EPJ Web of Conferences **251**, 02029 (2021) https://doi.org/10.1051/epjconf/202125102029
CHEP 2021

The Cherenkov Telescope Array production system prototype for large-scale data processing and simulations

Luisa Arrabito[1,*], *Johan* Bregeon[2], *Patrick* Maeght[1], and *Michèle* Sanguillon[1] for the CTA Consortium
Andrei Tsaregorodtsev[3] for the DIRAC Consortium

[1]Laboratoire Univers et Particules, Université de Montpellier Place Eugène Bataillon - CC 72, CNRS/IN2P3, F-34095 Montpellier, France
[2]Laboratoire de Physique Subatomique et Cosmologie, Université Grenoble Alpes, CNRS/IN2P3, 38026 Grenoble, France
[3]Centre de Physique des Particules de Marseille, 163 Av de Luminy Case 902, CNRS/IN2P3, 13288 Marseille, France

Abstract. The Cherenkov Telescope Array (CTA) is the next-generation instrument in the very-high energy gamma ray astronomy domain. It will consist of tens of Cherenkov telescopes deployed in 2 arrays at La Palma (Spain) and Paranal (ESO, Chile) respectively. Currently under construction, CTA will start operations around 2023 for a duration of about 30 years. During operations CTA is expected to produce about 2 PB of raw data per year plus 5-20 PB of Monte Carlo data. The global data volume to be managed by the CTA archive, including all versions and copies, is of the order of 100 PB with a smooth growing profile. The associated processing needs are also very high, of the order of hundreds of millions of CPU HS06 hours per year. In order to optimize the instrument design and study its performances, during the preparatory phase (2010-2017) and the current construction phase, the CTA consortium has run massive Monte Carlo productions on the EGI grid infrastructure. In order to handle these productions and the future data processing, we have developed a production system based on the DIRAC framework. The current system is the result of several years of hardware infrastructure upgrades, software development and integration of different services like CVMFS and FTS. In this paper we present the current status of the CTA production system and its exploitation during the latest large-scale Monte Carlo campaigns.

1 Introduction

The CTA production system is in charge of handling the future data processing and Monte Carlo (MC) simulations of the CTA observatory [1]. The prototype presented in this paper is based on the DIRAC framework [2] [3] and has been used since several years to handle the massive Monte Carlo simulations for the CTA consortium on the EGI grid [4] [5]. The current system makes use of all main DIRAC components and is deployed on 5 core servers at two main data centers. On the top of the native DIRAC software, we have also developed an

*e-mail: arrabito@in2p3.fr

extension specific to CTA, which mainly consists of an API allowing users to easily configure and submit CTA workflows.

In addition, we have contributed to DIRAC core software by developing a new component to further automatize the workflow management. This component, called *Production System* in the DIRAC jargon (not to be confused with the CTA production system mentioned above), is described in detail in [6] and has been integrated in one of the major releases in 2020. Besides the DIRAC framework, the current prototype also relies on CVMFS (CernVM File System)[1] for the distribution of CTA software and on FTS (File Transfer Service)[2] for bulk data transfers.

In Section 2, we describe in more detail the server installation of the current production system. The computing model used to handle past and current CTA massive Monte Carlo productions on the EGI grid is presented in Section 3. Then, in Section 4 we will outline the main concepts of the workflow management with DIRAC and in particular its application to the use case of CTA. Finally we will present our conclusions on the developed prototype and our plans to update the system for the next-coming CTA operations.

2 CTA-DIRAC infrastructure

DIRAC software is designed with a Service Oriented Architecture (SOA). It is composed of a number of modules, called *Systems*, each one dedicated to a specific set of functionalities. The CTA prototype makes use of all the main DIRAC Systems. Among them we mention: the Workload Management Systems (WMS) for job scheduling, the Data Management System (DMS), the Request Management System (RMS), the Transformation and the Production Systems for the workflow management (TS and PS).

Each DIRAC system is composed of services, agents and databases, storing different kinds of information necessary to the functioning of the system. All services and agents are installed on 4 core servers, 2 hosted at CC-IN2P3 (Lyon) and 2 at PIC (Barcelona), while the DIRAC web portal is installed on a dedicated server at CC-IN2P3. At each hosting data center, databases are installed on a high-availability infrastructure, also providing a daily backup. At CC-IN2P3, a MariaDB Galera cluster composed of 3 nodes, also shared with other experiments, hosts the DMS, RMS, TS and PS databases. Among these, one of the most critical is the *File Catalog DB* which contains the references of all CTA files stored on the grid. At PIC, WMS and Accounting databases are hosted on a dedicated MariaDB server. Since databases keep the state of different systems, in case of failure of one or more core servers, the corresponding services can be easily restored.

More recently, we have also installed a new DIRAC component, the Monitoring System, which uses an Elasticsearch DB backend at CC-IN2P3 and which collects different types of information for diagnostic purposes. Currently we have two use cases for the Monitoring System. In the first case we collect various job parameters (CPU, RAM, disk usage, etc.) to perform some statistical analysis on jobs, while in the second case we collect some metrics about the server usage by the different DIRAC components (CPU load, number of threads, RAM, disk, etc.), useful to monitor the health of the server infrastructure.

Finally, for CTA software distribution we rely on the CVMFS system. The CTA repository is hosted on a Stratum-0 server at CC-IN2P3 and on 2 Stratum-1 servers, one at CC-IN2P3 and one at DESY Zeuthen.

[1] https://cernvm.cern.ch/fs/

[2] https://information-technology.web.cern.ch/services/file-transfer

3 Computing Model

The computing model described below has been adopted since 2012 to handle large-scale simulation campaigns for the CTA consortium on the grid. About 20 EGI sites currently support the CTA Virtual Organization and 7 of them also provide permanent storage (Storage Elements, SE) for a total of 5 PB of disk. Among these storage sites, 3 of them also provide tape storage for a total of about 2 PB. On average, the number of available cores is about 10000.

These resources are provided on a best-effort basis, so we cannot afford to keep several replicas of the data. As far as MC datasets are concerned, this is not a big issue for the moment. In the current model, we store all data on disk with a single replica distributed among 7 Storage Elements. Then, we regularly transfer old datasets from disk to tape, before finally removing them.

The computing model foreseen for the future CTA operations is a distributed model, but with a lower number of data centers, 4 in the current baseline, providing pledged resources and with an agreed service level. There will be several replicas and versions of the different datasets, whose number varies according to the stage of the processing. To reduce costs, data accessed less frequently will be regularly stored on tape.

As of today, we run MC simulation jobs at all 20 grid sites. The data produced by these jobs are uploaded on one of the 7 available disk SEs according to configurable shares. The destination SE is chosen randomly among the list of the available SEs with different probability weights assigned to each SE by the production manager. Then, MC analysis jobs processing the data produced by simulation jobs, are run at some selected sites having good network connectivity with the storage sites. Currently, to select these analysis sites, we don't rely on precise measurements of network bandwidth, but simply on the experience of some past productions where we observed systematic timeouts during the data access from specific remote sites. The job throttling rules to assign jobs to the different sites are easily adjusted during operations through DIRAC configuration. The production manager specifies which sites are eligible to run jobs that need to access a given SE. Different association rules can be defined for different job types, which are used to characterize jobs (e.g. analysis jobs using a particular software). This is done through the DIRAC web interface by configuring a plugin of the Transformation System ('ByJobTypeMapping' plugin). These job throttling rules concern the centrally managed large-scale productions, while for users jobs the default rule consists of assigning jobs to the site hosting the input data.

In order to facilitate the software deployment, we have started to run analysis jobs in singularity containers. Container images are in advance deployed on CVMFS so that analysis jobs can directly access them.

In parallel to these central production activities, CTA members can access grid resources through the CTA-DIRAC client and run their specific productions or analysis. For this purpose, the CTA-DIRAC extension provides a set of python scripts to configure and submit the most common CTA workflows.

4 Workflow Management

In order to execute Monte Carlo production and analysis workflows on a distributed infrastructure in an automated way, we have been relying on several DIRAC components. In this section we will explain in more detail the role of the highest level components that are in charge of orchestrating the workflows execution, i.e. the Transformation and the Production Systems. We will see in particular that one of the key aspects of these systems is to rely on data characterization by custom metadata (data-driven system). The support of custom

metadata in the DIRAC File Catalog (DFC) and their application to the use case of CTA is presented in the next section.

4.1 Data characterization

All data stored on the grid are referenced in the DIRAC File Catalog. The DFC has both Replica and Metadata Catalog functionalities. Files are registered in the DFC under a unique namespace together with their low-level metadata (size, checksum, etc.). As Metadata Catalog, it supports user-defined metadata, that are useful to characterize data for provenance or processing purpose. These metadata are defined in the DFC as key-value pairs on which users can perform queries to retrieve the corresponding list of files. As an example, to characterize CTA simulation datasets, we have defined a number of metadata keys, e.g.:

```
zenithAngle, azimuthAngle, primaryParticle, site, arrayLayout,
dataLevel, airshowerSimProg, airshowerSimProgVersion, telSimProg, etc.
```

Then, production jobs register output files in the DFC and update the corresponding metadata values. Users can use these metadata to query the DFC and retrieve the list of files corresponding to their selection. As we will see in Section 4.2, the same types of queries are also used by the Transformation System to select the files to be processed at different stages of a given workflow.

Another useful feature of the DFC is the support of *datasets* as aliases to metadata queries. Instead of writing full long queries, users can simply ask to retrieve the files associated to a given dataset, which has been previously defined. The content of a dataset is dynamically updated to reflect the actual content of the DFC. For each MC production, we usually define a number of datasets corresponding to the most common queries executed by users. Currently, more than 21 million of file replicas are registered in the DFC and MC data are grouped into 555 datasets. Datasets are thus an efficient and simple means to make large data products available to users.

4.2 Transformation and Production Systems

Workflow management functionalities are realized in DIRAC by two systems, the Transformation System (TS) and the Production System (PS). In this section, we will present the general concepts of these systems and how we use them to implement CTA workflows.

Workflows in DIRAC are seen as a set of data transformations handled by the TS. As illustrated in Figure 1, each transformation is composed of a number of *tasks* processing a given dataset. In the case of processing workflows, tasks are *jobs* that the TS submits to the WMS. Additionally, the TS can handle data management transformations and in this case tasks are *requests* submitted to the RMS.

A transformation is defined by a *task template* and a *data filter* which selects the dataset to process. The task template provides the general description of tasks in terms of the application to be executed and its parameters, CPU requirements, etc. Tasks defined by this template are all identical except for one varying parameter, which can be the input file for the task or a parameter of the application to be executed. For instance, Monte Carlo transformations do not have any input data, but tasks differ for an application parameter. Conversely, data processing transformations consist of all identical tasks, but processing different input files. Data filters are expressed as DFC queries on metadata or directly by *datasets*, as explained in Section 4.1.

On the top of the TS, the Production System is in charge of connecting several transformations together for the execution of whole workflows. Its main functionality is, starting from a

formal description of a given workflow, to create and monitor all the different transformations of which it is comprised. A production is defined as a series of production steps, where each production step is defined by the transformation description and the eventual specification of a connected transformation. We have already seen that transformations are defined by specifying a data filter to select the input data to be processed (Input Data Filter). Now, for a given transformation it's also possible to define data filters that characterize the output data produced by the corresponding tasks (Output Data Filter). In the PS, two transformations are connected together if part of the data matching the Output Data Filter of one transformation also match the Input Data Filter of the second transformation (cf. Figure 2). Starting from the workflow description, the PS verifies that the transformation definitions and their connections are valid and then creates the corresponding transformations.

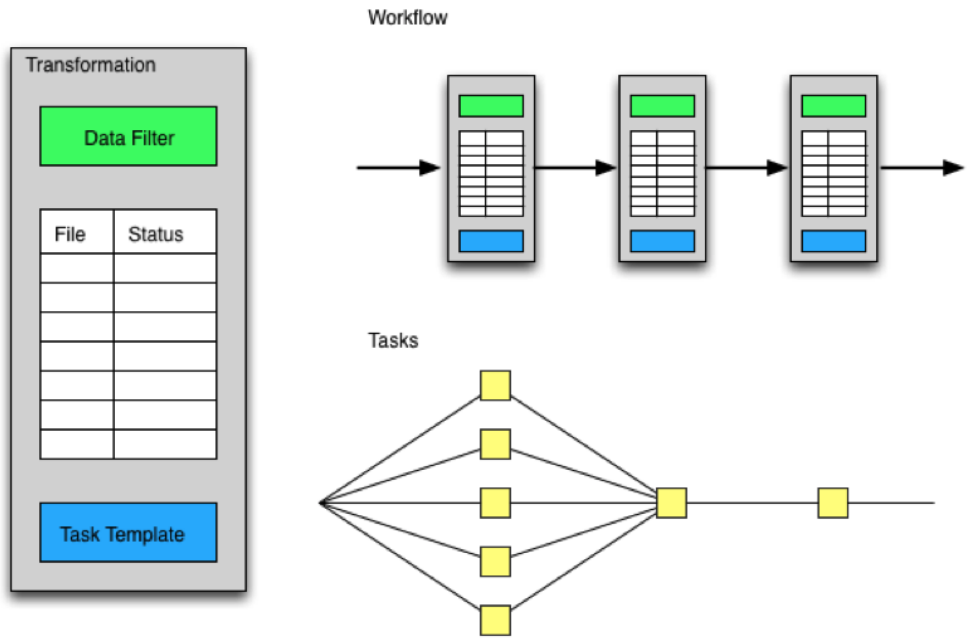

Figure 1. Workflows in DIRAC are series of data transformations. Each transformation is composed of several tasks and is defined by a task template and a data filter.

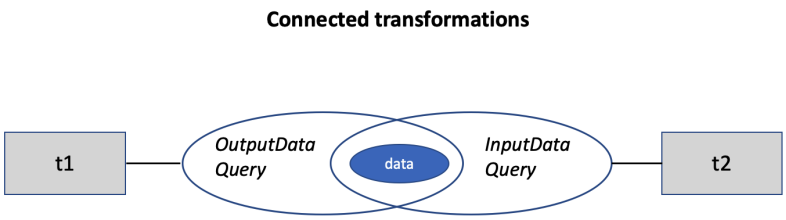

Figure 2. Example of transformations connected by data filters.

4.3 Application to CTA workflows

The first large scale application of the Production System was for the handling of the recent MC production ('prod5') that we have performed for the CTA consortium in 2020-2021. The whole prod5 workflow consists of several steps. The first steps correspond to MC production and are the most CPU intensive ones, while the last ones correspond to MC analysis, i.e. the processing of the data produced by the previous steps, and are less CPU demanding. Currently, MC production steps and the first stage of MC analysis are executed on the grid, while the final steps are executed on local clusters. Similarly to most HEP (High Energy Physics) workflows, prod5 workflow consists of a series of applications, where each application produces some data that are processed by another, as shown in the center of Figure 3. In Figure 3, we also show how we have implemented this workflow by means of the Transformation System. First of all we have defined a task template for the execution in sequence of the `CORSIKA` application [7] and of two instances of the `sim_telarray` application [8] for 2 different values of the Night Sky Background (NSB) parameter. With this template we have then defined a first transformation $t0$. Two files are thus produced by each task of $t0$, one for each NSB value. A second task template is used to describe tasks executing the `EventDisplay_stage1` application [9] . With this second template and with two data filters allowing respectively the selection of datasets produced for 2 different NSB values i.e., $DL0\_NSB1$ and $DL0\_NSB5$, we have then defined two processing transformations ($t1$ and $t2$). The whole workflow is thus composed of 3 transformations. It should be noticed however, that this workflow has to be executed for 4 different types of primary particles injected in the simulation and for 2 different pointing directions (azimuth angle = 0 and 180 degrees). This brings the total number of transformations to 24. Within the Production System, we have thus grouped 24 transformations into 8 productions, one for each configuration (particle type, pointing direction), as illustrated in Figure 4.

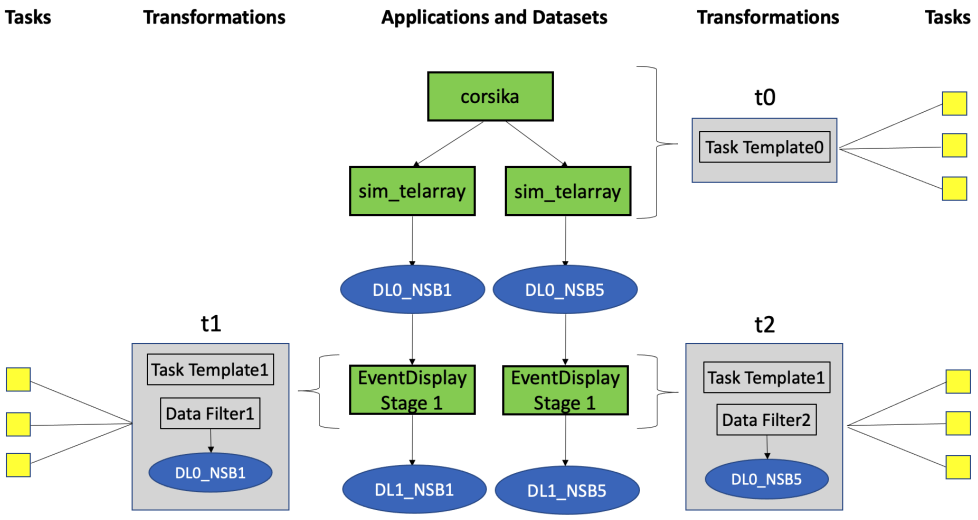

Figure 3. Example of workflow used during prod5 composed of 3 transformations.

Native DIRAC tools support the definition of transformations with data filters expressed as metadata queries. We found it convenient to develop some interfaces in the CTA-DIRAC extension to define transformations where data filters are specified directly by dataset names

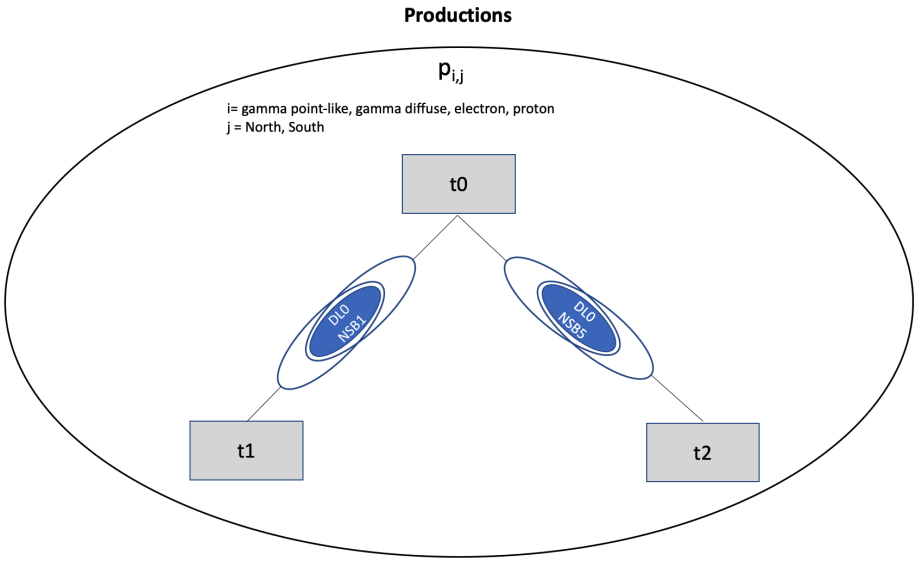

Figure 4. Example of a production executed during the prod5 simulation campaign. In total 8 productions were needed, one for each combination of incident particle and pointing direction.

(cf. Section 4.1). This feature largely simplifies the workflow description that is injected into the Production System.

Thanks to the Production System and the developed interfaces, we can easily define and monitor the progress of complex workflows composed of several processing steps, having different input parameters and resulting in more than 10 millions of jobs. In Figure 5, we show the number of concurrent running jobs since the start of prod5 in July in 2020. In particular, one can see that during the first phase of prod5 (July-August 2020), MC production and analysis jobs were running in parallel. This is due to the data-driven behavior of the Transformation System, so that as soon as MC data are produced, analysis jobs immediately start processing these data. Final results are thus available to users with minimal delay.

5 Data Management

In parallel to the production activity, we also regularly perform large-scale data management operations. During the last year, in order to free disk space we had first cleaned some old simulation datasets (about 530 TB) and then started transferring other datasets (about 1 PB) from disk to tape. The latter operation is still on-going and concerns old datasets that are very rarely accessed and that we plan to remove in the medium term. A single replica of these data is distributed among 7 Storage Elements. The goal is to move these data from the disk to tape by distributing them among the 3 tape Storage Elements available to CTA.

In order to efficiently transfer this large volume of data with minimum human intervention, we have again been relying on the TS. Indeed, the TS can also handle data management transformations using the RMS as backend. The RMS collects requests from various clients (jobs, users or DIRAC components). Requests of different type (client-service communications, removal, replication, etc.) are stored in a *Request DB* and then processed by a dedicated agent.

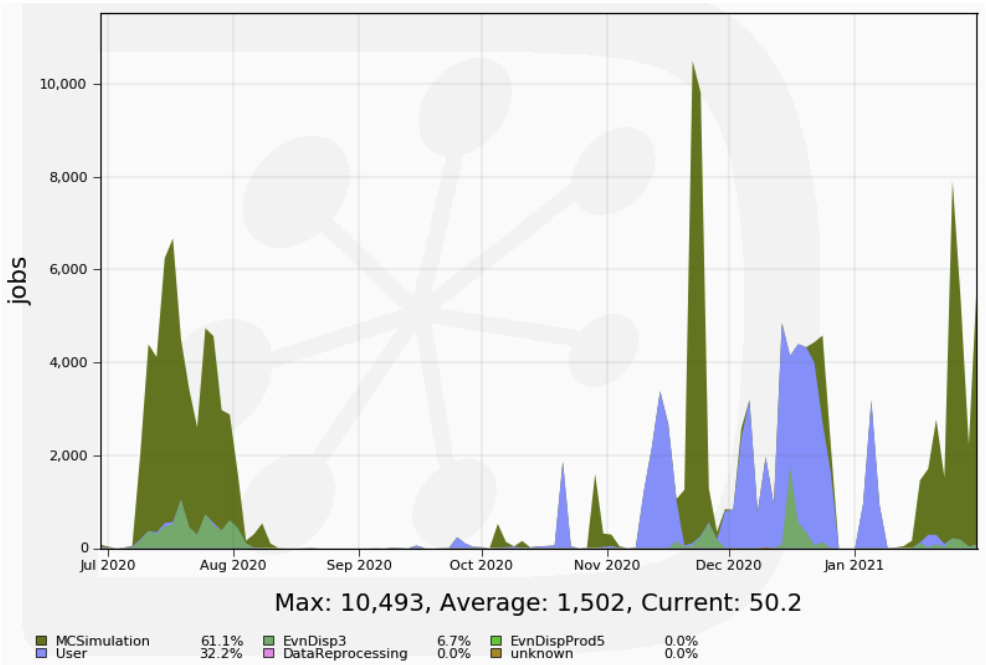

Figure 5. Running jobs from July 2020 to February 2021 grouped by job type: MC simulation jobs in dark green; analysis jobs in light green and users jobs in violet. During the first phase in July-August 2020, MC production and analysis jobs have been running in parallel through the Production System. The second phase of prod5 has been running late November 2020 and the third phase at the end of January 2021. Users' jobs also run in parallel to central production activities.

For the present use case of data transfer, we have created 'Moving' transformations, where the corresponding requests are composed of 2 operations: the first operation replicates the file from a source SE to a destination SE, while the second removes the file replica from the source SE. Exactly as for processing transformations, data management transformations are defined by a task template and a data filter. In this case the task template describes the type of request that must be executed (e.g. 'Moving' request), the source and the destination SE, while the data filter selects the files to be treated. In order to automatically handle and easily monitor the transfer of such a large data volume, we have thus selected about 50 *datasets* to transfer and for each one we have created a 'Moving' transformation.

6 Conclusions and perspectives

In order to handle large-scale data processing during the upcoming CTA operations, we have developed a production system prototype based on the DIRAC framework. It makes use of the main DIRAC components for workload and data management. This prototype has been successfully exploited since 2012 to manage all the main Monte Carlo campaigns of the CTA consortium.

In order to automatize as much as possible the execution of complex workflows, we have also contributed to the DIRAC core development with a new component (*Production System*). Moreover, we have developed a DIRAC software extension specific to CTA, which consists of a set of interfaces to facilitate the configuration and the submission of CTA workflows.

The current prototype also relies on other services, such as a CVMFS repository dedicated to CTA and the FTS instance at CERN.

From the beginning of the operations with this prototype, we have executed more than 18 million jobs corresponding to about 100 million of HS06 hours per year. These jobs have been transferring more than 4 PB per year between the different sites. Currently, about 4.4 PB of MC and users data are permanently stored on disk, distributed among 7 grid Storage Elements. We also regularly perform massive data management operations. In late 2020 we have started the migration of about 1 PB of old MC data from disk to 3 Storage Elements with tape library. All the files produced during these campaigns have been registered in the DIRAC File Catalog, which currently contains more than 21 million files. In order to characterize this large number of files, we have also defined in the DFC a number of metadata keys, whose values are automatically updated by jobs when they register their output files. In order to simplify the DFC queries, we have also grouped all production files into several *datasets* (about 550) sharing the same metadata. Finally, we have presented in more detail how we have defined and executed the CTA workflows of the latest MC campaign (prod5) using the DIRAC Production System.

The next step will consist of adapting the current prototype to the computing model that will be adopted for the future CTA operations. The computing model is planned to rely on 4 data centers equally sharing data storage and processing. Moreover, contrary to the current model, we plan to keep two versions of all the reduced data and one version of MC data. Then, for each version there will be 2 copies on tape and 1 copy on disk, with the exception of raw data for which, to reduce costs, only 10% will be kept on disk. It has to be noticed however that the exact number of versions and copies at each level of the data reduction might slightly change according to the effective data volume and processing time during future operations.

Thanks to the flexibility of the DIRAC framework, we will be able to easily adapt our current prototype to the new computing model and the underlying infrastructure. During the past years we have already successfully integrated new grid sites using different Computing Element types (e.g. CREAM, ARC, HT-Condor) as well as standalone clusters and cloud resources. In particular, cloud integration was tested with commercial and academic clouds in the context of the HNSciCloud project (see [6] for more details). The distribution of the processing workload can also be easily configured in our system and we regularly change it according to the available resources. Finally, thanks to the modular architecture of DIRAC, it is possible to use only its WMS component and interface it with an external data management system (or archive system in the CTA jargon). Some tests are currently on-going in this direction within the consortium to interface the DIRAC WMS with Rucio [10], as data management solution.

This work was conducted in the context of the CTA Consortium. We gratefully acknowledge financial support from the agencies and organizations listed here: http://www.cta-observatory.org/consortium_acknowledgments.

References

[1] *Science with the Cherenkov Telescope Array* (WORLD SCIENTIFIC, 2018), ISBN 9789813270091, http://dx.doi.org/10.1142/10986

[2] A. Casajus et al. (DIRAC), J. Phys. Conf. Ser. **396**, 032107 (2012)

[3] F. Stagni, A. Tsaregorodtsev, L. Arrabito, A. Sailer, T. Hara, X. Zhang (DIRAC), J. Phys. Conf. Ser. **898**, 092020 (2017)

[4] A. Acharyya et al., Astropart. Phys. **111**, 35 (2019), 1904.01426

[5] T. Hassan et al., Astropart. Phys. **93**, 76 (2017), 1705.01790

[6] L. Arrabito et al. (CTA Consortium, DIRAC Consortium), EPJ Web Conf. **214**, 03052 (2019)
[7] D. Heck, J. Knapp, J.N. Capdevielle, G. Schatz, T. Thouw (1998)
[8] K. Bernlöhr, Astropart. Phys. **30**, 149 (2008), 0808.2253
[9] G. Maier, J. Holder, PoS **ICRC2017**, 747 (2018), 1708.04048
[10] M. Barisits, T. Beermann, F. Berghaus, B. Bockelman, J. Bogado, D. Cameron, D. Christidis, D. Ciangottini, G. Dimitrov, M. Elsing et al., Computing and Software for Big Science **3**, 11 (2019)