Journal of Physics: Conference Series **513** (2014) 062038 doi:10.1088/1742-6596/513/6/062038

The GridKa Tier-1 Computing Center within the ALICE Grid Framework

WooJin J. Park[a,*], Jung Christopher[a,*], Andreas Heiss[a], Andreas Petzold[a], Kilian Schwarz[b,*]

[*](for the ALICE Offline Collaboration)
[a] Karlsruhe Institute of Technology, Karlsruhe, 76131, Germany
[b] GSI Helmholtzzentrum für Schwerionenforschung, Darmstadt, 64291, Germany

E-mail: WooJin.Park@kit.edu

Abstract. The GridKa computing center, hosted by Steinbuch Centre for Computing at the Karlsruhe Institute for Technology (KIT) in Germany, is serving as the largest Tier-1 center used by the ALICE collaboration at the LHC. In 2013, GridKa provides 30k HEPSPEC06, 2.7 PB of disk space, and 5.25 PB of tape storage to ALICE. The 10Gbit/s network connections from GridKa to CERN, several Tier-1 centers and the general purpose network are used by ALICE intensively. In 2012 a total amount of ∼1 PB was transferred to and from GridKa. As Grid framework, AliEn (ALICE Environment) is being used to access the resources, and various monitoring tools including the MonALISA (MONitoring Agent using a Large Integrated Services Architecture) are always running to alert in case of any problem. GridKa on-call engineers provide 24/7 support to guarantee minimal loss of availability of computing and storage resources in case of hardware or software problems. We introduce the GridKa Tier-1 center from the viewpoint of ALICE services.

1. Introduction

The GridKa computing center, hosted by SCC (Steinbuch Centre for Computing [1]) at KIT (Karlsruhe Institute for Technology [2]) is a member of the WLCG (Worldwide LHC Computing Grid) and the largest German Tier-1 center in WLCG. It supports all LHC experiments as well as non-LHC experiments in high energy physics, nuclear physics and astrophysics such as Auger, BaBar, Belle, Belle2, CDF, Compass, and D0. The available resources in 2013 are 135k HEPSPEC06, 11 PB of disk space and 17 PB of tape storage, and it accounts for ∼14% of the whole WLCG resources. Table 1 shows the allocated resources for each LHC experiment. ALICE and ATLAS have the biggest shares of GridKa's resources.

2. ALICE Grid Computing

ALICE (A Large Ion Collider Experiment [3]) is one of the four main experiments at the LHC (CERN) and is the dedicated heavy-ion experiment. More than 1,200 people from 132 institutes in 36 countries are collaborating to study the new state of matter called Quark-Gluon Plasma (QGP), the energy density of 10 GeV/fm^3 and the temperature of 200 MeV or higher, created during the ultra-relativistic collisions of heavy ions. ALICE is producing an overall data volume of ∼5 PB every year.

Experiment	CPU(HS06)	Disk(PB)	Tape(PB)	WLCG(%)
ALICE	**30,000**	**2.7**	**5.3**	**25**
ATLAS	39,875	4.1	5.0	12.5
CMS	17,500	2.6	5.0	10
LHCb	19,200	1.5	1.8	16.7

Table 1. GridKa resources for each LHC experiment. ALICE is highlighted in bold.

2.1. Data Production and Analysis Framework

Data processing in ALICE is based on the AliRoot [4] offline framework. It uses the ROOT [5] system as a foundation on which the framework for simulation, reconstruction and analysis is built. This framework is based on the Object-Oriented programming paradigm, written in C++, and it provides common tools for processing ALICE data in an efficient way. To run a given analysis algorithm using the framework, one needs to create a ROOT macro that can be easily assembled.

ALICE also developed the "analysis train", which is a way to run analysis in an efficient way over a large part or the full dataset. The train is assembled from a list of modules that are sequentially executed by the common object. All tasks will process the same dataset of input events, share the same event loop and possibly extend the same output with their own information produced in the event loop. It provides a framework to optimize CPU/IO ratio, accessing data via a common interface and making use of GRID infrastructures.

2.2. AliEn Middleware

The access to the GRID resources and data is provided by the AliEn (ALICE Environment [6]) middleware. AliEn has been developed based on web services and standard protocols. Within AliEn, all jobs are inserted in a global queueing system and distributed to the computing resources for optimal resource load. The system takes care of job splitting and execution and tries to optimize network traffic. Each computing site can be seen and used as a single entry. The middleware also offers a virtual file and metadata catalogue. Using the AliEn API, ALICE users can transparently access datasets on the Grid as if there were local files.

From the ROOT prompt, users can authenticate themselves, access distributed datasets or request the execution of their algorithms across distributed datasets and retrieve the resulting objects in interactive or batch session. When required, the datasets can be replicated and cached. The result of each job is optionally validated and the final output from many concurrent jobs is merged together and presented to the user as a dataset in his/her portion (directory) of the global logical file namespace (AliEn File Catalogue).

The concept of a master job combined with job splitting mechanism allows the handling of large scale productions (bulk job submission with a single command). For example, in a Monte-Carlo production a single master job can be spawned into hundreds of sub-jobs with different random seeds. For jobs which require specific input structure, e.g. analysis tasks, the input data file list is taken as a basis to split the jobs according to the file location.

The Authentication Service is responsible for checking users credentials. AliEn uses the SASL [7] protocol for authentication and implements several SASL mechanisms (GSSAPI using Globus/GSI, AFS password, SSH key, X509 certificates and AliEn tokens). Upon successful authentication, a Proxy Service acquires and holds the real database handle on behalf of a user and returns a temporary access token which the user has to present in order to re-connect to the database. The token remains in user possession and is valid for a limited period of time.

2.3. MonALISA Monitoring Service

MonALISA (MONitoring Agents using a Large Integrated Services Architecture [8]) is the main monitoring framework as a part of the AliEn monitor module. It collects the monitoring information and publishes it via web service. The framework is based on Dynamic Distributed Service Architecture and is able to provide complete monitoring, control and global optimization services for the entire AliEn grid system. It also provides the integrated web-based interface, where one can easily monitor the system information for computer nodes and clusters, network information (traffic, flows, connectivity, topology), the performance of applications, jobs or services, and the end-to-end performance measurement.

3. GridKa Tier-1 Center within the ALICE Framework

GridKa center is serving as the largest Tier-1 site for ALICE and is providing ∼25% of the whole Tier-1 requirements. In this section, we will describe how GridKa is configured and how it works within the ALICE Grid framework.

3.1. Computing Resources for ALICE

GridKa is providing 30k HEPSPEC'06 of computing power for ALICE, plus 2.7 PB of disk and 5.25 PB of tape space for ALICE. The VOBOX has been upgraded recently to the WLCG VOBOX, which provides an EMI UI, a GSI-OpenSSH service and a proxy renewal and VO agent service. A total of 8 CREAMs (Computing Resource Execution And Management) are running for the job management operation at the CE level. WNs are installed with Scientific Linux 6.4 and their characteristics are collected by the CEs that publish all information to the BDII (Berkeley Database Information Index) services. UniVa GridEngine [9] is being used as a batch-queuing system and is able to manage the whole cluster. ALICE software packages are distributed via CVMFS (Cern Virtual Machine File System) network file system.

3.2. Storage

Both the disk and the tape elements are using the XRootD [10] middleware. 10 XRootD servers are running behind the disk storage element (ALICE::FZK::SE), and 2 redirectors connected to all file servers are on top of them. About one quarter of the disk space is being used as a tape buffer, serving a tape element (ALICE::FZK::TAPE), and another two redirectors are running for the tape system. The disk space is partitioned in about 100 mount points, each having a size between 15 TB and 60 TB, and all underlying file systems are set up with the GPFS (General Parallel File System [11]) file system.

The tape-backed storage element uses the "Migration-Purge-Staging Support (MPS)", which was designed as enhanced support of the Open Storage System (OSS) scalable file system as an XRootD extended feature. It allows the system to transfer files from disk to tape and back as well as to purge archived files from disk. The MPS daemon executes scripts which have been adjusted to call local migration and staging commands. At GridKa, the Tape Staging Service (TSS), a locally developed software, communicates with the Tivoli Storage Manager (TSM) back-end. This back-end uses the IBM Enterprise Removable Media Manager (eRMM) [12], a service virtualizing three TSM libraries.

3.3. Network

The GridKa core network is organized with one backbone. The WNs are aggregated per rack with a switch via 2×10 Gbit/s uplink Ethernet to the backbone. File servers are directly connected to the backbone, in order to allow full utilization of the 10 GE interface through the backbone. The edge router is connected to the 10 GE LHC OPN(Large Hadron Collider Optical Private Network), which is a spanned layer-2 tunnel network between all Tier-1 centers within the LHC project and the Tier-0 center at CERN.

The firewall (for traffic through the general internet uplink to X-WIN [13]) has been upgraded in the calendar week 28 in 2013, and is now able to cope with 100 Gb/s throughput. It brings significant improvement to the job efficiency, especially when the jobs are reading (or writing) the data from (or to) remote sites.

3.4. Monitoring

The MonALISA service is the first place to monitor ALICE services at GridKa. It is already providing many essential information on the site facilities, network and many ongoing tasks in real time. However, more detailed monitoring inside GridKa is needed to diagnose individual components and track down (possible) hidden problems. The Icinga [14] and Ganglia [15] monitoring systems are chosen for these purposes, and are intensively used. Usage of the CPU, memory, disk and internal network for every single computing node as well as the status of the cluster and batch jobs are monitored and summarized in the GridKa internal webpages using the Ganglia web interface. Icinga makes it possible to test all resources and services, and sends alarms in case of failures. It also keeps watching over any conceivable network resource, it notifies the user of errors and recoveries, and it generates performance data for reporting. These features are connected to the on-call alarm system, so that GridKa resources are at all time supervised by the highly trained engineers.

3.5. ALICE On-site Representative

GridKa is supporting all LHC experiments, and each VO requires a different infrastructure and workflow. Different services and middleware components make the communication and maintenance difficult and induced the establishment of full-time positions for experiment representatives since 2010. They are well connected to both GridKa local staff and the experiment. Their major task is the communication between the experiment and the site admins and the services administration. Based on their experience and knowledge in both fields, they interface between GridKa and the experiment in an appropriate way. The representative participates all the GridKa and ALICE regular meetings, and communicates with local staffs at GridKa with all parties within the ALICE experiment and with other grid sites and ALICE users. The ALICE representative is also serving as an XRootD administrator for a successful operation of the ALICE storage services.

Year	CPU(HS06)	Disk(PB)	Tape(PB)
2015	27.5 k	2.625	5.3
2016	32.5 k	2.975	8.7
2017	40 k	3.825	12.2

Table 2. ALICE requirements at GridKa Tier-1 center

4. Performances in 2013

About 11% of the total ALICE jobs ran at the GridKa center in 2013. They correspond to the 3762 averaged number of active jobs and 22.3 million hours of CPU time. The overall job efficiency (defined as the ratio between CPU time and wall clock time) is 82%. Since the efficiency has been increased after the firewall upgrade, we expect much higher averaged efficiency value in the second half of 2013. GridKa has served an averaged XRootD traffic of 713 MB/s to the WAN in 2013, but this includes several months of imperfect XRootD services due to the migration of the file servers.

There was an issue of the frequent failure of the ALICE functional tests on the XRooD servers, when 4 XRootD file servers were upgraded to new and powerful machines. The problem was fixed by removing the option for the distributed filesystem in the server configuration in early September 2013. The availability of the XRootD service is 100% since then.

5. Summary and Outlook

ALICE Grid Computing at the GridKa Tier-1 center has been running successfully for several years, thanks to the very well established services and frameworks. The support of ALICE computing groups and the endless efforts of the GridKa team are an indispensable factor for this great success.

A series of recent upgrades to the internal firewall and XRootD file servers have improved the performance and the efficiency of ALICE jobs. The continuous updates of the software packages and the framework support maximized the utilization of GridKa resources.

ALICE is increasing the requirements for the computing resources every year, and GridKa has pledged them. 40k HEPSPEC06 CPU capacity, 3.825 PB of disk and 12.2 PB of tape space is anticipated in 2017 (see Table 2). Besides the technical aspects, GridKa is also aiming for a strong collaboration with other Tiers and the ALICE group.

References

[1] Steinbuch Centre for Computing, http://www.scc.kit.edu
[2] Karlsruhe Institute for Technology, http://www.kit.edu
[3] ALICE Experiment, http://aliceinfo.cern.ch
[4] AliRoot, http://aliweb.cern.ch/Offline/AliRoot/Manual.html
[5] ROOT, Rapid Object-Oriented Technology, http://root.cern.ch
[6] ALICE Environment, http://alien2.cern.ch
[7] Simple Authentication and Security Layer, http://asg.web.cmu.edu/sasl/
[8] MonALISA, http://monalisa.caltech.edu/monalisa.htm
[9] UniVa GridEngine batch-queuing system, http://www.univa.com/products/grid-engine
[10] XRootD project, http://xrootd.org
[11] IBM GPFS file system, http://www-03.ibm.com/systems/software/gpfs/
[12] IBM eRMM, http://www-935.ibm.com/services/de/de/it-services/enterprise-removable-media-manager-ermm.html
[13] X-WIN, the German national research and education network, https://www.dfn.de/en/xwin/
[14] ICINGA, http://www.icinga.org/
[15] Ganglia Monitoring System, URL http://ganglia.sourceforge.net
[16] LHC-OPN, https://twiki.cern.ch/twiki/bin/view/LHCOPN