have been ingested.

The evaluation experience using Rucio has been positive and is now formalised through the H2020 ESCAPE project, the European Science Cluster for Astronomy and Particle Physics ESFRI research infrastructures[23]. Rucio is the primary candidate for the data management and orchestration service in ESCAPE. The main findings from the test include the arduous need for X.509 certificates across storage systems, which is now being addressed via alternatives such as token-based authentication and authorisation. Also, an in-depth look at the ELK monitoring and dashboards will be performed to see if they are useful and where they need to be extended. Another major point is the integration with the DIRAC WMS system, matching the Belle II needs, for a full end-to-end use case. Another use case will be similar to LHC Tier-0 processing with event-driven data management and processing and will be tested. The inclusion of Australian storage for long-distance tests and a focus on network optimisation is also upcoming through the ESCAPE project.

9 XENON

The XENON dark matter experiment is operated in the underground research facility Laboratory National del Gran Sasso (LNGS) in Italy. It is aiming to directly detect weakly interacting massive particles (WIMPs). The experimental setup of XENON1T is a two-phase time projection chamber (TPC), which is filled with 3.2 tons of ultra pure liquid xenon. Nuclear interactions introduce an ionisation and scintillation signal, that is recorded with 248 photo-multiplier tubes [24].

The first stage are raw data, which are distributed with Rucio among grid computing facilities within the European Grid Infrastructure (EGI) and the Open Science Grid (OSG) in

the US. Connected resources within EGI are CNAF (Bologna), CCIN2P3 (Lyon), NIKHEF and SURFsara (Amsterdam) and the Weizmann Institute (Rehovot). The CI Connect infrastructure connects SDSC's Comet Supercomputer and the HPC campus resources to OSG. Second and third stage are processed data and minitrees which are kept at the Research Computing Centre (RCC) in Chicago. The RCC is also the main data analysis centre of the XENON collaboration. A Rucio independent tape copy is kept at the Paralleldatorcentrum (PDC) in Stockholm. XENON1T has taken more than 800 TB of raw data and ran multiple re-processing campaigns for improving data quality in ongoing data analysis tasks. [25]

The upcoming XENONnT upgrade will take 1 petabyte per year. Processing and Monte Carlo simulation campaigns are planned at the major infrastructures EGI an OSG. The new stream processor (STRAX) will generate multiple data products, based on intermediate steps in the processing logic such as hit and event finding, position reconstruction and further high level data products. The aDMIX tool integrates Rucio in the XENONnT data flow and data product locations are registered in the run database of XENONnT. All data products are distributed within Rucio to the connected grid computing facilities for storage. Tape storage will be integrated in Rucio this time and therefore dedicated grid locations are reserved to store the raw data product. XENONnT is the first hard Python3 dependency on Rucio.

STRAX processes the major data products at the LNGS already before distributing them with aDMIX and Rucio. Reprocessing of data products with STRAX are initiated with the job submission tool OUTSOURCE and distributed again with aDMIX. For analysts, the RCC in Chicago is the main data analysis centre and provides user access to high level data products at a near location with a Rucio location. Analysts can also define and produce their own data products for analysis purposes outside the run database or grid storage at any time. All analysis tools are shipped out in singularity containers with CVMFS to the connected computing facilities and the data analysts at RCC.

10 Lessons learnt

Distilling the experiences from development to deployment to support allows to distinguish between positive and negative feedback. The main positive feedback from the communities can be summarised into four topics.

1. A frequently voiced opinion was that Rucio was easy to integrate into existing infrastructure and software. The main reasons mentioned were the availability the Python clients, since most of the processing frameworks seem to be Python compatible, and the availability of the HTTP REST-based API, which allows to use the system from non-Python compatible systems.

2. The easy automation of a community's dataflow using rules and subscriptions results in large time savings for the operators. Where previously transfers had to be meticulously followed by humans, Rucio has automated this process reliably.

3. The trust in the automation stems from the comprehensive monitoring, which gives insights into the transfers, but also allows to easily discover yet unknown properties of a community's distributed system, such as redundant dataflows. These monitoring possibilities were particularly appreciated by the communities' computing managements.

4. Finally, the stand-out feature was that it is easy to contribute new code and extensions to the upstream system. With an open and inclusive software engineering process the Rucio code can be adapted and enhanced such that every community benefits.

On the other hand, there have been major challenges in the adoption processes of Rucio. The main negative feedback which had to be addressed can be summarised into three topics.

1. The installation procedure of Rucio is convoluted and requires expert knowledge. The reason for this was that Rucio's deployment was tailored for the peculiarities in the ATLAS experiment, including several scripts, database jobs, and customisations which were difficult to port to other environments. This has been addressed recently through full containerisation of Rucio and support for Kubernetes-based deployment.

2. The configuration of the system is complex and relies on too many ambiguous properties. Again, this stems from the needs of ATLAS, which requires fine-grained customisation of many of the parameters and options in Rucio. This has been addressed recently through a complete redesign of the configuration mechanism. The configuration will be improved and simplified even further, such that a functional Rucio system setup can work out of the box without major configuration.

3. The most voiced challenge was that the documentation is too dispersed and out-of-date for many of the features. This has been addressed through automatic generation of documentation where possible, such as the API documentation, and through a dedicated documentation rewrite effort via the Google Season of Docs programme. A new community knowledge base, similar to StackOverflow, will be established as well for public contribution to recipes and best practices.

The software engineering process has successfully become a community-driven development. Requirements, features, issues, and releases are always publicly discussed, e.g., in weekly meetings, on GitHub, or on Slack. The core team usually only provides guidance for architecture, design, and tests. Normally, 1-2 people from a particular community then take responsibility to develop the software extension and take charge of its ongoing maintenance. Communities are helping each other across experiments, which has become particularly effective across time zones due to US involvement. Recent improvements in automation and containerisation of the development environment also lowered the barrier of entry for newcomers, such that the core team then only takes care of the management and packaging of the releases. Examples of ongoing community-driven developments include alternative third-party copy, Data Carousel improvements, quality of service support, token-based authorisation and authentication with storage, or integration with software defined networks.

11 Summary and conclusions

In the last few years, several experiments and communities have started to evaluate Rucio, and some have already adopted it for production use. Experts from the AMS[26] and XENON collaborations have been the early adopters and thus also contributed to the improvements in the software development process. The adoption by CMS was a decisive moment for the project, and led to a complete overhaul of the software release strategy. In the meantime, strong US and UK participation for support, development, and deployment boosted the collaborative nature of the project, yielding successful integrations with existing software and computing infrastructures. The emerging strong cooperation between HEP and multiple other fields, notably neutrino and astronomy, have led to growing interest from more diverse range of sciences, which now stimulates community-driven innovations to enlarge functionality and address common needs. In conclusion, Rucio is a successful collaborative open-source project that is rapidly developing into a common standard for scientific data management.

References

[1] G. Apollinari, O. Brüning, T. Nakamoto and L. Rossi, "High Luminosity Large Hadron Collider HL-LHC," CERN Yellow Rep. (2015) no.5, 1 doi:10.5170/CERN-2015-005.1 [arXiv:1705.08830 [physics.acc-ph]].

[2] A. Tonazzo [DUNE Collaboration], "Status of DUNE," SciPost Phys. Proc. **1** (2019) 043. doi:10.21468/SciPostPhysProc.1.043

[3] A. Chrysostomou, R. Bolton, G. R. Davis, "The square kilometre array: challenges of distributed operations and big data rates," Proc. Observatory Operations: Strategies, Processes, and Systems VII; Volume 10704:1070419 (2018). doi:10.1117/12.2309554

[4] M. Barisits *et al.*, "Rucio - Scientific data management," Comput. Softw. Big Sci. **3** (2019) no.1, 11 doi:10.1007/s41781-019-0026-3

[5] ATLAS Collaboration, JINST 3 (2008) S08003

[6] P. Svirin *et al.*, "BigPanDA: PanDA Workload Management System and its Applications beyond ATLAS," EPJ Web Conf. **214** (2019) 03050. doi:10.1051/epjconf/201921403050

[7] F. Stagni *et al.* [DIRAC Collaboration], "DIRAC in Large Particle Physics Experiments," J. Phys. Conf. Ser. **898** (2017) no.9, 092020. doi:10.1088/1742-6596/898/9/092020

[8] T. Kuhr [Belle II Software Group], "Belle II at the Start of Data Taking," EPJ Web Conf. **214** (2019) 09004. doi:10.1051/epjconf/201921409004

[9] CMS Collaboration, JINST 3 (2008) S08004

[10] M. Giffels, Y. Guo, V. Kuznetsov, N. Magini and T. Wildish, "The CMS Data Management System," J. Phys. Conf. Ser. **513** (2014) 042052. doi:10.1088/1742-6596/513/4/042052

[11] Y. Iiyama, C. Paus and M. Goncharov, "Dynamo - The dynamic data management system for the distributed CMS computing system," Slides, Contribution ID:328, CHEP 2016 Conference, San Francisco, October 8-14, 2016

[12] A. A. Ayllon, M. Salichos, M. K. Simon and O. Keeble, "FTS3: New Data Movement Service For WLCG," J. Phys. Conf. Ser. **513** (2014) 032081. doi:10.1088/1742-6596/513/3/032081

[13] J. Kero *et al.*, "EISCAT 3D: the next generation international atmosphere and geospace research radar," Proc. First NEO and Debris Detection Conference, Darmstadt, Germany, 22-24 January 2019. https://hdl.handle.net/10037/15341

[14] J. Aasi *et al.*, "Advanced LIGO," Class. Quantum Grav. 32, 074001 (2015). doi:10.1088/0264-9381/32/7/074001

[15] F. Acernese *et al.*, "Advanced Virgo Status," J. Phys. Conf. Ser. **1342** (2020) no.1, 012010.

[16] Apache Kafka, https://kafka.apache.org/, 2020

[17] R. Pordes *et al.*, "The Open Science Grid," J. Phys. Conf. Ser. **78** (2007) 012057. doi:10.1088/1742-6596/78/1/012057

[18] D. Williams [IceCube Collaboration], "Status and prospects for the IceCube Neutrino Observatory," Nucl. Instrum. Meth. A **952** (2020) 161650. doi:10.1016/j.nima.2018.11.109

[19] Nautilus - Pacific Research Platform, http://pacificresearchplatform.org/nautilus/, 2020

[20] D. Thain, T .Tannenbaum and M .Livny, "Distributed Computing in Practice: The Condor Experience" Concurrency and Computation: Practice and Experience 17:2-4 (2005) 323

[21] P. Buncic, C. Aguado Sanchez, J. Blomer, L. Franco, A. Harutyunian, P. Mato and Y. Yao, "CernVM: A virtual software appliance for LHC applications," J. Phys. Conf. Ser. **219** (2010) 042003. doi:10.1088/1742-6596/219/4/042003

[22] O. Andreassen, C. Charrondière and A. De Dios Fuente, "Monitoring Mixed-Language Applications with Elastic Search, Logstash and Kibana (ELK)," doi:10.18429/JACoW-ICALEPCS2015-WEPGF041

[23] ESCAPE - European Science Cluster of Astronomy & Particle physics ESFRI research infrastructures, https://projectescape.eu/, 2020

[24] E. Aprile *et al.*, "Dark Matter Search Results from a One Ton-Year Exposure of XENON1T," Phys. Rev. Lett. **121** (2018) 111302 doi:10.1103/PhysRevLett.121.111302

[25] D. Ahlin *et al.*, "The XENON1T Data Distribution and Processing Scheme," EPJ Web Conf. **214** (2019) 03015 doi:10.1051/epjconf/201921403015

[26] Nicola Tomassetti, "AMS-02 in space: physics results, overview, and challenges," Nucl. Part. Phys. Proc. **265-266** (2015) 245-247 doi:10.1016/j.nuclphysbps.2015.06.063