EPJ Web of Conferences **245**, 03029 (2020) https://doi.org/10.1051/epjconf/202024503029
CHEP 2019

Evolution of the WLCG Information Infrastructure

Julia Andreeva[1*], *Alexey Anisenkov* [2], *Alessandro Di Girolamo* [1], Alessandra Forti [3], Stephen Jones [4], *Balazs Konya* [5], *Andrew McNab* [3], *Panos Paparrigopoulos* [1]

[1]CERN, European Organization for Nuclear Research, Switzerland
[2] Joint Inst. for Nuclear Research, Russia
[3] University of Manchester, UK
[4] University of Liverpool, UK
[5] Lund University, Sweden

Abstract. The WLCG project aimed to develop, build, and maintain a global computing facility for storage and analysis of the LHC data. While currently most of the LHC computing resources are being provided by the classical grid sites, over the last years the LHC experiments have been using more and more public clouds and HPCs, and this trend will certainly continue. The heterogeneity of the LHC computing resources is not limited to the procurement mode. It also implies variety of storage solutions and types of computer architecture which represent new challenges for the topology and configuration description of the LHC computing resources. The WLCG Information infrastructure has to evolve in order to meet these challenges and to be flexible enough to follow technology innovation. It should provide a complete and reliable description of all types of the storage and computing resources to ensure their effective use. This implies changes at all levels, starting from the primary information providers, through data publishing, transportation mechanism and central aggregators. This paper describes the proposed changes in the WLCG Information Infrastructure, their implementation and deployment.

1 Introduction

For their computing activities the experiments of the Large Hadron Collider (LHC) [1] rely on WLCG [2] a global collaboration of more than 170 computing centres in 42 countries, linking up national and international grid infrastructures, including EGI [3], OSG [4] and NorduGrid[5]. Over the last years the resources of the classical WLCG sites are complemented with resources provided by commercial clouds and HPCs. Moreover, new types of computing and storage architectures require changes in the workload and data management systems of the experiments. Resources provided by WLCG distributed sites are more homogeneous and better controlled compared to those which are used opportunistically.

* Corresponding author: Julia.Andreeva@cern.ch

The fraction of the resources used opportunistically is steadily increasing. Similarly, there is a constant trend for growth of their heterogeneity. The more dynamic and heterogeneous nature of the provided resources dictates a need for the evolution of the WLCG Information Infrastructure (II).

The first section of this paper overviews the current components of the WLCG II and their limitations. The following sections describe the main challenges for the II evolution, requirements for the new II generation and implementation of the new components.

2 Current state of the WLCG Information Infrastructure

2.1 Overview of the current components of the WLCG II

WLCG II consists of several components which handle static and/or dynamic information. Some of them like GocDB [6] or REBUS [7] represent central repositories, others like BDII[8] have a complex distributed and hierarchical structure. In addition to the components provided centrally, LHC experiments developed their own topology and configuration systems which are experiment-specific and work only in the scope of the experiment.

2.1.1 Centrally provided components

Static information related to the sites and services of the EGI and NorduGrid infrastructures is provided by the GocDB system. GocDB contains site information like site names, country where the site is located and its geographical coordinates, information about site support unit and some other site attributes. GocDB also represents a central repository of static information describing services hosted by the sites. Main service related attributes are: service type, service endpoint and its status. GocDB also keeps a track of service/site downtimes. For sites belonging to the OSG infrastructure there is an OSG repository of the topology information [8].

Collaboration with WLCG is typically formalised through negotiating and signing a Memorandum of Understanding (MoU) [9]. The level of provided service and amount of the provided resources is agreed between WLCG and a particular federation which is normally organized at the national level. The level of service of a particular site is defined by the *tier* the site is allocated to. The amount of resources provided by federations is defined by so-called pledges which are agreed on annual basis for CPU and storage, including tape and disk media. Information related to MoU agreement is described in REBUS system. REBUS contains topology information defining which sites belong to which federation, which *tier* a given site represents for a given experiment as well as pledge values.

BDII stands for Berkeley Database Information Index, it is a distributed system which has three levels of hierarchy: service, site and top level BDII. Every service hosted by a site is supposed to provide information for BDII. In contrast with REBUS and GocDB, BDII contains both static and dynamic information, as for instance the number of running jobs or the number of jobs in the queue. BDII has been developed having in mind a fully distributed operational model and a certain dependency of the operational tasks on the dynamic information.

2.1.2 Experiment-specific information systems

In addition to the central components of the WLCG II, every LHC experiment developed its own topology and configuration system. Such systems aim to collect and consolidate data coming from the central sources and complement it with the configuration which is required for experiment workload and data management systems. Examples of information which might be specific for a given experiment are: access protocols and location for input and output data at a particular site, configuration parameters used for job submission to a given computing resource, site naming convention used in the experiment scope, etc. Therefore, all experiment systems re-implement similar tasks of data collection and consolidation and then enrich the obtained data with their specific configuration.

2.1.3 Current limitations

WLCG relies on several GRID infrastructures: EGI, OSG and NorduGrid. Central components described above are mainly used by the EGI component of WLCG, while for example, OSG stopped using BDII several years ago. This implies additional complexity for the information handling which depends on the location of a particular site or service. Moreover, data reliability is often an issue. Data is coming from multiple data sources which sometimes contradict to each other. It is not easy to understand which source is correct and how the inconsistencies can be fixed. This task has to be performed by every LHC experiment individually and there is no central service which can validate data centrally and provide it to all interested clients.

Overall complexity of the II, in particular its BDII component is another drawback. In particular, it represents a problem for site administrators who need to support BDII services at the site and have to make sure that information providers are properly configured. For example, many sites did not succeed to enable proper description of the WLCG storage services in BDII. As a result, none of the LHC experiments rely on BDII for storage service topology description and for this purpose use experiment-specific systems.

3 Major challenges and conditions to be considered

3.1 Challenges to be addressed

One of the main challenges for WLCG II is an ability to react and adapt quickly to the technology evolution, changes in the procurement modes and computing models of the experiments. Heterogeneous resources available to the experiments have to be described in the II in order to be integrated with the experiment data and workload management systems. Moreover, a large fraction of those resources is used opportunistically, therefore, it is not realistic to enforce everywhere systems like BDII. Rather some lightweight approach has to be used for service and resource description.

Another important goal is the improvement of data reliability. Data coming from multiple sources easily become contradictory. For every use case and data type a canonical information source has to be defined and a mechanism for data validation has to be provided.

Resource optimization plays an important role for the overall WLCG service evolution. Therefore, WLCG II should handle information required for data optimization. An example of this type of information is the quality of service information (QoS) for storage services. QoS information allows experiments to optimize the usage of the provided storage. Another example is the network matrix topology which is needed to monitor performance of various network channels and to take educated decisions regarding data transfers.

All those challenges have to be addressed in a common way, for the benefit of all experiments, to avoid any duplication of the development and operational effort.

3.2 Conditions to be considered

Initial design of the WLCG II took into account the fully distributed operational model of EGI. In contrast with EGI, WLCG operational model is a centralized one. This can simplify the implementation of data flows and data consolidation procedures.

Another important factor which should be considered, is the dependency of the experiment workflows on the dynamic information, as for example, the number of jobs in the queue. Analysis of the requirements performed by the WLCG Information Evolution task force confirmed that LHC experiments mainly rely on static and semi-static information. This can simplify the implementation of primary service-level data sources which are described in the next section.

One of the positive trends in the recent evolution of the computing systems of the LHC experiments is the sharing common solutions, like for example, Rucio [10] for the data management. This implies an opportunity to implement common data models and interfaces.

Finally, new developments can benefit from extensive experience accumulated while developing and operating experiment-specific information systems. For example, the development of the Computing Resources Information Catalogue (CRIC) [11] has been inspired by the success of the ATLAS Grid Information System (AGIS) [12].

4 WLCG II evolution

There are three main requirements which should be considered in the design of the new generation of WLCG II: flexibility, extensibility and data reliability.

Comparison of data quality between centrally provided information systems and experiment-specific ones is not in favour of central systems. There are several reasons why experiment-specific systems like AGIS or DIRAC [13] are more successful in providing reliable data. First of all, data is being constantly validated by its continuous usage. The content is defined by the needs of operations and is limited to information which is being actively used. Normally, experiment-specific systems provide functionality allowing to change faulty information in place without waiting for data to be corrected at the primary source and then propagated through the complete chain, since it can take long time and therefore has negative impact on operations. All those principles have to be considered while developing new components of the WLCG II.

In order to cope with ever-changing computing environment, extensibility and flexibility are mandatory for the WLCG Information Infrastructure. How these requirements are ensured by the design and implementation of the new generation of the information infrastructure will be described below.

4.1 New components

New components of the WLCG II can be split in two categories: primary data sources describing service-level information and central aggregators. In the current implementation static and dynamic service-related information is provided via BDII. In addition, service endpoints are recorded in GocDB. As already mentioned above, usage of BDII is limited to the classical EGI GRID sites, it is not used neither for the OSG component of the WLCG

infrastructure nor for the new resource types, like HPC. In the new implementation, the description of storage and compute resources will be provided through Storage Resource Reporting (SRR) and Compute Resource Reporting (CRR). SRR and CRR mainly provide static topology and configuration information. This information is presented in a JSON format and should be accessible through the http protocol. SRR and CRR files are testable against a version controlled JSON schema. The URLs of the SRR and CRR files are being recorded in central systems like CRIC and GocDB. The content of the files represents minimal, clean and consolidated data set which can be easily provided for any storage or compute resource and can be generated from a simple configuration file. Current SRR and CRR format can be extended to ensure flexibility of the information system considering topology and configuration data content.

4.2 SRR and CRR structure

Both SRR and CRR structures consist of three sections:

- Common description of the resource including attributes like: unique identifier of the resource (computing cluster or storage capacity), name of the site owning the resource, implementation, implementation version, resource capacity, status and time stamp when the file has been generated. In SRR case, the implementation is defined by the middleware of the underlying storage service, for example, EOS[14], dCache [15], DPM [16], etc. In CRR, the implementation generally points to the type of the batch system managing the computing cluster, for example, LSF, THCondor [17], etc.
- Section describing interfaces which are used in order to access the resource. In SRR it describes the set of storage protocols, for example, xrootd, http, srm, etc. In CRR it describes the set of computing elements which interface a given computing resource
- Section describing internal organization of the resource in terms of storage shares (for SRR) or queues (for CRR)

Unlike the CRR, SRR can contain some dynamic information consisting of storage accounting data. Storage middleware providers have enabled generation of SRR including storage accounting information. Massive deployment campaigns aiming to deploy storage versions with enabled SRR generation functionality are currently ongoing on the WLCG infrastructure.

4.3 Computing Resource Information Catalogue (CRIC)

CRIC represents a new component of the WLCG II. It belongs to the second type of category mentioned above - central aggregator. CRIC is a high level information system providing both the topology of the WLCG infrastructure and other resources used by the LHC experiments (HPC, clouds, etc.) and experiment-specific configuration required to exploit these resources according to the experiments computing models. CRIC was inspired by AGIS, which has evolved towards a common solution. Being generic enough, CRIC can be used not only by the LHC experiments but also by other communities, and beyond the scope of a single experiment, for central WLCG operations.

The system collects information from a variety of information sources, allows to consolidate and validate data and then provides it in a consistent way to all interested clients. Described data flow is shown in figure 1.

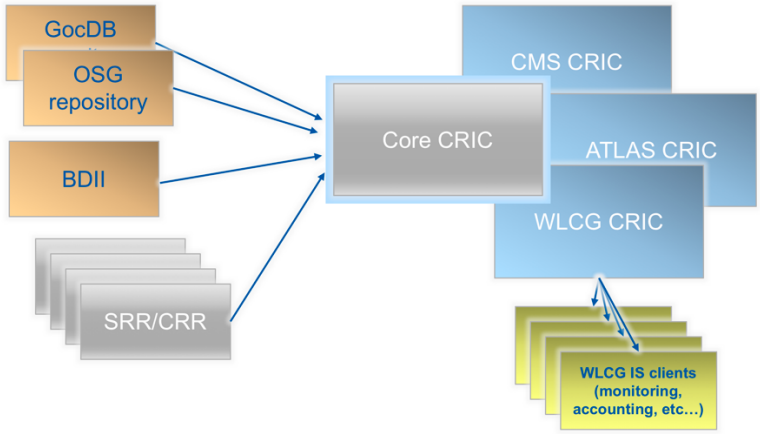

Fig. 1. CRIC data flow.

4.3.1 CRIC implementation

One of the main CRIC design principles is: "Share what is common, customize what is specific". The system is plugin-based which allows straightforward customization to address various experiment requirements via the dedicated experiment instances. Modular architecture based on Django framework enables a lego bricks-like approach. It implies usage of shared building blocks which enables common look and feel and ensures optimization of the development process.

CRIC is very flexible regarding:

- Primary information sources and corresponding collectors
- Authentication authorization methods and utilization of Permissions, Roles and Groups at various level
- Customized UIs and APIs
- Customized data models and configuration

The system can be easily extended in order to follow technology evolution and changes in the experiment applications.

Every CRIC instance consists of the core part which describes resources provided by the infrastructure (central rectangle in figure 2) and one or more plugins (blue rectangles in figure 2) containing configuration and data models specific to a particular community or a particular activity which exploits provided resources.

There are several CRIC instances which are currently in production or exist as prototypes:

- CRIC for CMS experiment
- CRIC for WLCG central operations
- CRIC for ATLAS experiment
- CRIC for third party copy functional tests

All REBUS functionalities have been implemented in CRIC instance for WLCG central operations. By summer 2020 REBUS will retire.

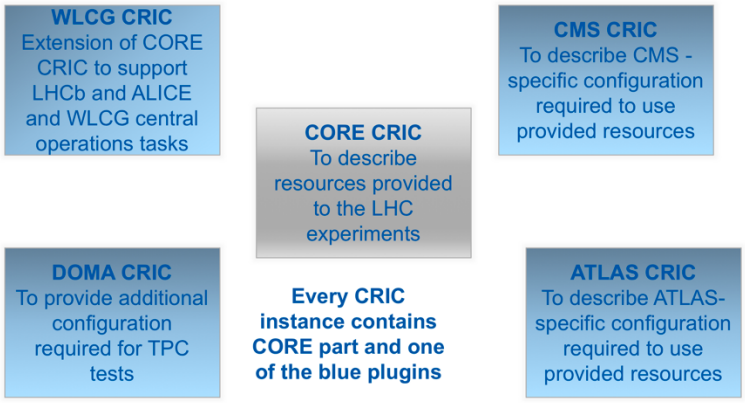

Fig. 2. Structure of the CRIC instances.

5 Conclusions

WLCG II evolves towards an implementation which allows to address the needs of the computing infrastructure which is becoming more dynamic and heterogeneous. New components aiming to improve data quality, flexibility and extensibility of the information infrastructure have been developed and are being deployed. These components are designed in a way that can be easily adapted for the new types and architectures of the storage and compute resources and are generic enough to be used beyond the WLCG scope.

References

1. LHC: https://home.cern/science/accelerators/large-hadron-collider
2. I. Bird, *Computing for the Large Hadron Collider*, Annual Review of Nuclear and ParticleScience, **61**, 99-118 (2011)
3. EGI: https://www.egi.eu/
4. OSG : https://opensciencegrid.org/
5. Eerola, Paula; et al. (2003), *The NorduGrid production Grid infrastructure, status and plans*. Proceedings of Fourth IEEE International Workshop on Grid Computing: 158–165
6. G. Mathieu, A. Richards, J. Gordon, C.D.C. Novales, P. Colclough, and M. Viljoen. *GOCDB, a topology repository for a worldwide grid infrastructure*, J. Phys.: Conf. Ser. **219**, 062021 (2010)
7. REBUS: https://gstat-wlcg.cern.ch/apps/topology/
8. OSG Topology Interface: https://my.opensciencegrid.org/
9. M. Schulz, L. Field, M. Alandes, *Grid Information Systems: Past, Present and Future*, 2020 in preparation for the proceedings of CHEP 2019 Conference, Adelaide (Australia)
10. WLCG MoU: https://wlcg.web.cern.ch/mou
11. Barisits, M., Beermann, T., Berghaus, F. et al, *Rucio: Scientific Data Management Computing and Software for Big Science*, 3, 11 (2019)
12. A. Anisenkov , J. Andreeva, A. Di Girolamo, P. Paparrigopoulos, B. Vasilev, *A unified topology system for a large scale, heterogeneous and dynamic computing infrastructure*, 2020 in preparation for the proceedings of CHEP 2019 Conference, Adelaide (Australia)

12. A. Anisenkov, A.D. Girolamo and M.A. Pradillo, *AGIS: Integration of new technologies used in ATLAS Distributed Computing,* J. Phys.: Conf. Ser.898, 92023 (2017)
13. F. Stagni, A. Tsaregorodtsev, C. Haen, P. Charpentier, Z. Mathe, W. J. Krzemien and V. Romanovskiy , *LHCb and DIRAC strategy towards the LHCb upgrade*, CHEP 2018, EPJ Web of Conferences **214**, 03012 (2019)
14. G. Bitzes, F. Luchetti, A. Manzi, M. Patrascoiu, A. J. Peters, M. K. Simon, E. A. Sindrilaru, *EOS architectural evolution and strategic development direction 2020,* in preparation for the proceedings of CHEP 2019 Conference, Adelaide (Australia)
15. T. Mkrtchyan et al, *dCache - storage for advanced scientific usecases and beyond*, CHEP 2018, EPJ Web of Conferences **214**, 04042 (2019)
16. F. Furano, O.Keeble, A. Manzi, G. Bitzes, *A milestone for DPM*, CHEP 2018, EPJ Web of Conferences **214**, 04018 (2019)
17. B Bockelman, T Cartwright, J Frey, E M Fajardo, B Lin, M Selmeci, T Tannenbaum and M Zvada, Commissioning the HTCondor-CE for the Open Science Grid, Journal of Physics: Conference Series, Vol. **664**, (2015)