Towards sustainability: An interoperability outline for a Regional ARC based infrastructure in the WLCG and EGEE infrastructures

Towards Sustainability: An Interoperability Outline for a Regional ARC based infrastructure in the WLCG and EGEE infrastructures

L Field[1], M Gronager[2], D Johansson[2] and J Kleist[2]

[1]European Organization for Nuclear Research, CERN, CH-1211, Geneve 23, Switzerland
[2]Nordic DataGrid Facility, NORDUnet A/S, Kastruplundsgade 22, 1st, 2770 Kastrup, Denmark

E-mail: gronager@ndgf.org

Abstract. Interoperability of grid infrastructures is becoming increasingly important in the emergence of large scale grid infrastructures based on national and regional initiatives. To achieve interoperability of grid infrastructures adaptions and bridging of many different systems and services needs to be tackled. A grid infrastructure offers services for authentication, authorization, accounting, monitoring, operation besides from the services for handling and data and computations. This paper presents an outline of the work done to integrate the Nordic Tier-1 and 2s, which for the compute part is based on the ARC middleware, into the WLCG grid infrastructure co-operated by the EGEE project. Especially, a throughout description of integration of the compute services is presented.

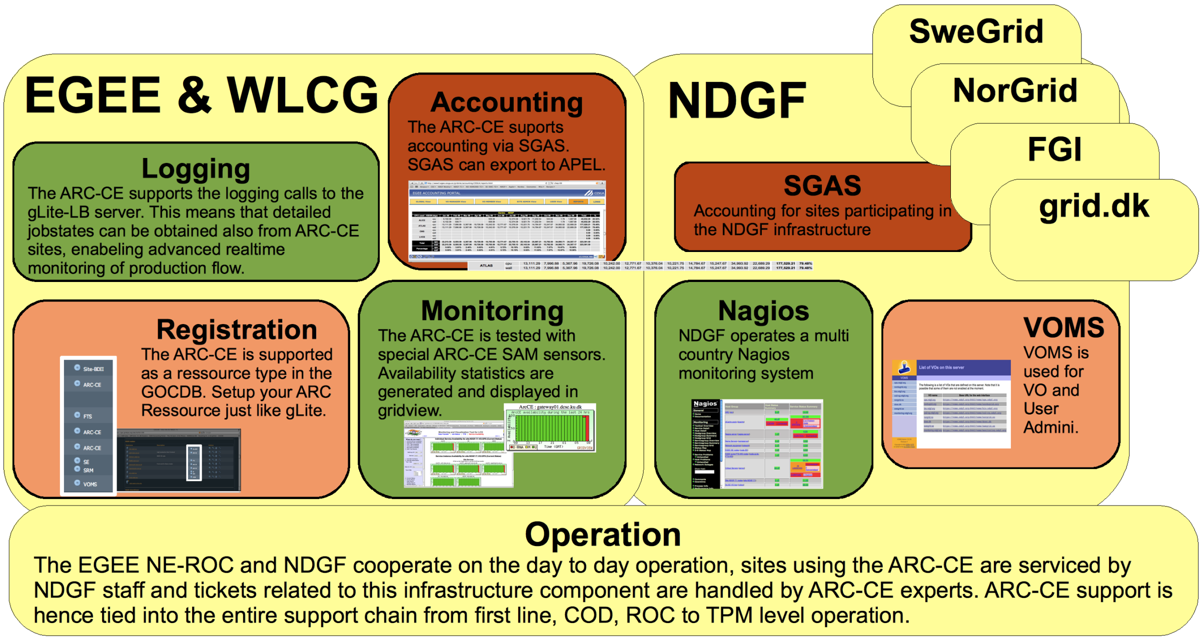
1. Introduction

The EGEE[2] projects (I, II and III) have successfully created a common European Grid Infrastructure and together with the WLCG[1] project the fabric for running the CERN experiment computations. This includes registration of services, indexing, monitoring, accounting and a well functional pan European operation team. The goal of the EGEE-III project is to ensure that a graduate transition towards a more sustainable model for operating a European Grid Infrastructure, a model that builds on the various national and regional funded grid infrastructures, and hence ties these together to a pan European grid.

In many of the bigger European countries national grid initiatives exists, like NGS[3] in the UK and D-Grid[4] in Germany, however, for the smaller countries regional collaborations have shown quite successful like e.g. the south east European SEE-GRID[5] and the Nordic NDGF[6].

This paper will focus on how the Nordic grid infrastructure operated by NDGF is a role model for how to integrate a regional or national grid into the European Infrastructure Fabric. The Infrastructure operated by NDGF is today fully interoperable with the WLCG and the EGEE Infrastructure. All services are covered and sites and users see a seamless integration of the two grids. This has happened even though the Nordic infrastructure uses the ARC[7] middleware and is not formally an EGEE partner.

The main motivation for working on integration of the Nordic infrastructure with EGEE is a wish to enable Nordic e-Science users to participate in European research collaborations, but most urgently the WLCG Tier-1 and Tier-2s in the Nordic countries. As the NDGF infrastructure from the outset has both operationally and technically been different from the one employed by EGEE, NDGF have made changes to accommodate a tighter integration with the EGEE infrastructure. In the following we describe the operational and technical setup that now allows Nordic researchers familiar with the NDGF infrastructure to enter into European collaborations, and European researchers to use the Nordic infrastructure provided by their Nordic research colleagues.
The paper is organized with two main sections covering the operational and technical setup that has been mended in order for the two infrastructures to interoperate. We start out by the most human relation intensive effort, namely the operation.



The ARC-CE suports accounting via SGAS. SGAS can export to APEL.

The ARC-CE supports the logging calls to the gLite-LB server. This means that detailed jobstates can be obtained also from ARC-CE sites, enabeling advanced realtime monitoring of production flow.

Accounting for sites participating in the NDGF infrastructure

The ARC-CE is supported as a ressource type in the GOCDB. Setup your ARC Ressource just like gLite.

The ARC-CE is tested with special ARC-CE SAM sensors. Availability statistics are generated and displayed in gridview.

NDGF operates a multi country Nagios monitoring system

VOMS is used for VO and User Admini.

The EGEE NE-ROC and NDGF cooperate on the day to day operation, sites using the ARC-CE are serviced by NDGF staff and tickets related to this infrastructure component are handled by ARC-CE experts. ARC-CE support is hence tied into the entire support chain from first line, COD, ROC to TPM level operation.

Figure 1. The operational setup with the bridged services for WLCG, EGEE and NDGF. A more throughout explanation of the different services can be found in the following sections.

2. Operation

In order for two infrastructures to interoperate it is extremely important to have a high degree of communication between the operation teams. The NDGF Operation team have hence participated in the weekly WLCG-EGEE-OSG operation meetings[8] the last couple of years and have the last year taken part in the rotating CIC on Duty[1], further, integration between the NDGF operation team and the Northern part of the EGEE NE-ROC was initiated in October 2008 and today the two teams are fully

[1] CIC-On-Duty is the rotating Core-Infrastructure-Center on Duty in the EGEE Operation model. The COD takes care of monitoring alarms at sites and issuing tickets.

integrated. This means that Nordic Grid operation is handled every second week by NDGF and every second week by SNIC[9], which constitutes the Nordic part of the NE-ROC. The joint Nordic Operation team hence maintains Nordic ARC as well as Nordic and Baltic gLite[10] resources. This joint operation setup ensures that competences on operation of ARC as well as gLite are shared by a large number of people and also it paves the way for sustainable future Nordic Grid Operation as only half of the operation team are dependent of EU funding.

The operation of the ARC-CEs[2] is now also shared between the SNIC and NDGF operation teams, which hence responds to GGUS[11] tickets as a virtual site as well as a ROC, ensuring that operational informations flows between the infrastructures. Last but not least, participation in common conferences and events have been crucial for the success of the interoperation. Without good human relations and knowledge of each others organizational structures the process becomes too cumbersome.

NDGF uses the NorduGrid ARC middleware for handling computations and also contributes heavily to maintenance of the production release of the software. At the operational level this means that NDGF also takes the responsibility of ensuring that bugs reported on the production release are solved and NDGF provide support for Nordic as well as non-Nordic sites wishing to use ARC for managing computations.

3. Technical Services

Technically a number of adjustments to the NDGF infrastructure have been needed to achieve interoperability. In this section the road towards interoperation will be described service by service.

3.1. Registration of Services

The first step towards interoperation is to enable registration of new services provided in the regional infrastructure, and to do this with a proper mapping to the already existing services in EGEE. EGEE registers services in the GOCDB[12] (Grid Operation Center Database), this gives a single point of entry to what services are running where and who to contact on a management and technical level as well as in case of security issues. For the NDGF case a virtual site was created (The Tier-1 site: NDGF-T1). A lot of services was already in common; storage elements (dCache[13]), catalog service (LFC[14]), VOMS[15], etc. However, the ARC method dynamic service indexing, the ARC-GIIS[7], as well as a different compute element, the ARC-CE was not supported in the GOCDB. A new Compute Element type, ARC-CE, was created in the GOCDB and the NDGF CEs registered.

3.2. Indexing of Services

The Globus Meta Data Service[16] consisting of top level GIIS and site level GRIS[7] services are not supported by any service in the EGEE infrastructure and hence it was decided to setup a special BDII[17] for NDGF, dynamically reading the contend of the GRIS'es on the ARC-CEs based on the list of CEs provided by the GIIS'es. The service was setup quite early, and it enabled visualization of the resources in the different grids and as the ARC-CEs were now visible for the EGEE services they could start to interact with them. The BDII concept have recently been adapted by the ARC-CE and hence the ARC-CE as of release 0.8 also supports rendering directly to the GLUE[18] schema and setting up a special site BDII is hence no longer needed.

[2] The ARC-CE will be explained in a succeeding section.

3.3. Monitoring

The first EGEE Infrastructure fabric service to interact with the NDGF services was the Service Availability Monitoring, SAM[19]. SAM executes tests every 3rd hour towards the different sites registered and marked for monitoring in the GOCDB, by querying the individual services listed in the site BDII. For the NDGF case, SAM tests for index, storage, catalogue could run right from the beginning, whether the ARC-CE node type was not supported by SAM. SAM is build modular to enable testing of different services, and hence a new sensor suite for the ARC-CE was developed. The mapping of the different tests was brought up on WLCG-MB level and a working group sat down to review the tests ensuring a fair translation between the tests for the different CE types.

3.4. Accounting

The next step in integration of the NDGF infrastructure into the EGEE, was to gather and export accounting from NDGF pr VO to the EGEE Accounting Portal[20]. The EGEE Accounting Portal uses the APEL database as back-end, and direct DB insertion is provided pr site. The NDGF infrastructure is accounted using SGAS[21] (SweGrid Accounting System) and an automatic script for exporting the accounting info gathered in SGAS to APEL[22] was setup. The is an added value of this approach; Many national states does not allow for accounting info on the level per person to be exported outside the country borders. Hence a federated approach only submitting accounting info on a relevant level will be required for many states.

3.5. Job Submission

Cross grid job submission is not strictly needed for infrastructure interoperation. The user might very well opt for having several user interfaces installed, and many of the large WLCG VOs chooses this. However, cross grid job submission does indeed promote a more seamless integration between the infrastructures.

The ARC-CE is different from the gLite-CE and gLite-CREAM-CEs various ways. In principle it originates from a quite different approach to grids in the Nordic countries than in the rest of Europe; The grid should not be a separate infrastructure from the HPC infrastructure. This means that grid jobs could and should also be used a low priority back-filling jobs on large super computer installations. Further, there should be no reason to have all massively parallel computers also accessible via the grid. However, if a resource can also be accessed via grid middleware there is a need to be able to protect the resource from unwanted use patterns. All these requirements have been the design criteria for the ARC-CE.

The workflow conducted on the Compute Element should hence resemble the work done by a well behaved super computer user, just with a lot of the steps automated. A well behaved user would on his frontend do the following as part of data analysis task:

- Compile and optimize the source
- Installation of the optimized code on the resource
- Handling of job data, download / upload remote files – put it on the shared cluster filesystem
- Running the job

Further, as it is an automated resource other workflow optimizations becomes easy to include:

- Caching of data
- Intelligent transfer retries
- Configurable resource throttling

Finally, the Compute Element offers a unified interface for the user for great variety of batch systems. Currently ARC supports PBS, Torque, LSF, LoadLeveler, EASY, SLURM, SGE, Condor and plain fork[7].

By automating, and being able to limit the resource usage to avoid cluster congestion the ARC-CE uses the resource in a highly efficient way. E.g. the data handling service in the ARC-CE, see fig. 2, with pre loading of job data to the shared cluster file system, the caching of often used data files and automatic upload of data once the job has for the ATLAS experiment showed 10-15% better resource utilization as compared to having the jobs fetching the data them self keeping the CPU idling and making the job much more vulnerable towards transfer errors and service glitches.

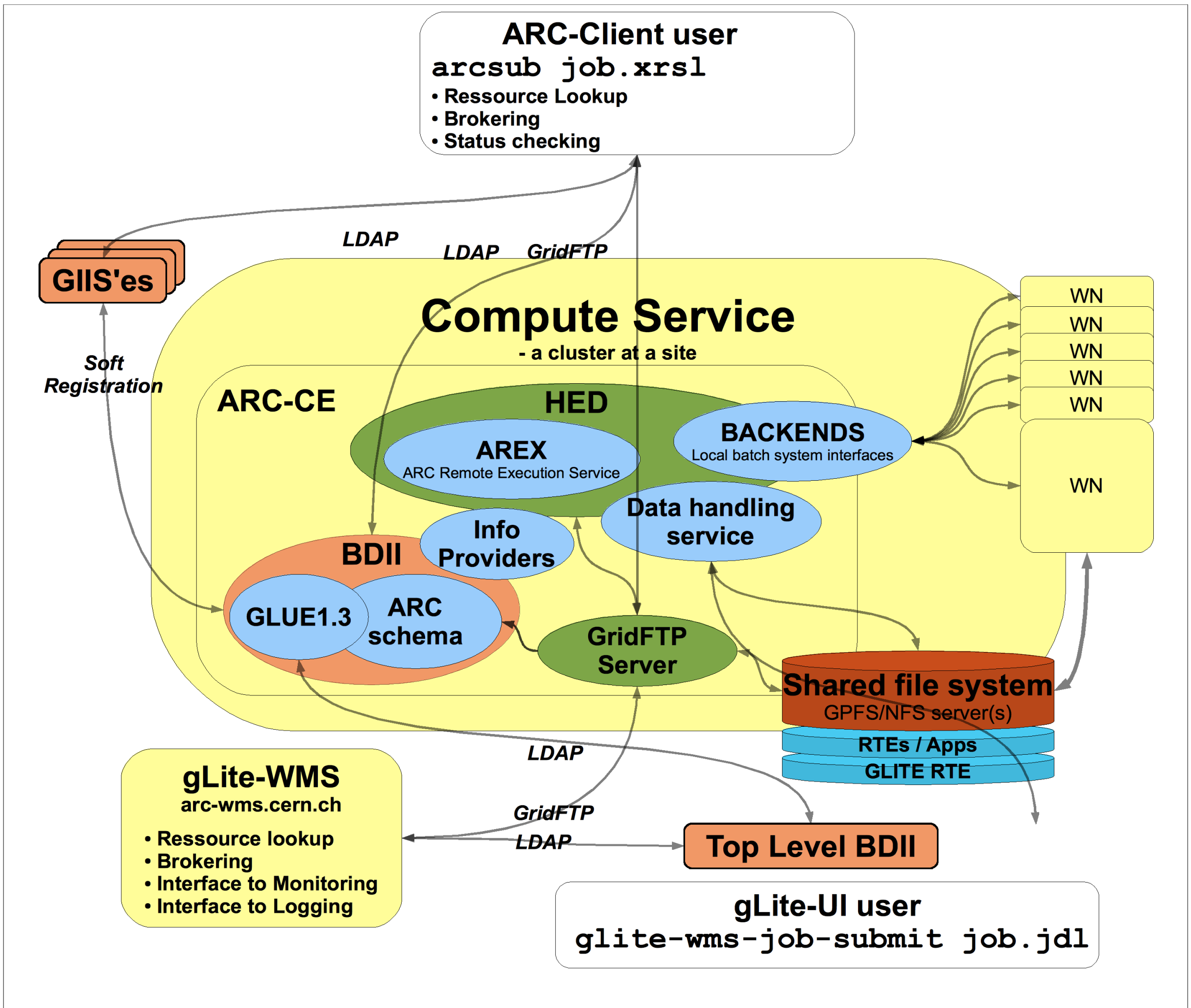


Figure 2. The ARC-CE and the interface to users and resources.

The Execution Service manages the job workflow and ensures that jobs are not started before all input files are present, further, it throttles the number of concurrent up and downloading streams to optimize the network usage and it also manages a transfer retry policy, making the resource robust towards failing storage elements. By separating data handling from job execution the resource can keep a queue of ready jobs, and the cluster can effectively work for several days completely disconnected from the network.

The information system in ARC has traditionally been based on Globus MDS, and the arc-schema for representing cluster information. However, ARC has now moved to use the BDII for information publishing and several redundant standard LDAP servers for resource discovery. For the representing cluster information the ARC-CE now also supports direct publishing of GLUE, making a drop in replacement of a gLite flavor CE with an ARC-CE very simple.

The ARC-CE are accessed directly by the ARC client and not as for gLite where an intermediate resource broker is used (gLite-WMS[23]). The ARC client hence does the brokering. However, for a user to submit a job to the NDGF infrastructure based on ARC-CEs support for submission from the gLite-WMS either directly or through some kind of gateway was needed. Various schemes have been explored for this, and the current setup is based on direct submission from the gLite-WMS to the ARC-CEs. This work has been conducted within SA3 in EGEE-II and reported earlier[24]. Today, this is part of the standard gLite-WMS and used by e.g. the CMS experiment for running the production[25].

The latest version of the ARC client also supports submission directly to the gLite-CREAM-CE hence making cross grid submission possible also in the ARC->gLite direction.

3.6. Data Management

The storage infrastructure used by NDGF is based on dCache. As with compute resources the storage infrastructure is also distributed, this is realized by having dCache storage pools at the participating centers and with only the central dCache services placed at the Nordic DataGrid Facility. NDGF together with the dCache core partners contributed to dCache with features that makes the operation of dCache in such a distributed setting efficient and from the outside - seamless. For most of the time sites operate their storage pools autonomously, but if a site need to bring down a pool for a longer period of time, NDGF operators can instantiate the replication of the data stored at the pool. The choice of what data is stored at specific pools are also under the governance of the owner of the storage pools and only data belonging to an approved VO is stored. From time to time coordinated upgrade of the dCache installations is needed, this is scheduled in collaboration with the participating centers.

3.7. Logging

The final service added was detailed logging. The logging information in the EGEE grid is mainly used for detailed monitoring of jobs like the Real Time Monitor[26], and for debugging scenarios. Instead of installing the logging clients on all the ARC-CEs a hook in ARC for directly exporting the detailed job states to the gLite Logging and Bookkeeping server was chosen.

4. Conclusions

This paper has presented a list of services that are all part of a modern grid infrastructure. All these services need to be bridged in order for two infrastructures to become fully interoperable. The purpose of this paper has been to provide an outline and an example for how this can be done for other infrastructures as well, but most importantly an outline of the interoperation work between the NDGF and the WLCG and EGEE infrastructures has been provided.

The paper has also outlined to function of the ARC-CE as compared to the gLite-flavor CEs and explained how the ARC-CE can utilize existing shared supercomputer resources as grid resources in a non invasive and protective way, not jeopardizing the local use.

Finally, it is suggested that site with larger shared resources becomes part of the European wide grid using the ARC-CE and the interoperation work as described in this paper.

References

[1] J.Shiers, *The Worldwide LHC Computing Grid (worldwide LCG),* Computer Physics Communications, Volume 177, Issues 1-2, July 2007, Pages 219-223

[2] E.Laure, B.Jones, *Enabling Grids for e-Science: The EGEE Project*, EGEE-PUB-2009-1, http://www.eu-egee.org/

[3] National Grid Services, http://www.grid-support.ac.uk/

[4] W.Gentsch, *D-Grid, an E-Science Framework for German Scientists,* Proceedings of The Fifth International Symposium on Parallel and Distributed Computing, 2006. ISPDC '06, http://www.d-grid.de/

[5] South East European GRid-enabled eInfrastructure Development, http://www.see-grid.org/

[6] Nordic DataGrid Facility, http://www.ndgf.org/

[7] M.Ellert et al., *Advanced Resource Connector middleware for lightweight computational Grids,* Future Generation Computer Systems **23** (2007) 219-240.

[8] EGEE Operations meeting, http://indico.cern.ch/categoryDisplay.py?categId=258

[9] Swedish National Infrastructure for Computing, http://www.snic.vr.se/

[10] gLite – Lightweight Middleware for Grid Computing, http://glite.web.cern.ch/glite/

[11] Global Grid User Support, http://www.ggus.org/

[12] GOCDB, https://goc.gridops.org/

[13] P.Fuhrmann and G.Volker, *dCache, Storage System for the Future,* Proceedings of 12th International Euro-Par Conference, Lecture Notes in Computer Science, Volume 4128, Springer Verlag, http://www.dcache.org/

[14] J. P. Baud, J. Casey, S. Lemaitrel, and C. Nicholson, *Performance analysis of a file catalog for the LHC computing grid*, Proc. IEEE 14th Int. Symp. High Performance Distributed Computing, HPDC-14, 2005, pp. 91–99,

[15] R.Alferi et.al., VOMS, *an Authorization System for Virtual Organizations,* Lecture Notes in Computer Science, Volume 2970, Springer Verlag, http://hep-project-grid-scg.web.cern.ch/hep-project-grid-scg/voms.html

[16] C.Kesselman et.al., *Grid Information Services for Distributed Resource Sharing,* Proc. of 10th IEEE International Symposium on High Performance Distributed Computing, 2001, http://www.globus.org/toolkit/mds/

[17] Field L and Schultz M W, *Scalability and Performance Analysis of the EGEE Information System,* Proc. of CHEP 2004, Journal of Physics: Conference Series **119** (2008)

[18] http://infnforge.cnaf.infn.it/glueinfomodel/

[19] GridView - Monitoring and Visualization Tool for LCG, http://gridview.cern.ch/

[20] EGEE Accounting Portal, http://www3.egee.cesga.es/

[21] E. Elmroth, F. Galán, D. Henriksson and D. Perales. *Accounting and Billing for Federated Cloud Infrastructures.* In Juan. E. Guerrero (ed), Proceedings of the Eighth International Conference on Grid and Cooperative Computing (GCC 2009), IEEE Computer Society Press, pp. 268 - 275, 2009., http://www.sgas.se

[22] Byrom R et al. *APEL: An implementation of Grid accounting using R-GMA,* http://www.gridpp.ac.uk/abstracts/allhands2005/apel.pdf

[23] http://glite.web.cern.ch/glite/packages/R3.0/deployment/glite-WMS/glite-WMS.asp

[24] Grønager M et al, *Interoperability between ARC and gLite - Understanding the Grid-Job Life Cycle*, pp.493-500, 2008 Fourth IEEE International Conference on eScience, 2008

[25] Linden T et al, *Grid Interoperation with ARC middleware for CMS experiment,* to be published

[26] GridPP Real Time Monitor, http://gridportal.hep.ph.ic.ac.uk/rtm/