The Cherenkov Telescope Array production system for Monte Carlo simulations and analysis

The Cherenkov Telescope Array production system for Monte Carlo simulations and analysis

L Arrabito[1], K Bernloehr[2], J Bregeon[1], P Cumani[3], T Hassan[3], A Haupt[4], G Maier[4], A Moralejo[3], N Neyroud[5] for the CTA Consortium
F Stagni[6], A Tsaregorodtsev[7] for the DIRAC Consortium

[1]Laboratoire Univers et Particules, Université de Montpellier Place Eugène Bataillon - CC 72, CNRS/IN2P3, F-34095 Montpellier, France
[2]Max-Planck-Institut für Kernphysik, P.O. Box 103980, D-69029 Heidelberg, Germany
[3]Institut de Fisica d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology, Campus UAB, 08193 Bellaterra Barcelona, Spain
[4]Deutsches Elektronen-Synchrotron, Platanenallee 6, 15738 Zeuthen, Germany
[5]Laboratoire d'Annecy-le-Vieux de Physique des Particules, Université de Savoie, CNRS/IN2P3, F-74941 Annecy-le-Vieux, France
[6]PH Department, CH-1211 Geneva 23 Switzerland
[7]Centre de Physique des Particules de Marseille, 163 Av de Luminy Case 902, CNRS/IN2P3, 13288 Marseille, France

E-mail: `arrabito@in2p3.fr`

Abstract. The Cherenkov Telescope Array (CTA), an array of many tens of Imaging Atmospheric Cherenkov Telescopes deployed on an unprecedented scale, is the next-generation instrument in the field of very high energy gamma-ray astronomy. An average data stream of about 0.9 GB/s for about 1300 hours of observation per year is expected, therefore resulting in 4 PB of raw data per year and a total of 27 PB/year, including archive and data processing. The start of CTA operation is foreseen in 2018 and it will last about 30 years. The installation of the first telescopes in the two selected locations (Paranal, Chile and La Palma, Spain) will start in 2017. In order to select the best site candidate to host CTA telescopes (in the Northern and in the Southern hemispheres), massive Monte Carlo simulations have been performed since 2012. Once the two sites have been selected, we have started new Monte Carlo simulations to determine the optimal array layout with respect to the obtained sensitivity. Taking into account that CTA may be finally composed of 7 different telescope types coming in 3 different sizes, many different combinations of telescope position and multiplicity as a function of the telescope type have been proposed. This last Monte Carlo campaign represented a huge computational effort, since several hundreds of telescope positions have been simulated, while for future instrument response function simulations, only the operating telescopes will be considered. In particular, during the last 18 months, about 2 PB of Monte Carlo data have been produced and processed with different analysis chains, with a corresponding overall CPU consumption of about 125 M HS06 hours. In these proceedings, we describe the employed computing model, based on the use of grid resources, as well as the production system setup, which relies on the DIRAC interware. Finally, we present the envisaged evolutions of the CTA production system for the off-line data processing during CTA operations and the instrument response function simulations.

1. Introduction

The CTA [1] computing model is based on the use of a distributed computing infrastructure, where a few first-class computing centers will share the global storage and computing loads. The current baseline foresees 4 computing centers, with one of them playing the role of 'Science Data Management Center' (SDMC), in charge of the orchestration of the activities of all the contributing computing centers. In particular, the SDMC will run both the Computing Resource Management (CRMS) and the Archive systems. The CRMS is responsible for the *workload management*, i.e. the management of the whole job life cycle, starting from the job submission by the user, passing through the job brokering, the execution on the computing infrastructure and ending with the result retrieval. At a higher level, the CRMS also manages the different steps of the data processing (*workflow management*), where each step is composed of a large number of jobs, processing a given dataset. Finally, the Archive system is in charge of all data management operations, i.e. data ingestion, retrieval, transfer and replication, also providing catalog functionalities.

A production setup prototype, based on the DIRAC framework [2][3], has been deployed in 2011 to perform CTA Monte Carlo (MC) studies on the EGI (European Grid Infrastructure) grid. Indeed, during the past CTA preparatory phase (2010-2014) and the current pre-construction phase, massive Monte Carlo (MC) simulations have been required to guide various choices in the CTA design. The first big MC campaign, called 'prod2', was run in 2012-2013, with the goal of selecting the best site candidate in each hemisphere, to host the CTA telescopes. Once the two sites have been selected, i.e. Paranal, Chile and La Palma, Spain, a new MC campaign, called 'prod3', was started to determine the optimal array layout at each site. For the South site at Paranal, 892 telescope positions were considered, as well as different types of telescopes and cameras, resulting in 3092 simulated telescopes. In order to make sure that the conclusions were robust, two fully independent analysis chains, Eventdisplay [4] and MAGIC Reconstruction Software (MARS) [5] (in use respectively within the VERITAS [6] and MAGIC [7] collaborations), have been used to process the full prod3 dataset for a large number of telescope configurations.

In these proceedings, we describe the production setup, called 'CTA-DIRAC', used to handle these MC simulations, with particular reference to the computing and operational challenges represented by the prod3. The current production setup covers both the CRMS and archive functionalities, but in this paper we will focus only on the two aspects of the CRMS, i.e. the workload and workflow management.

The paper is organized as follows. In section 2.1, we describe the computing model adopted for the prod3 campaign and which is based on the use of grid resources. In section 2.2, we give some details about the physical implementation of the CTA-DIRAC production setup. In section 3, we focus on the workflow management of the prod3. Finally in section 4, we discuss the computing challenges of the prod3 and the results obtained with our production setup.

2. Production setup

2.1. Computing model

Since in the current pre-construction phase, CTA has not yet dedicated resources at 4 computing centers, as planned in the computing model, we have been relying on the EGI grid resources to perform the MC studies. The grid resources available to the CTA Virtual Organization consist of about 20 sites spread in 7 countries. Even if in the final model we will have a reduced number of computing centers, which won't be necessarily grid based, the use of the grid infrastructure allowed us to validate our production setup in a distributed environment.

The computing model used to handle these MC activities is very simple. From the computing point of view, a typical MC workflow consists of two main steps: 1) shower generation and simulation of the telescope array response, i.e. 'production' jobs; 2) reconstruction and analysis, i.e. 'analysis' jobs. Production jobs run at all available sites, while their output data are stored on the fly only at 6 sites, with a share proportional to their storage capacity. Production data thus form a single distributed replica. In order to minimize data transfers, analysis jobs were originally run exclusively at the sites hosting their input data. However, with this approach the available resources were limited and the analysis progress was too slow. In order to speed up the analysis, we then allowed analysis jobs to run also at sites having a good network connectivity with the sites hosting the data. For the typical throughput of the prod3 (see Fig. 3), we haven't observed any increase in failures due to these data transfers. Moreover, since the memory requirements of the two analysis softwares (Eventdisplay and MARS) are quite different, we have also assigned a specific pool of resources to each type of analysis jobs. The produced analysis data are again stored at the same 6 sites used for the production. Finally, these high level data are retrieved by end users to perform the final steps of the analysis on their local resources.

2.2. CTA-DIRAC setup

The developed production setup fully relies on the DIRAC interware, which is designed to handle the computing activities in a distributed infrastructure. In particular, the DIRAC Workload Management System (WMS) and the Transformation System (TS), cover the functionalities required by the CTA CRMS. The TS is in charge of creating the jobs associated to a given high level workflow description (see section 3). Once the jobs are created, these are submitted to the WMS, which in turn submits them to the computing infrastructure. In order to store the production data on the grid storage elements, we have been using the DIRAC Data Management System. All production files have been registered in the DIRAC File Catalog, together with some useful meta-data information. As an example, in order to characterize the datasets of the prod3, we have defined about 10 meta-data, covering both physical information (simulation conditions) and provenance information (*e.g.* software program, release, etc.).

The choice of DIRAC was driven by several considerations, related to its advanced functionalities, modularity, flexibility and scalability. In particular, the DIRAC WMS implements the pilot mechanism [8], which allows to optimize the resource usage. Moreover, the DIRAC capability to target different types of computing resources (grids, clouds, batch systems, etc.), is of particular importance for a long life project like CTA.

The DIRAC architecture is based on a Service Oriented paradigm, where each System is composed of services, agents and databases. The current CTA-DIRAC server setup is the result of a few hardware upgrades done during the last 4 years. Today, it is composed of 5 core servers, hosted at CC-IN2P3[1], PIC[2] and DESY[3] computing centers, with 16 to 32 cores and 8 to 32 GB of RAM. Each server runs one or more DIRAC systems (services and agents), while the databases are hosted on two dedicated servers (at CC-IN2P3 and PIC). Finally, an additional server runs the DIRAC web portal. In addition to this hardware setup, we have also developed a minimal 'CTADIRAC' software extension, which essentially includes interfaces to easily configure and submit CTA productions.

2.3. Software deployment

For the software deployment of CTA applications at grid sites, we have setup a CVMFS (CernVM File System) [9] installation, composed of a stratum-0 server (at CC-IN2P3) and two stratum-1 servers (at CC-IN2P3 and DESY). Since most of the sites were already configured to use CVMFS, it has been straightforward to enable them to access to the CTA CVMFS repository.

[1] `http://cc.in2p3.fr/`

[2] `http://www.pic.es/`

[3] `http://gridcenter.desy.de/`

For the few remaining sites, not supporting CVMFS or not having CVMFS configured for CTA, the software releases are installed at jobs startup.

3. Workflow management

For the management of the multiple steps composing the prod3 campaign, i.e. production and analysis, we have been fully relying on the DIRAC TS. From the computing point of view, the production part is relatively simple, since it doesn't require any input file and it consists of two applications that can be executed sequentially within the same job. On the other hand, the analysis workflow is more complex, since it's composed of several steps running different applications and requiring input files.

In order to illustrate how we have been using the TS to manage the analysis workflow, we briefly describe here one of the two analysis chains run during the prod3, i.e. the MARS workflow. Even if this workflow is not the final one that will be used for the CTA data processing, the main functional blocks will very likely remain unchanged. The Fig. 1 shows the main steps composing the MARS workflow in terms of software applications (*e.g.* chimp, stereo, flux, sensitivity) and input/output datasets. Multivariate event classification algorithms (*e.g.* random forest [10]) are used for the energy estimation and background suppression, so that some steps are dedicated to the training while others to the analysis. The initial input dataset is split into several samples according to the primary particle type (*e.g.* proton, gamma, gamma-diffuse, electron). Each sample is then subdivided into subsamples to be used either for the training or for the analysis.

The building unit of the TS is the 'transformation', which is a collection of many identical jobs with a varying parameter. A transformation is essentially defined by a sequence of applications that must be executed within the same job (*transformation body*) and by a varying parameter. Moreover, several plugins can be defined to group input files and to select destination sites according to different criteria. The simplest example of transformations are those associated to MC production jobs. Indeed, in this case, jobs don't require any input file and they simply differ by a 'run number'. On the other hand, analysis jobs consist in executing a given sequence of applications on an input dataset, which is typically defined by a meta-data query condition (provided files are registered in a meta-data catalog). In this case the varying parameter is represented by the input file. In order to execute the whole MARS workflow, about 10 transformations are required. As an example, the fourth block in the Fig. 1 requires 3 transformations, running *chimp* and *stereo* on 3 different datasets (80 % of *proton*, 80 % of *gamma* and the full *electron* dataset), while the fifth block requires a transformation running the *flux* application on the datasets resulting from the previous step.

The automatization of the workflow execution is obtained in the TS, by implementing a data-driven paradigm. Indeed, once a transformation is defined with an associated meta-data query condition, as soon as new matching files are registered in the catalog (for instance produced by another transformation), they are automatically attached to the transformation. The corresponding jobs are then created to process those files. Thanks to this data-driven model, it's possible to automatize the execution of a workflow composed of several applications, each one taking as input the output of another application. The production manager should simply create all the necessary transformations describing a given workflow and let the system manage the whole process.

In the case of the MARS workflow, we were not able to apply this procedure, but we rather created and monitored the necessary transformations step by step. One of the main reasons is that some steps still require a human intervention, as for the validation of intermediate results or for the eventual tuning of the applications parameters. Moreover, some steps are triggered, not only by the presence of the input data files, but also by the presence of some auxiliary files (*e.g.* produced by the training steps) to be used as common inputs by all the jobs of a given transformation. In section 5, we will discuss our plans to improve the current procedure and make it fully automatized.

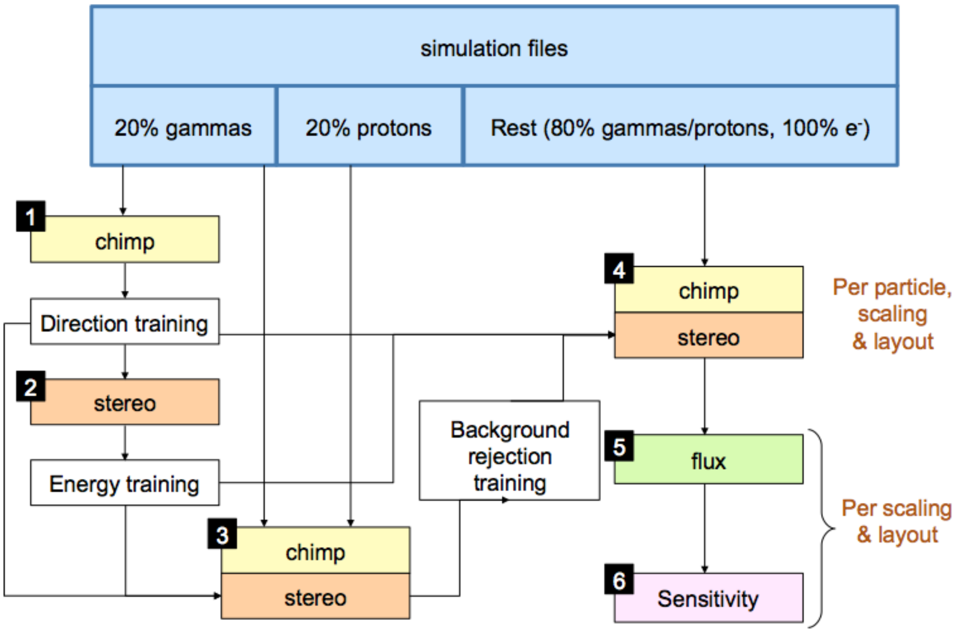

Figure 1. MARS analysis workflow. The initial simulation dataset is subdivided into several subsamples, that are used as input for the different steps of the workflow. Each colored box represents a step of the workflow, running one or two applications sequentially, on a given dataset. White boxes represent the output of the training steps.

4. Computing challenges

The CTA-DIRAC production setup, described in section 2.2, has been successfully exploited to run several massive MC simulations during the past 4 years. In order to illustrate the typical computing challenges of CTA MC simulations, we will consider the example of the most recent prod3. The campaign started in July 2015 and, even if the bulk production has been completed, some analyses are still in progress. As shown in Fig. 2, it consisted in several thousands of running jobs over periods of 4-5 months, reaching peak values of 12000 concurrent jobs. The overall CPU consumption is of 210 M HS06 hours with a total volume of 1.9 PB produced data. The typical throughput obtained during the prod3 is also shown in Fig. 3, as a function of the transfer status, corresponding to more than 10 PB globally transferred (see Fig. 4). Finally, there are currently about 7 M of files registered in the DIRAC File Catalog issued by the prod3, over a total number of 25 M of files.

The major problems encountered during this production have not been related to the CTA-DIRAC production setup, but to the high memory requirements of some specific analysis jobs (*e.g.* more than 8 GB) as well as to the large output to be written on the local disk (*e.g.* more than 50 GB). This problem was overcome by creating different types of transformations, according to the memory requirements of the associated application, and mapping each type to a given pool of resources.

5. Evolutions of the production setup

The overall behavior of the CTA-DIRAC production setup has been satisfying for the purpose of the prod3 and beyond, both from the functional and the performance point of view. Concerning

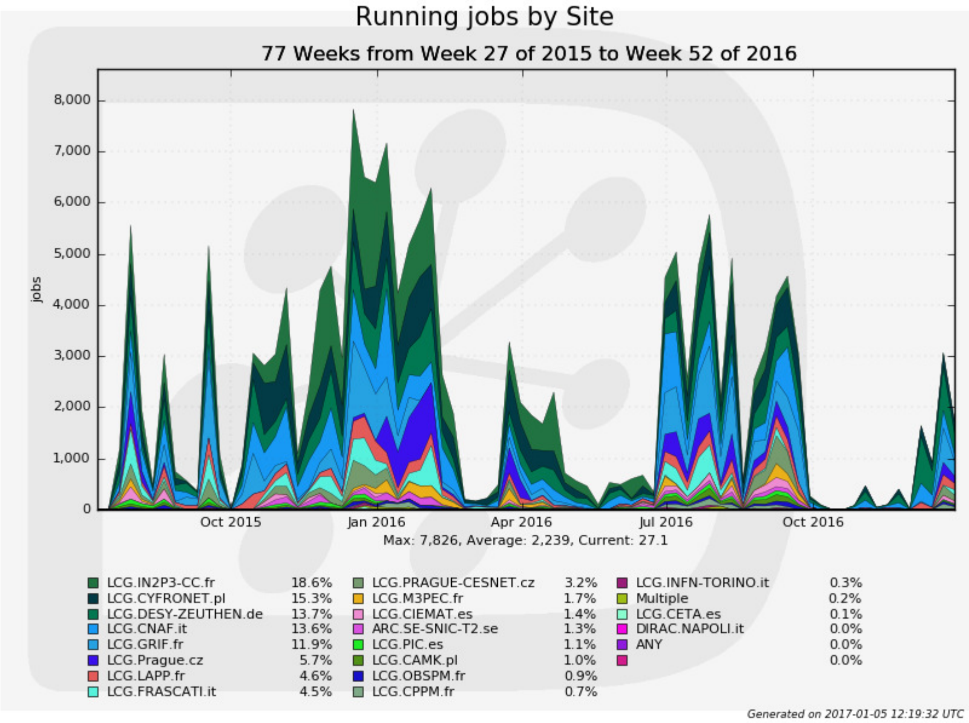

Figure 2. Running jobs during the prod3 MC campaign (July 2015 - December 2016). The long period of sustained activity, from November 2015 to March 2016, corresponds to the bulk part of the production and analysis for the array layout optimisation at the Paranal site. Peaks of 12000 concurrent jobs have been reached during this period, but they are smoothed in the picture. Job sharing among about 10 grid sites is represented by the colour legend.

the general performances of the various DIRAC systems, the hardware upgrade of the DIRAC servers (see section 2.2), solved the slowness problems observed in the early phase of the prod3. Then, even if the DIRAC WMS scalability and functionalities have already been largely demonstrated by other experiments (*e.g.* LHCb [11]), we have confirmed that with our server installation we are able to fulfill the requirements of the future CTA data-processing. Indeed, the peak value of the number of cores needed for the CTA data-processing is of about 10000, which has already been surpassed during the prod3. Moreover, we have also successfully tested the flexibility of the DIRAC WMS regarding the integration of new types of computing resources. Indeed, after an initial usage of exclusively CREAM [12] resources, we have been able to easily integrate new types of resources that became available to CTA, like ARC resources as well as a Torque cluster. In future, we also plan to integrate cloud resources.

Concerning the TS performances, an improvement was obtained, by developing a new functionality, that we have successfully adopted in the last phase of the prod3. This functionality introduces the use of meta-data filters, allowing to reduce the load on the File Catalog. The details of this development are explained in [13].

Finally, even if we have been able to manage different types of workflows, we still need to further automatize the whole workflow execution. In order to achieve this goal, we plan several improvements. First of all, we aim to develop a mechanism to group together several transformations belonging to the same workflow. Indeed, when running several workflows in parallel, the production manager should be able to easily monitor the progress of each workflow, rather than of the individual transformations. Moreover, we envisage to generalize the data

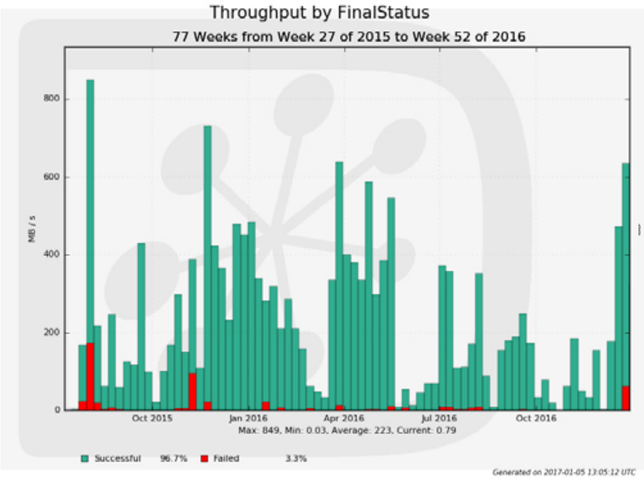

Figure 3. Throughput during the prod3 MC campaign. Successful transfers are represented in green, while failed transfers are in red.

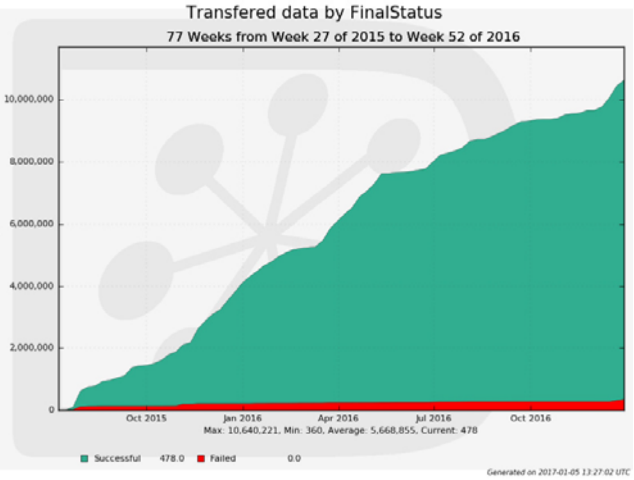

Figure 4. Cumulative data transferred during the prod3 MC campaign. In total, about 11 PB have been transferred. They include all the transfers between the processing sites and the storage locations, both for production and analysis jobs.

driven model to make it an event driven model. The idea is that, not only the presence in the catalog of matching files, would trigger the jobs creation, but also any other kind of predefined event. These may be, the availability of a new software release in the central repository, the completion of a given transformation or whatever other meaningful event.

Finally, in order to further improve the TS performances, we also plan to introduce a Message Queueing (MQ) system to trigger the agent execution, thus partially replacing the polling mode [13]. The use of MQ would also benefit to the transition from a data-driven to an event-driven model. Indeed, messages carrying any useful information could be sent by various components to trigger the jobs creation.

6. Conclusions

In order to manage the CTA production activities over a distributed infrastructure, we have developed a prototype based on the DIRAC interware. This prototype has been extensively exploited during the past four years, in the context of the MC simulations devoted to the CTA site selection and to the array layout optimization. Even if the obtained results are satisfying, both in terms of performances and functionalities, some improvements are still needed in view of the upcoming real data processing. Indeed, CTA will start collecting the first telescope data already in 2018, while science operations are expected to start in 2020. These improvements mainly consist in further automatizing the whole workflow management, as described in section 5. In particular, we aim to develop a high level system, on the top of the TS, allowing to easily describe and manage complex workflows, as those that will be used for the CTA data-processing. Finally, the goal is that this system should be general enough, that could be adopted also by other communities.

Acknowledgments

We gratefully acknowledge financial support from the following agencies and organizations: Ministerio de Ciencia, Tecnología e Innovación Productiva (MinCyT), Comisión Nacional de Energía Atómica (CNEA), Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina; State Committee of Science of Armenia, Armenia; Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ), Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Brasil; Croatian Science Foundation, Croatia; Ministry of Education, Youth and Sports, MEYS LE13012, LG14019, 7AMB14AR005 and the Czech Science Foundation grant 14-17501S, Czech Republic; Ministry of Higher Education and Research, CNRS-INSU and CNRS-IN2P3, CEA-Irfu, ANR, Regional Council Ile de France, Labex ENIGMASS, OSUG2020 and OCEVU, France; Max Planck Society, BMBF, DESY, Helmholtz Association, Germany; Department of Atomic Energy, Department of Science and Technology, India; Istituto Nazionale di Astrofisica (INAF), Istituto Nazionale di Fisica Nucleare (INFN), MIUR, Italy; ICRR, University of Tokyo, JSPS, Japan; Netherlands Research School for Astronomy (NOVA), Netherlands Organization for Scientific Research (NWO), Netherlands; The Bergen Research Foundation, Norway; Ministry of Science and Higher Education, the National Centre for Research and Development and the National Science Centre, Poland; Slovenian Research Agency, Slovenia; MINECO support through the National R+D+I, Subprograma Severo Ochoa, CDTI funding plans and the CPAN and MultiDark Consolider-Ingenio 2010 programme, Spain; Swedish Research Council, Royal Swedish Academy of Sciences, Sweden; Swedish National Infrastructure for Computing (SNIC), Sweden; Swiss National Science Foundation (SNSF), Ernest Boninchi Foundation, Switzerland; Durham University, Leverhulme Trust, Liverpool University, University of Leicester, University of Oxford, Royal Society, Science and Technologies Facilities Council, UK; U.S. National Science Foundation, U.S. Department of Energy, Argonne National Laboratory, Barnard College, University of California, University of Chicago, Columbia University, Georgia Institute of Technology, Institute for Nuclear and Particle Astrophysics (INPAC-MRPI program), Iowa State University, Washington University McDonnell Center for the Space Sciences, USA. The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n^o 262053.

References

[1] Actis M et al. (CTA Consortium) 2011 Design concepts for the Cherenkov Telescope Array CTA: an advanced facility for ground-based high-energy gamma-ray astronomy *Experimental Astronomy* **32** 193-316

[2] Casajus A et al. 2012 Status of the DIRAC Project *Journal of Physics: Conference Series* **396** 032107

[3] Tsaregorodtsev A et al. 2014 DIRAC Distributed Computing Services *Journal of Physics: Conference Series* **513** 032096

[4] `https://znwiki3.ifh.de/CTA/Eventdisplay%20Software`, (last access date: 3rd Feb. 2017)

[5] Zanin R et al. 2013 MARS, the MAGIC analysis and reconstruction software *Proc. 33rd Int. Cosmic Rays Conf. (Rio de Janeiro)* 0773

[6] `http://veritas.sao.arizona.edu/`, (last access date: 3rd Feb. 2017)

[7] `http://magic.mppmu.mpg.de/`, (last access date: 3rd Feb. 2017)

[8] Stagni F et al. 2015 Pilots 2.0: DIRAC pilots for all the skies *Journal of Physics: Conference Series* **664** 062061

[9] `https://cernvm.cern.ch/portal/filesystem`, (last access date: 3rd Feb. 2017)

[10] Breiman L 2001 Random forests, Machine Learning vol 45 issue 1 pp 5-32 doi: 10.1023/A:1010933404324

[11] Stagni F et al. 2012 LHCbDirac: distributed computing in LHCb *Journal of Physics: Conference Series* **396** 032104

[12] `https://wiki.italiangrid.it/CREAM`, (last access date: 3rd Feb. 2017)

[13] Arrabito L et al. 2015 Prototype of a production system for Cherenkov Telescope Array with DIRAC *Journal of Physics: Conference Series* **664** 032001