Journal of Physics: Conference Series **396** (2012) 042032 doi:10.1088/1742-6596/396/4/042032

ALICE Grid Computing at the GridKa Tier-1 Center

Abstract. The GridKa center at the Karlsruhe Institute of Technology is the largest ALICE Tier-1 center. It hosts 40,000 HEPSEPC'06, approximately 2.75 PB of disk space, and 5.25 PB of tape space for the 'A Large Ion Collider Experiment' (ALICE), at the CERN Large Hadron Collider (LHC). These resources are accessed via the AliEn (ALICE Environment) middleware. The storage is divided into two instances, both using the storage middleware xrootd. We will focus on the set-up of these resources and on the topic of monitoring. The latter serves a vast number of purposes, ranging from efficiency statistics for process and procedure optimization to alerts for on-call duty engineers.

C Jung[1], A Petzold[1], C-E Pfeiler[1] and K Schwarz[2]

[1] Steinbuch Centre for Computing, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

[2] GSI Helmholtz Centre for Heavy Ion Research, Darmstadt, Germany

E-mail: christopher.jung@kit.edu, andreas.petzold@kit.edu, christoph-erdmann.pfeiler2@kit.edu, k.schwarz@gsi.de

1. Introduction

1.1. ALICE Grid Computing

The ALICE experiment is one of the four detectors at CERN LHC. It has been designed to study the physics of strongly interacting matter at extreme energy densities. The ALICE experiment will primarily focus on heavy-ion collisions (in particular Pb-Pb), but also has a physics program for proton-proton and proton-ion collisions.

The overall data volume per year is estimated to be about 5 PB without replicas. Half of this is RAW data, the other half consists of first and second level RAW reconstruction, simulated data and user generated data. RAW data and user files have two replicas, first and second level RAW reconstruction data have three replicas.

Within ALICE, the Tier-1 centers store parts of the RAW data as well as parts of the first and second level RAW reconstruction data. Typically, the computing jobs running at Tier-1 centers are second pass reconstruction, Monte Carlo simulations, and user analysis. Currently, there are six ALICE Tier-1 centers worldwide [1].

1.2. GridKa Tier-1 Center

The GridKa Tier-1 center is hosted by the Steinbuch Centre for Computing (SCC) at the Karlsruhe Institute of Technology (KIT). It was established in 2002 as a regional computing center for the LHC experiments. Today, GridKa supports all four LHC experiment VOs, seven additional VOs from High Energy Physics and Astroparticle Physics (Auger, BABAR, Belle, Belle2, CDF, Compass, and D0) and several VOs from different fields of science besides physics. ALICE and ATLAS have the biggest shares in GridKa's resources (see Table 1). For historical

	CPU capacity [HEPSPEC'06]	disk space [TB]	tape space [TB]
ALICE	40,000	2,700	5,250
ATLAS	32,380	3,375	4,500
CMS	15,000	2,200	4,600
LHCb	19,200	1,610	1,050

Table 1: GridKa's pledges for the April 2012 milestone for the LHC VOs.

reasons, the ALICE site name for GridKa is FZK (Forschungszentrum Karlsruhe and University of Karlsruhe merged to form KIT in 2009).

2. ALICE Grid Computing at the GridKa Tier-1 Center

ALICE uses several grid services at GridKa, which will be discussed in the following subsections. All LHC VOs use gLite-based services [2], yet they do not all use the very same services, e.g. ALICE does not use File Transfer Service (FTS) or Workload Management System (WMS).

2.1. Computing resources

Within the AliEn framework, three kinds of grid nodes are used for submitting and running computing jobs:

- VO box,
- Computing Resource Execution and Management Computing Element (CREAM CE) and
- Worker Node (WN).

The VO box provides the gLite user interface and runs one gLite (proxy renewal) as well as five AliEn services (CMreport, ClusterMonitor, PackMan, AliEn CE, MonALISA), which are responsible for direct job submission into the cluster, for monitoring, and for software installation. Additionally, a user interface for AliEn is provided. This grid node is a critical component for ALICE grid computing. A few local monitoring scripts run on this machine, mainly collecting information for debugging purposes (i.e. translation of AliEn job IDs into CREAM CE job IDs).

At GridKa, there are two independent clusters of compute nodes. For both clusters, Portable Batch System Pro (PBS) [3] is used for job scheduling and fair share. Each computing cluster can be accessed by several CREAM CEs (see Figure 1). Within the AliEn Framework, the VO box directly submits the jobs to the CREAM CEs.

The compute nodes are gLite WNs. The experiment software directory is mounted on all WNs. In the near future, experiment software will be distributed by an AliEn torrent service.

2.2. Storage

At GridKa, there are two storage elements for ALICE, both using xrootd middleware [4]. About three quarters of the overall ALICE disk capacity at GridKa are installed in the disk-only ALICE::FZK:SE, while one quarter is used in the tape-backed ALICE::FZK:TAPE.

For GridKa storage, all underlying file systems are set up with General Parallel File System (GPFS) [5]. The ALICE disk space is partitioned in about 100 mountpoints, with each having a size between 15 TB and 33 TB. Although each mountpoint is mounted on at least two file servers, each mountpoint is exported to xrootd only on one file server. Each xrootd instance has two redirectors, which are connected to all file servers. A schematic depiction is given in Figure 2. On ALICE::FZK:SE, the redirectors run on two file servers, while on ALICE::FZK:TAPE, they run on dedicated virtual machines. As redirectors do not consume many computing resources,

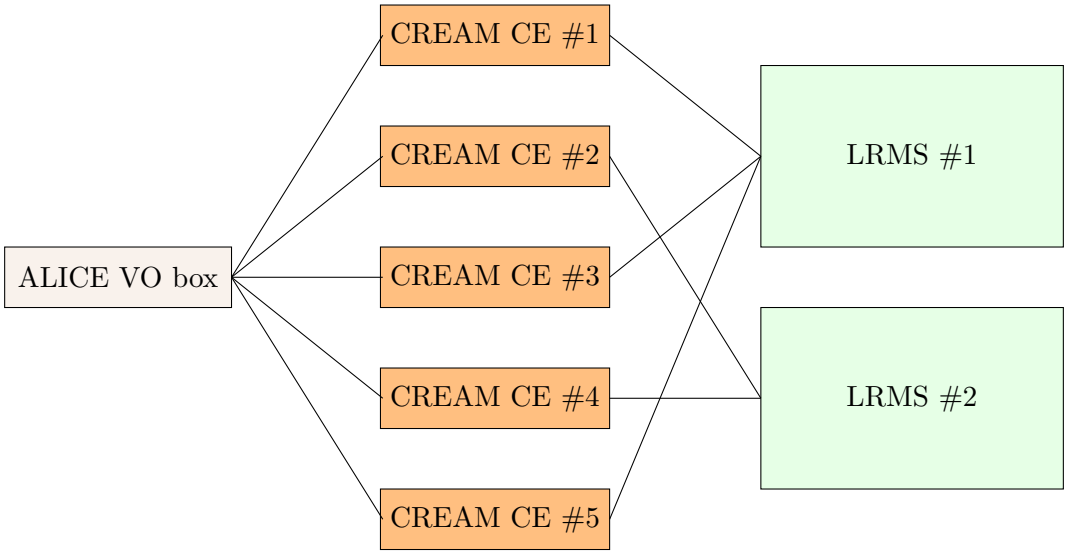

Figure 1: The VO box submits the jobs directly to the CREAM CEs, which schedule the jobs on one of the clusters.

no deficiency in performance has been noticed. On most of the file servers, the xrootd name space directory is located on a file system which is also used for storing data; on new machines, dedicated mount points for the meta data directory have been set up.

The tape-backed storage element uses the xrootd extended features supplement "Migration-Purge-Staging Support" (MPS) for transferring files from disk to tape and back as well as for purging archived files from disk. The system has been set up to purge archived files from a file system when it is at least 90% full. In this case, the least recently used files are deleted until 20% of the file system are free again.

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) [6], a service virtualizing three TSM libraries.

2.3. Monitoring

Both performance monitoring and status monitoring are performed on two levels: ALICE higher level service monitoring uses MonALISA, machine monitoring uses Icinga [7] and Ganglia [8].

Status monitoring is most important for discovering and reporting major problems, e.g. services not running or partitions running full. MonALISA allows site admins to subscribe to e-mail alerts in case of a status change. It also summarizes the status of each ALICE grid site on one web page [9] (see Figure 3). Icinga can produce an overall status by combining the states of several services. At GridKa, the overall status of critical services is used for alerting 24/7 on-call operators via text message.

The quantitative state of services, e.g. data throughput or number of computing jobs running, can be checked in MonALISA and Ganglia. In MonALISA, a site administrator can compare the performance of his/her site with the performance of other sites. In the GridKa Ganglia plots, the state of ALICE services and the behavior of ALICE jobs (i.e. CPU/walltime efficiency) can be compared with the values of other VOs. Discrepancies in either system are often clear indicators for problems or possibilities for optimization.

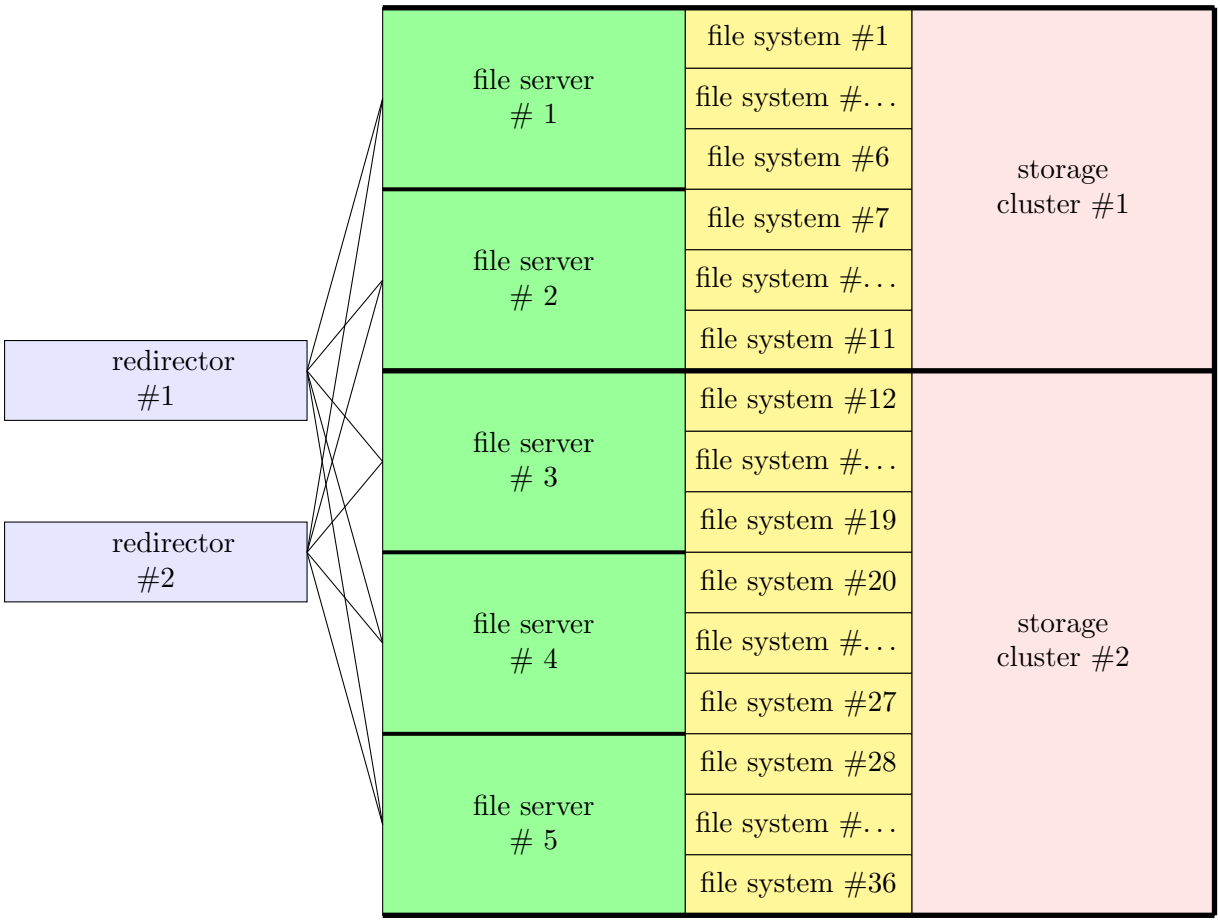

Figure 2: This schematic drawing describes the set up of the file systems and servers for xrootd at GridKa. Each storage cluster is partitioned into several file systems, while each file system is exported to xrootd only on one file server. Both redirectors are connected to all file servers.

2.4. ALICE on-site representation

In 2010, full-time positions for experiment representatives for ALICE, ATLAS, and CMS were established at the Tier-1[10]. These representatives are fully integrated into GridKa staff and their experiments. Their major tasks are communication between the experiment and the site admins as well as service administration. They serve as principal point of contact for the site admins for all matters related to the VO, ranging from job behavior to storage system setup, and communicate with all parties within the experiment, e.g. computing experts at CERN, other grid sites and users. The experiment representatives at GridKa are in close contact with each other, often analyzing problems together, allowing them to quickly assess the impact on the experiments and the Tier-1 center.

The ALICE experiment representative's regular duties include xrootd administration, participation in all regular GridKa meetings, the monthly ALICE grid phone meeting and participation in the monthly grid computing meeting at GSI (the biggest German ALICE group and a major ALICE Tier-2 site are situated there) as well as contribution to the organization of the international grid summer school "GridKa School" [11] at KIT. In early 2012, the ALICE experiment representative headed the local organization of the second worldwide ALICE Tier-1/2 Workshop [12].

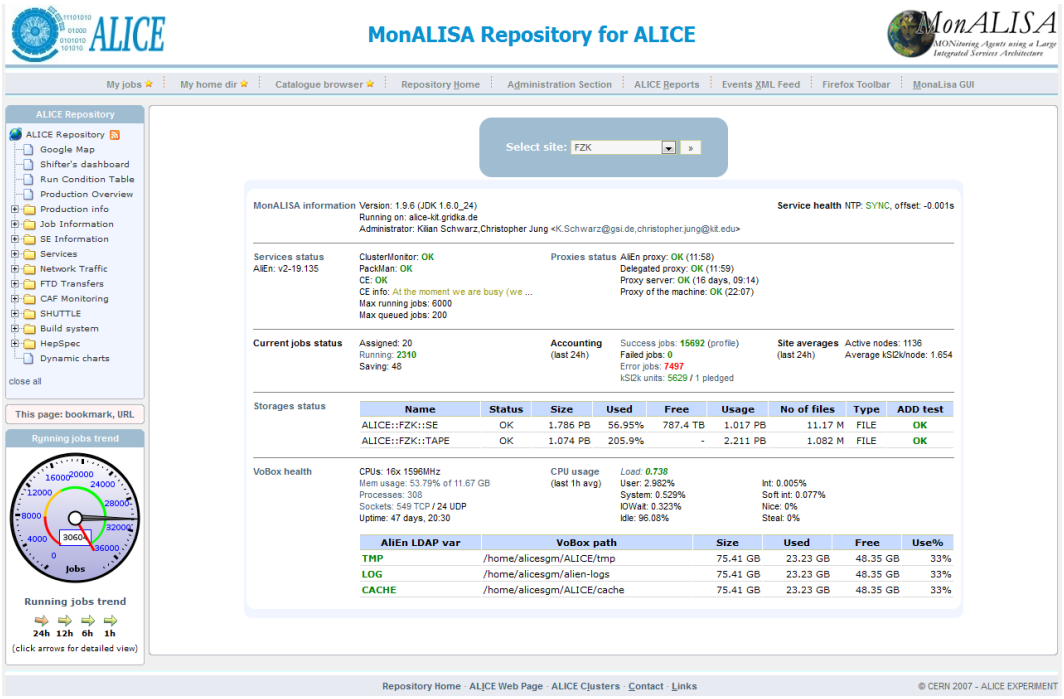

Figure 3: The MonALISA site summary for GridKa gives an overview on the status of AliEn services, current job status, storage instances, and VO box.

2.5. Experience

The computing jobs differ in behavior, mainly depending on if they are production or user analysis jobs. Production and central data analysis jobs usually have a high CPU efficiency (87% and 75%, respectively, for November 2011 to April 2012 at GridKa [13]) and can produce large amounts of data output. The overall CPU efficiency is clearly lower (54% [13]), as user analysis jobs run chaotically, may contain ineffective or wrong user code, and may actually use only little of the data read from the Storage Element. The new LEGO train framework [14] for analysis trains will combine many tasks of user jobs, will test user code before being sent on the grid, and will thereby improve the overall CPU efficiency.

Storage is the most challenging service at the moment. This is due to the large amount of data that production and user jobs want to access. Disk space is usually nearly fully used. This results in newly installed disk space attracting all new data, which causes jobs analyzing these data to access the corresponding disks all at the same time, thus slowing down the system. Also, there is an imbalance in the use of ALICE::FZK:SE and ALICE::FZK:TAPE. The former is being used constantly, while the latter is used during reproduction campaigns and for archival with few exceptions. Therefore, more than 600 TB of disk space were moved from ALICE::FZK:TAPE to ALICE::FZK:SE in early 2012. When the disk-space in ALICE::FZK:SE is completely used, it can have a bothersome effect: jobs writing their output to other grid sites, causing a very high load on the WN clusters firewall and affecting other VOs' jobs and transfers.

An ALICE experiment representative being at site has paid off. The communication with ALICE computing experts at CERN and with other ALICE site admins has been expanded. Being a member of both the GridKa administration team and the ALICE collaboration, the inside knowledge allows to proactively contribute to many topics in order to obviate disturbances. The close collaboration with the other experiment representatives at GridKa has brought profit to all three VOs and GridKa, as VOs are often affected by the same issues.

3. Summary and Outlook

ALICE Grid Computing at the GridKa Tier-1 Center has been running highly successfully for several years. This is mainly due to stable AliEn services, to the smooth operation of the Tier-1 by its local operators and to good collaboration between CERN experts and GridKa operators. The latter has been facilitated by the installation of an on-site ALICE representative.

The main goal for the future is the more efficient use of resources. The CPU efficiency of the user jobs can be improved, e.g. by employing the LEGO train framework and by training users on effective programming and use of grid resources. Better distribution of data on the disk-only Storage Element would improve the overall data throughput; the use of central tools and xrootd mechanisms is preferred to setting up local tools. Using prestaging from tape in production could contribute to more effective use of the tape-backed Storage Element.

Besides the technical aspects, the communication and collaboration between the major ALICE computing centers can be intensified. The second ALICE Tier-1/2 workshop concluded several options for synergy effects, despite the centers differing in their setups. GridKa aims to strengthen its collaboration with ALICE computing centers in Germany and its neighboring countries as well as with the other Tier-1 centers.

References

[1] Worldwide LHC Computing Grid Memorandum of Understanding URL `http://lcg.web.cern.ch/lcg/mou.htm`

[2] gLite - Lightweight Middleware for Grid Computing URL `http://glite.cern.ch`

[3] PBS Works - Enabling On-Demand Computing URL `http://www.pbsworks.com/`

[4] Homepage of XRootD URL `http://xrootd.slac.stanford.edu`

[5] IBM General Parallel File System URL `http://www-03.ibm.com/systems/software/gpfs`

[6] IBM Enterprise Removable Media Manager URL `http://www-935.ibm.com/services/de/igs/pdf/br-stor-enterprise-remove-mm-en.pdf`

[7] ICINGA homepage URL `http://www.icinga.org/`

[8] Ganglia Monitoring System URL `http://ganglia.sourceforge.net`

[9] MonALISA Repository for ALICE URL `http://www.alimonitor.cern.ch/map.jsp`

[10] Jung C, Petzold A and Zvada M Experiment representation at the WLCG Tier-1 center GridKa to be published in Proceedings of EGI Community Forum 2012

[11] GridKa School homepage URL `http://www.kit.edu/gridka-school`

[12] Second ALICE Tier-1/2 Workshop homepage URL `http://www.kit.edu/gridka-school`

[13] Betev L 2012 private communication

[14] Grigoras C, Gheata A and Grosse-Oetringhaus J F Analysis Trains: The LEGO Framework URL `http://indico.cern.ch/getFile.py/access?contribId=45&sessionId=5&resId=1&materialId=slides&confId=138478`