Integration of Rucio in Belle II

Cédric Serfon[1,*], *Ruslan* Mashinistov[1], *John Steven* De Stefano Jr[1], *Michel* Hernández Villanueva[2], *Hironori* Ito[1], *Yuji* Kato[3], *Paul* Laycock[1], *Hideki* Miyake[4], and *Ikuo* Ueda[4]

[1]Brookhaven National Laboratory, Upton, NY, USA
[2]University of Mississippi, MS, USA
[3]KMI - Nagoya University, Nagoya, Japan
[4]High Energy Accelerator Research Organization (KEK), Japan

Abstract. The Belle II experiment, that started taking physics data in April 2019, will multiply the volume of data currently stored on its nearly 30 storage elements worldwide by one order of magnitude to reach about 340 PB of data (raw and Monte Carlo simulation data) by the end of operations. To tackle this massive increase and to manage the data even after the end of the data taking, it was decided to move the Distributed Data Management software from a homegrown piece of software to a widely used Data Management solution in HEP and beyond : Rucio. This contribution describes the work done to integrate Rucio with Belle II distributed computing infrastructure as well as the migration strategy that was successfully performed to ensure a smooth transition.

1 Introduction

The Belle II experiment [1] on the SuperKEKB [2] accelerator at the High Energy Accelerator Research Organization (KEK) (Tsukuba, Japan) is an experiment dedicated to B physics. Belle II uses a Distributed Computing infrastructure with about 30 sites worldwide. Until recently, Belle II has been using a homegrown piece of software for its Distributed Data Management (DDM), part of an extension of Dirac [3] called BelleDIRAC [4]. By late 2018, it was realized that this software required significant performance improvements to meet the requirements of physics data taking and was seriously lacking in automation. At that time, a Distributed Data Management solution called Rucio [5], initially developed by the ATLAS collaboration [6], started to gain popularity in the wider HEP community. In the evaluation exercise, Rucio was found to provide all the missing features, including automation and scalability, that were needed for Belle II. Therefore, it was decided to start working on the integration of Belle II software with Rucio. This paper describes all the work done to integrate Belle II software with Rucio. In section 2, the old DDM system is briefly introduced. Sections 3 and 4 respectively detail the new developments and tests that were performed. The final migration that happened in January 2021 was also a complex task and is described in section 5.

*e-mail: cedric.serfon@cern.ch

2 Generalities about Belle II DDM

The Data Management part of BelleDIRAC [7, 8] provides the tools to register, read, transfer and delete files. It is integrated with the other components of BelleDIRAC and in particular the Workload Management system as shown in Fig. 1. Before the migration to Rucio, it used an external catalog called the LCG File Catalog (LFC) [9] which stores the file replicas. This catalog was widely used in the early 2010s, in particular by the LHC experiments, but now all of them moved to other solutions like the DIRAC File Catalog [10] or Rucio. Rucio is not only a file catalog, but an advanced DDM system that provides not only the functionalities of the old Belle II DDM system but also many others like replication policies, smart space usage, recovery tools, etc. all demonstrated at scales well beyond Belle II's needs. For instance the maximum daily volume of transferred data in Belle II during the first year of data taking was about 50TB with 0.2M files at peak, whereas ATLAS runs Rucio in production with a daily throughput of up to 4M files or 2 PB.

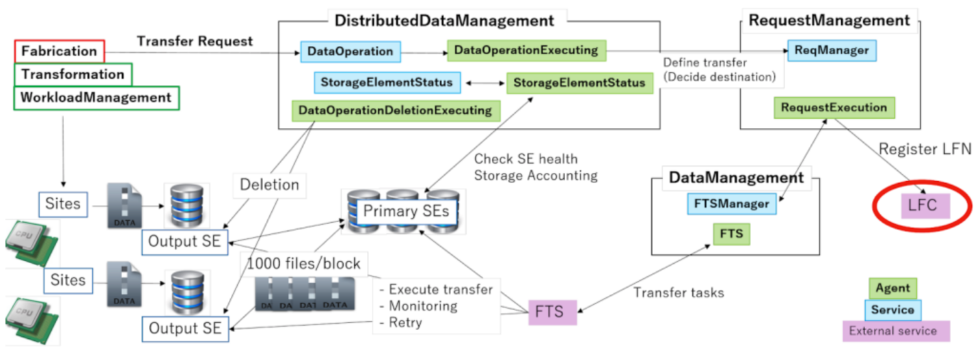

Figure 1. Schema of the DDM system before the transition to Rucio detailing its interactions with the Workload Management of BelleDirac and the external services (catalog, storage elements, File Transfer Service). Detailed description of the system can be found in [8].

3 Developments

3.1 Modification of the DDM API

BelleDIRAC DDM is based on a set of agents dedicated to the transfer of data (files or datablocks which are a collection of files) and on a remote procedure call (RPC) service that can be used by the applications or the end-users to query the status of the replications, shown in Fig. 1. With the migration to Rucio, the RPC Service was completely re-implemented in order to maintain the same APIs for the Belle II production and raw data management systems [11], while relying on the Rucio subscription mechanism to manage data. APIs used by the monitoring system were also adjusted to maintain the functionality of existing tools used by Data Production shifter as much as possible.

3.2 Rucio File Catalog plugin

As mentioned in section 2, before the migration to Rucio, DDM used LFC which is a hierarchical catalog that enables the organization of files into a directory structure. Each file in this structure has a Logical File Name (LFN). Each LFN can have a list of associated Physical File Names (PFN) corresponding to multiple copies, also known as replicas, of the same file across distributed storage. If an application or a user wants to locate a particular LFN, query must be made to the LFC to get the associated list of file replicas. To be able to use Rucio, a Rucio File Catalog (RFC) plugin was created in BelleDIRAC. More details about this plugin can be found in [12].

3.3 Monitoring

Rucio has no built-in monitoring for file transfers and deletion. Every collaboration that uses Rucio have developed their own monitoring. Fig. 2 shows for instance the monitoring infrastructure that is used by the ATLAS experiment and that is described in detail in [13]. The infrastructure relies on Apache Kafka [14] that collects the data feeds from Rucio and on a Apache Spark [15] cluster that does the aggregation and the enrichment of data. This whole infrastructure is heavy and does not suit the needs of a collaboration like Belle II. To overcome this, a simplified monitoring infrastructure (see Fig. 3) was developed for Belle II. This infrastructure relies on a new lightweight and horizontally scalable daemon called Hermes2. This daemon collects the different events produced by Rucio and stores them in its internal database, aggregates them and sends them into a list of different services that can be plugged into the daemon. The services currently supported are InfluxDB [16], ElasticSearch [17], ActiveMQ, and email.

For Belle II, two data sources are used : InfluxDB and ElasticSearch. They receive every event related to file transfers and deletions. These data sources are then used to build a Grafana [18] dashboard that allows the monitoring of all the transfers and deletion managed by Rucio. A snapshot of this dashboard can be seen on Fig. 4.

3.4 Chained subscriptions

Although Rucio has many of the requested features for Belle II, some workflows were not covered. One of them is the chained replication for RAW data from KEK to a disk endpoint at a RAW data centre (a set of sites dedicated to storing RAW data) and then from the disk area to the tape area of the same site. Another one is the export of calibration data, produced by the automated calibration system [19] to KEK disk endpoint then to its associated tape endpoint.

To achieve this, a new feature was added to Rucio subscriptions [20]. In Rucio a subscription is a tool that allows users to define the replication policy for future data. Each subscription has two parameters: the first one is a list of metadata that a Data IDentifier (DID), i.e. a file, dataset or container, must match and the second one is a list of independent replication rules. If a DID matches the list of metadata of the subscription, the rules corresponding to that subscription are created. The new feature, called a chained subscription, allows a condition to be applied between the rules created by the subscription, e.g. if the first rule is create on site A, then the second rule must be created on site B, as shown in Fig. 5.

4 Tests

4.1 Performance tests

In order to determine the size of the Rucio instance at BNL, performance tests were conducted. For these tests, a Rucio instance was setup using a dedicated database node and a

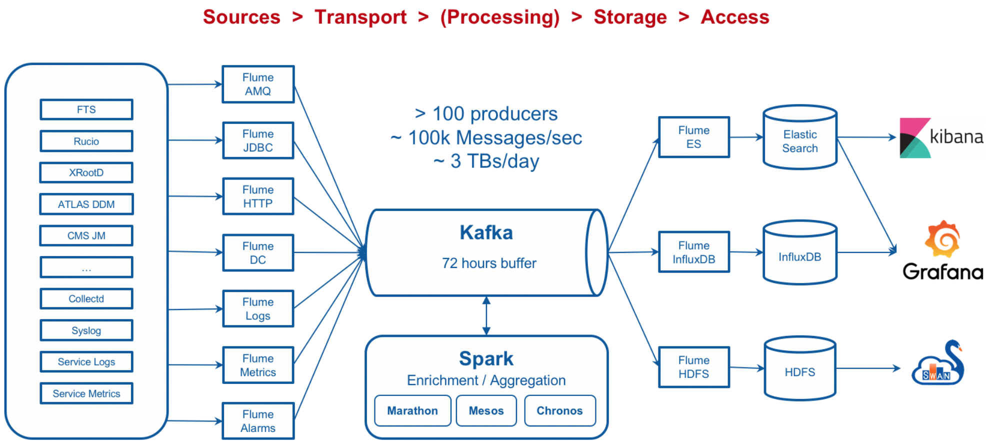

Figure 2. Monitoring infrastructure used for ATLAS. The whole infrastructure relies on a Kafka, a distributed event streaming platform, and on a Spark cluster that does the aggregation and enrichment of the data that is sent to different data sources.

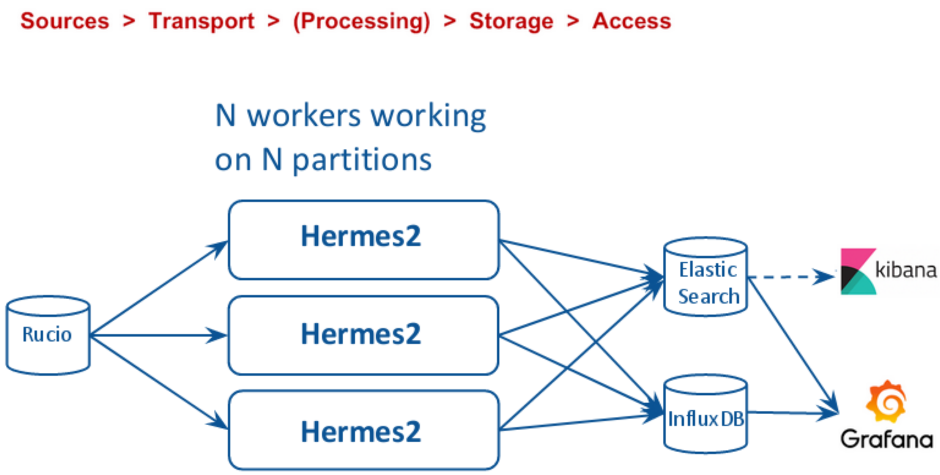

Figure 3. Monitoring infrastructure used for Belle II. The Hermes2 daemon collects Rucio messages and sends it to different services like ElasticSearch or InfluxDB that can be used as data sources for monitoring frontends. Multiple instances of the daemon can be started if needed, each instance running on a separate partition.

Rucio frontend. The instance was pre-populated with approximately 120 million files to simulate the number of files that will need to be managed. Following this initialisation procedure, insert, read and delete tests were performed to study the main database access patterns. The tests showed that with one frontend the insertion and read rates can reach 550 Hz, which is far beyond the expected rates required by Belle II. In addition, it showed that the botteneck was located on the frontends and not on the PostgreSQL backend.

Following these tests, it was decided to use two virtual machines to host the Rucio servers while the database host is a physical node with 200 GB of RAM running PostgreSQL. Two

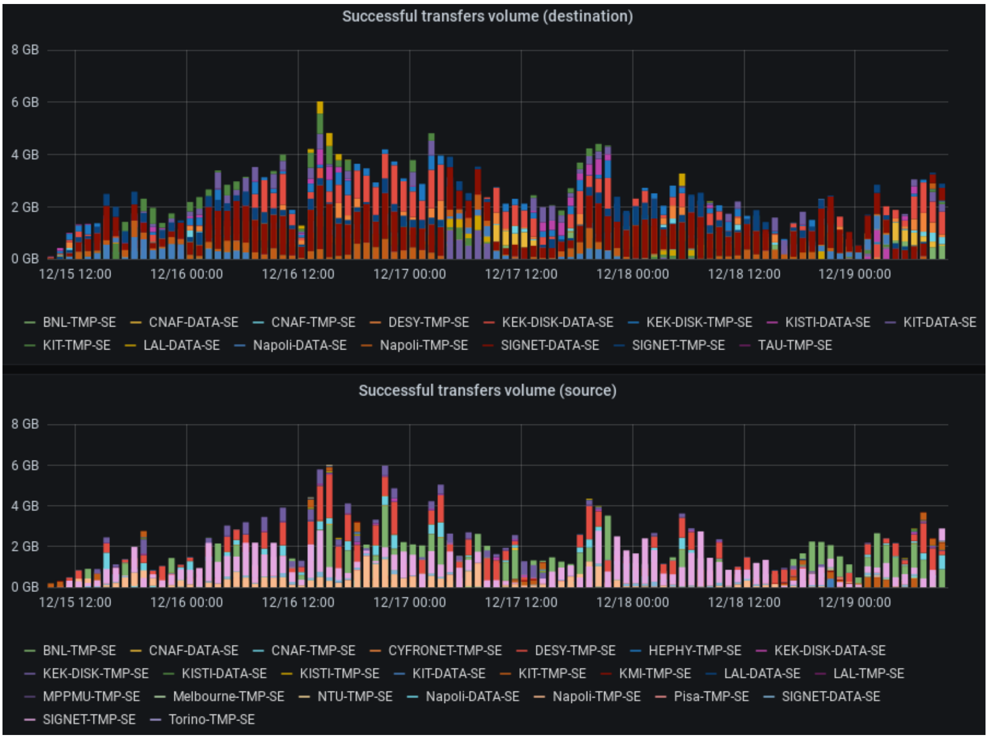

Figure 4. Snapshot of the dashboard monitoring for transfers and deletion. The top (resp. bottom) plot shows the volume of transfer to the destination (resp. source) versus time over a four days period with a one hour binning.

additional virtual machines to host the Rucio daemons complete the deployment configuration.

4.2 Functionality tests

After the initial implementation phase, the new DDM software components were developed and integrated into BelleDIRAC using the BelleDIRAC Fabrication system to check functionality, as this has the tightest coupling to the DDM. After the development phase, a six month certification period followed which was used to conduct performance and functionality checks of all of the major workflows which are:

- The export of RAW data from KEK to RAW data centres which is a critical part of Belle II computing. Using Rucio, this export is achieved using chained subscriptions. To test the workflow, a dedicated subscription was created. Datablocks were generated at KEK and shortly afterwards the subscriptions initiated the two step transfers as shown in Fig. 5.
- Monte Carlo production and distribution which relies heavily on DDM. The Fabrication system needs to get the location of the input data to broker the jobs and move data around. Each job needs to query Rucio for input data and to register new files. To test the whole workflow, several productions were launched and were successfully completed. To distribute data according to the defined policies, subscriptions were created. Different shares

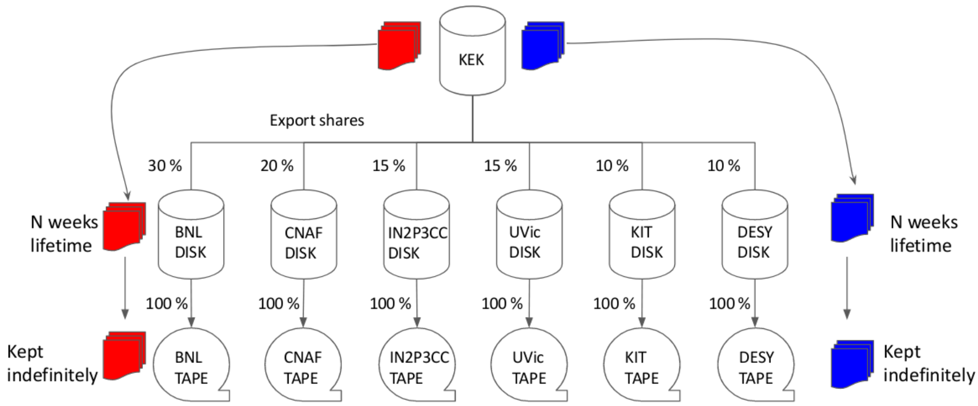

Figure 5. Schema explaining the concept of chained subscription. A new dataset is created and uploaded at KEK. If the dataset match the parameters of the subscription, Rucio will create a rule on one of the six RAW data centres according to the defined share, then it will create another rule on the associated tape endpoint.

are specified for the first steps of the production and the final step and the actual distribution is in good agreement with the shares as shown in Tables 1 and **??**.

- Finally, user analysis which is similar to the Fabrication system but has some significant differences, e.g. the account used by the users have not the same permissions as the production accounts. In order to have a realistic validation, real users were contacted and asked to run their analysis code on datablocks that were imported from LFC to Rucio specifically for this purpose.

Table 1. Distribution of datablocks produced during the certification tests for Monte Carlo production.

	First steps		Final step	
Site	Share expected	Actual number of datablocks	Share expected	Actual number of datablocks
BNL	14.3	157 (16.6%)	0	0 (00.0%)
CNAF	14.3	118 (12.5%)	11	7 (13.0%)
DESY	14.3	138 (14.6%)	0	0 (00.0%)
KEK	14.3	124 (13.2%)	22	16 (29.6%)
KIT	14.3	148 (15.7%)	12	6 (11.1%)
KMI	0.0	0 (00.0%)	5.5	0 (00.0%)
Napoli	14.3	119 (12.6%)	5.5	2 (03.7%)
SIGNET	14.3	138 (14.6%)	44	23 (42.6%)

5 Migration

5.1 Migration strategy

The migration to Rucio was a complex procedure that aimed to reach the final configuration shown in Fig. 6. Two migration strategies were evaluated:

- A two step migration: In the first step of this migration, the DDM is modified to delegate data movement to Rucio, while all other BelleDIRAC components continue to use the LFC for locating files. The second step is the migration from LFC to the Rucio File Catalog for all BelleDIRAC components. This strategy has the advantage that Rucio is used for transfers as soon as possible and before having the RFC plugin. However, the file replica information needs to be consistent in both Rucio and the LFC.
- In the second strategy considered, migration to Rucio only happens once all the components are ready. The disadvantage is that the lead time to using Rucio is longer, while the advantages include only having one migration.

It should be noted here that there was a strong desire to use Rucio as soon as possible and thus the first strategy was initially preferred. The two file catalog problem could be mitigated in the case of replication by using the DDM component itself to manage synchronisation. In the case of deletion, it was proposed to continue using the existing DDM implementation and ensure the LFC content (the only file catalog visible to other BelleDIRAC components) was correct and update Rucio asynchronously. However, it was eventually realised that, particularly in the case of deletion, it was really only a matter of time before the two file catalogs would be inconsistent, and the first strategy was eventually ruled out.

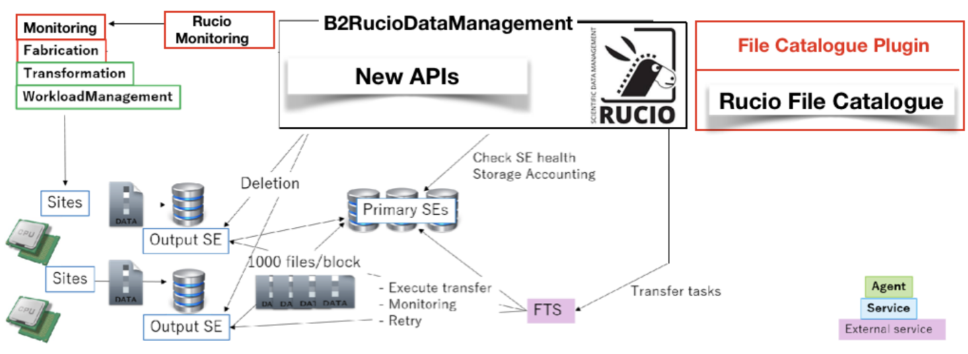

Figure 6. Schema of the DDM after the transition to Rucio detailing its interactions with the Workload Management of BelleDirac and the external services (storage element, File Transfer Service).

5.2 Migration tools and tests

To prepare the migration, a set of tools were created to import the content of the LFC into Rucio. The import procedure consists of different steps. In the first step a dump of the LFC at KEK was imported to Brookhaven National Laboratory (BNL) that hosts the Rucio server. This dump is then pre-processed to ease the insertion into Rucio. In the last step a set of scripts create all the files and their replicas, built the catalog hierarchy and finally created the rules. The scripts use multi-core concurrency to speed-up the import. Extensive tests were performed multiple times and showed that the whole LFC content could be imported in less than 24 hours.

5.3 Final migration

The final migration was scheduled between January 14th and January 18th 2021 (UTC) and necessitated a complete downtime of Belle II computing activities. These dates were chosen during the winter shutdown of the KEK accelerator in order not to disrupt the data taking and to reduce the effect on end-users, since the date overlaps a week-end. One of the major difficulties of this migration was that it involved people spread over four timezones: JST (UTC+9), CET (UTC+1), EST (UTC-5), CST (UTC-6), so good coordination was needed.

After a one day draining of the grid, all the Dirac services were switched off and the LFC hosted at KEK was set to read-only to prevent the addition of new files. Then the content of the LFC was dumped and exported to BNL where the Rucio instance is running. After this, the LFC dump was imported into the Rucio database using the tools mentioned previously. The whole import lasted about 24 hours as shown in Fig. 7. During this import a little more than 100 million file replicas were created and around 1 million replication rules were injected. No major issue was identified during this process thanks to the multiple tests described in previous subsection.

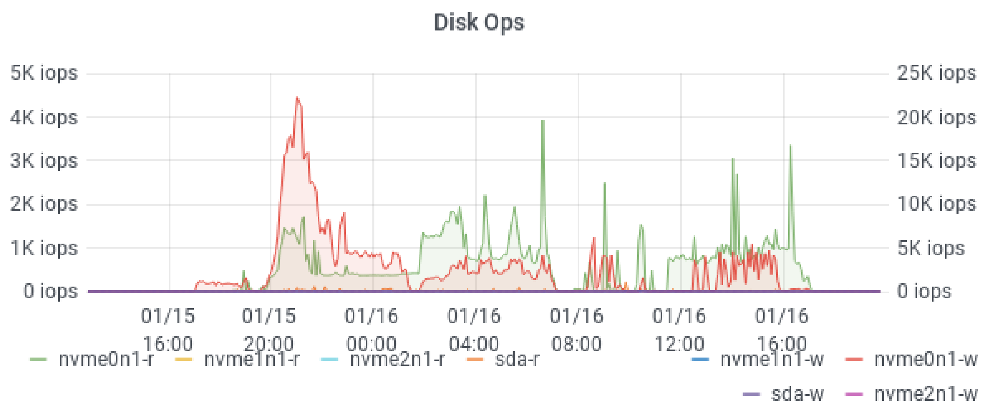

Figure 7. Number of Input/output operations per second on the PostgreSQL database used for Rucio during the import procedure of the LFC content.

After the whole LFC content had been imported to Rucio, the replication rules for the datablocks of active production were needed to be registered into the DDM service of BelleDIRAC so that when activity was resumed the Fabrication system was able to continue tracking its datablocks. Once the imports were done, and validated, the configuration of the BelleDIRAC production servers was changed to use Rucio instead of LFC, then user and production test jobs were sent. After the restart, a few small bugs that were not spotted during the certification process were identified and quickly fixed in the course of the day. The full restart was then postponed to January 19th, with one day delay with respect to the schedule.

During the next days, the system stayed under close monitoring from the Distributed Computing experts and a few minor bugs were identified and fixed, but none of them were critical. In the weeks following the transition, Belle II managed to achieve transfer rates similar to the ones from bigger collaborations like ATLAS (see Fig. 8).

6 Conclusion

The migration of Belle II to Rucio as Data Management software is a big achievement. It is the result of more than 2 years of work in evaluating, interfacing and testing the integration of Rucio with BelleDIRAC. The last step of this integration that consisted of importing the content of the old DDM into Rucio went smoothly for such a big change and was made

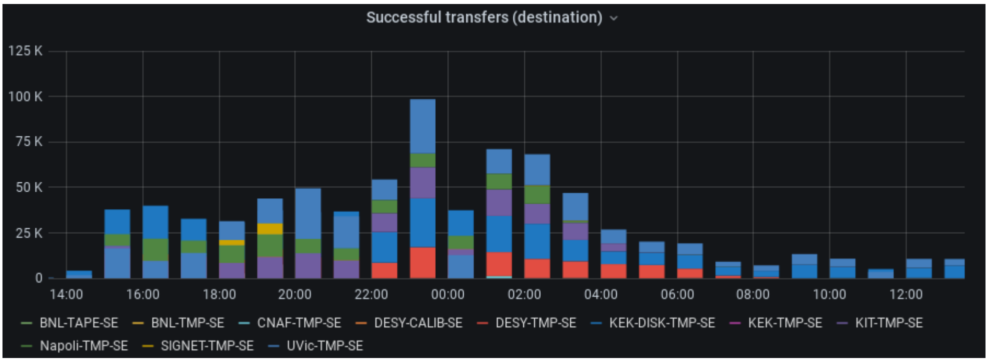

Figure 8. Number transfers over a 24 hours period on January 28th-29th. The number of files transferred over this period is of the same order than a normal day of transfers for ATLAS.

possible thanks to the large amount of preparatory work done beforehand. No critical issues have been reported since Rucio was put into production in mid-January 2021. Some of the new features provided by Rucio and that were not available in the old DDM are already being actively used by Distributed Computing experts and shifters.

Rucio will help to manage the big increase of data expected in the coming years by Belle II. We will be able to leverage the experience from the growing Rucio community and in return the developments performed for Belle II (e.g. the RFC plugin in Dirac) will benefit the wider community.

References

[1] T. Abe *et al.*, KEK-REPORT-2010-1, arXiv:1011.0352 (2010)

[2] K. Akai *et al.*, Nucl. Instrum. Meth. A **907**, 188-199 (2018)

[3] Federico Stagni, Andrei Tsaregorodtsev, André Sailer and Christophe Haen, "The DIRAC interware: current, upcoming and planned capabilities and technologies," EPJ Web Conf. **245** 03035 (2020). doi: 10.1051/epjconf/202024503035

[4] H. Miyake *et al.* [Belle-II computing group], "Belle II production system," J. Phys. Conf. Ser. **664**, no.5, 052028 (2015) doi:10.1088/1742-6596/664/5/052028

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

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

[7] Malachi Schram, "The data management of heterogeneous resources in Belle II," EPJ Web Conf. **214** 04031 (2019). doi:10.1051/epjconf/201921404031

[8] Siarhei Padolski, Hironori Ito, Paul Laycock, Ruslan Mashinistov, Hideki Miyake, Ikuo Ueda "Distributed data management on Belle II," EPJ Web Conf. **245** 04007 (2020). doi: 10.1051/epjconf/202024504007

[9] J.P. Baud, J. Casey, S. Lemaitre and C. Nicholson, "Performance analysis of a file catalog for the LHC computing grid", HPDC-14. Proceedings. 14th IEEE International Symposium on High Performance Distributed Computing, 2005., Research Triangle Park, NC, 2005, pp. 91-99, doi: 10.1109/HPDC.2005.1520941.

[10] A. Tsaregorodtsev *et al.* [DIRAC], "DIRAC file replica and metadata catalog", J. Phys. Conf. Ser. **396** (2012), 032108 doi:10.1088/1742-6596/396/3/032108

[11] Michel Hernández Villanueva and Ikuo Ueda, "The Belle II Raw Data Management System," EPJ Web Conf. **245** 04005 (2020). doi: 10.1051/epjconf/202024504005

[12] Cédric Serfon *et al.*, "The Rucio File Catalog in Dirac" CHEP 2021, these proceedings

[13] Thomas Beermann *et al.*, "Implementation of ATLAS Distributed Computing monitoring dashboards using InfluxDB and Grafana" EPJ Web Conf. **245** 03031 (2020). doi: 10.1051/epjconf/202024503031

[14] Apache Kafka: https://kafka.apache.org/ (accessed February 2021)

[15] Apache Spark: https://spark.apache.org/ (accessed February 2021)

[16] Influxdb: https://www.influxdata.com/ (accessed February 2021)

[17] Elasticsearch: https://www.elastic.co/elasticsearch (accessed February 2021)

[18] Grafana: https://grafana.com/ (accessed February 2021)

[19] F. Pham, D. Dossett and M. Sevior "Automated calibration at Belle II" CHEP 2021, these proceedings

[20] Martin Barisits *et al.*, "ATLAS Replica Management in Rucio: Replication Rules and Subscriptions" J. Phys.: Conf. Ser. **513** 042003 (2014) doi: 10.1088/1742-6596/513/4/042003