Integrating a heterogeneous and shared computer cluster into grids

V. Büge[1,2], U. Felzmann[1], C. Jung[1,2], U. Kerzel[1], M. Kreps[1], G. Quast[1], A. Vest[1]
[1] Institut für Experimentelle Kernphysik, University of Karlsruhe, Germany
[2] Institut für Wissenschaftliches Rechnen, Forschungszentrum Karlsruhe, Germany

Abstract

Integrating existing computer clusters at universities into grids is quite a challenge, because these clusters are usually shared among many groups. As an example, the Linux cluster at the "Institut für Experimentelle Kernphysik" (IEKP), University of Karlsruhe, is shared between working groups of the high energy physics experiments AMS, CDF and CMS, and has successfully been integrated into the SAMGrid for CDF and the LHC computing grid LCG for CMS while it still supports local users. This shared usage of the cluster effects heterogeneous software environments, grid middleware and access policies. Within the LCG, the IEKP site realises the concept of a Tier-2/3 prototype centre. The installation procedure and setup of the LCG middleware has been modified according to the local conditions. With this dedicated configuration, the IEKP site offers the full grid functionality such as data transfers, CMS software installation and grid based physics analyses. The need for prioritisation of certain user groups has been temporarily satisfied by supporting different Virtual Organisations within the same experiment. The virtualisation of the LCG components, which can improve the utilisation of resources and security aspects, will be implemented in the near future.

WHY DO UNIVERSITY GROUPS NEED GRID COMPUTING?

Current and future high energy physics experiments at Tevatron or the LHC need to deal with huge data production rates and event sizes. Therefore, the demand for computing power and mass storage has significantly increased. For example, the CMS detector has a final recording rate of 225 MB per second [1] which has to be stored online for later processing and data reconstruction. Corresponding Monte Carlo simulations have to be generated and stored as well. Currently, there are already:

- Simulated data in the LHC experiments, $\mathcal{O}$(100 TB)
- Real data in the HEP experiments CDF, D0, H1, ZEUS, BaBar, etc., $\mathcal{O}$(1 PB)

Processing power is widely available in the associated institutes, but the worldwide distributed datasets for analyses are only accessible using grid technologies. Therefore, the participating groups – particularly at universities – cope with these challenges using grid tools. This leads to an opportunistic or shared use of the resources between local users and grid users in the collaborating groups.

The main benefits of integrating an institute's cluster into grids are:

- minimisation of idle times
- interception of peak loads, e.g. before conferences
- shared data storage
- shared deployment effort of common services

COMPUTING ENVIRONMENTS AT UNIVERSITIES

In this section, the peculiarities of a typical university cluster are described using the representative example of the Linux cluster at the Institut für Experimentelle Kernphysik (IEKP) at the University of Karlsruhe.

In general, a typical university cluster has to cope with diverse challenges:

- A computer cluster at an institute usually has a heterogeneous structure in hardware, software, funding and ownership.
- Furthermore, it has to support multiple groups with different applications and sometimes conflicting interests.
- Typically, the cluster's infrastructural facilities have grown in the course of time. Thus, the cluster has developed its own characteristical history and resulting inhomogeneities.
- Moreover, a cluster is embedded in structures imposed by institute, faculty and university.

For these reasons, the integration of a university cluster into existing grids is not easy at all. Moreover, the idea of sharing resources is still not present in all minds.

Representative example: The IEKP at Karlsruhe

As an example, the IEKP Linux cluster, called "EKPplus", at Karlsruhe has successfully been integrated in the Sequential Access via Meta-data Grid (SAMGrid) [2] for CDF and the LHC Computing Grid (LCG) [3] for CMS, while it still supports local users. This shared usage of the cluster leads to heterogeneous environments concerning

- software and hardware
- local and grid users
- access policies
- grid middleware

A detailed description of the integration of the IEKP Linux cluster into the LHC Computing Grid can be found in [4] and references therein.

The specification of the IEKP Linux cluster components which is representative for many other institutes is shortly described in the following:

- There are one or more portal machines for each experiment (3 for CDF, 1 for CMS and 1 for AMS).
- Five file servers provide a disk space of about 15 – 20 TB in total.
- The local batch system consists of 27 computing nodes with a total of 36 CPU's.

The EKPplus cluster is independent of the desktop cluster. Figure 1 depicts the overall architecture of the EKPplus cluster and the integration of the SAMGrid and LCG components.

Peculiarities of a typical university cluster

The main issues to be considered when setting up and running an institute's cluster are described in the following, where the IEKP cluster is again taken as instance.

Network architecture The inner network of the cluster hosts the computing nodes, several file servers and a dedicated cluster control machine. This control machine takes care about local users, manages the job queues for the batch system and provides the root file system for the computing nodes.

The outer network consists of publicly accessible portals which serve as testbeds for the development of analysis software. Via multiple Ethernet cards, the portals are also connected to the inner private network. Thus, they offer access to the file servers and the usage of the worker nodes (WNs) via the local batch system.

Network protocols and services User accounts are exported by the cluster control machine to all nodes via the Network Information Service (NIS). File and root systems are exported via the Network File System (NFS) services. Furthermore, the protocols GSIFTP and SRM [5] are supported.

Local batch system The worker nodes are controlled by the Open Source PBS/Torque [6] batch system on the cluster control machine. The scheduler used to send a job to the next free worker node is the flexible system MAUI [7]. This system supports the fair share principle and is able to manage both, group and user fair share.

In addition to the local batch queues, one grid queue according to each Virtual Organisation (VO) supported by the LCG site is configured. These queues are dedicated to jobs submitted via the grid and have the same names as the respective VOs.

Firewall The grid components are placed behind the firewall of the institute. To allow external access to the services run by LCG, some ports of the firewall have to be opened to these dedicated hosts. The internal campus net is in general protected by the University's computing department and is switched off for the IEKP cluster net.

Desktop cluster The desktop cluster comprises user workstations which can be used as access points to the portal machines. The EKPplus cluster is connected to the IEKP desktop network by a 1 GBit connection.

Operating systems The software on the portal machines and worker nodes is experiment dependent. The operating system on all machines is Linux but the flavour is not identical on all components due to experiment specific extensions and modifications. At the IEKP, the following operating systems are used:

- The CDF portals use a Linux distribution based on Fermi RedHat 7.3.
- On the CMS portal, the operating system Scientific Linux CERN (SLC) 3 is used.
- The underlying operating system for the cluster control machine is Scientific Linux CERN 3, as well as for the LCG hosts.
- On the worker nodes, a slightly modified release of Scientific Linux CERN 3 is used since it is the only operating system under which all AMS, CDF, CMS software and grid software runs at the moment. Besides, the worker nodes run a 32-bit operating system; an upgrade to a 64-bit operating system is foreseen in the near future.

It is worth mentioning that no major problems occurred running different Linux distributions on the same cluster.

GRID MIDDLEWARE REQUIREMENTS AND SITE SPECIFIC GRID SERVICES

Grid middleware requirements

Integrating a computer cluster into grids requires that the grid middleware is able to adapt to the needs of the grid site, which are:

- Flexibility:
  The installation procedure and setup of the grid middleware should be modifiable according to local conditions.
- Interoperability:
  The grid middleware should be compatible with other grid middlewares.
- Dynamic:
  It must be possible to add or remove resources during the running grid service.
- Encapsulation:
  Experiment and analysis software must be shielded from changes in the underlying grid environment.
- Level of abstraction:
  The access to computing and storage resources must be independent of their physical location and local setup.

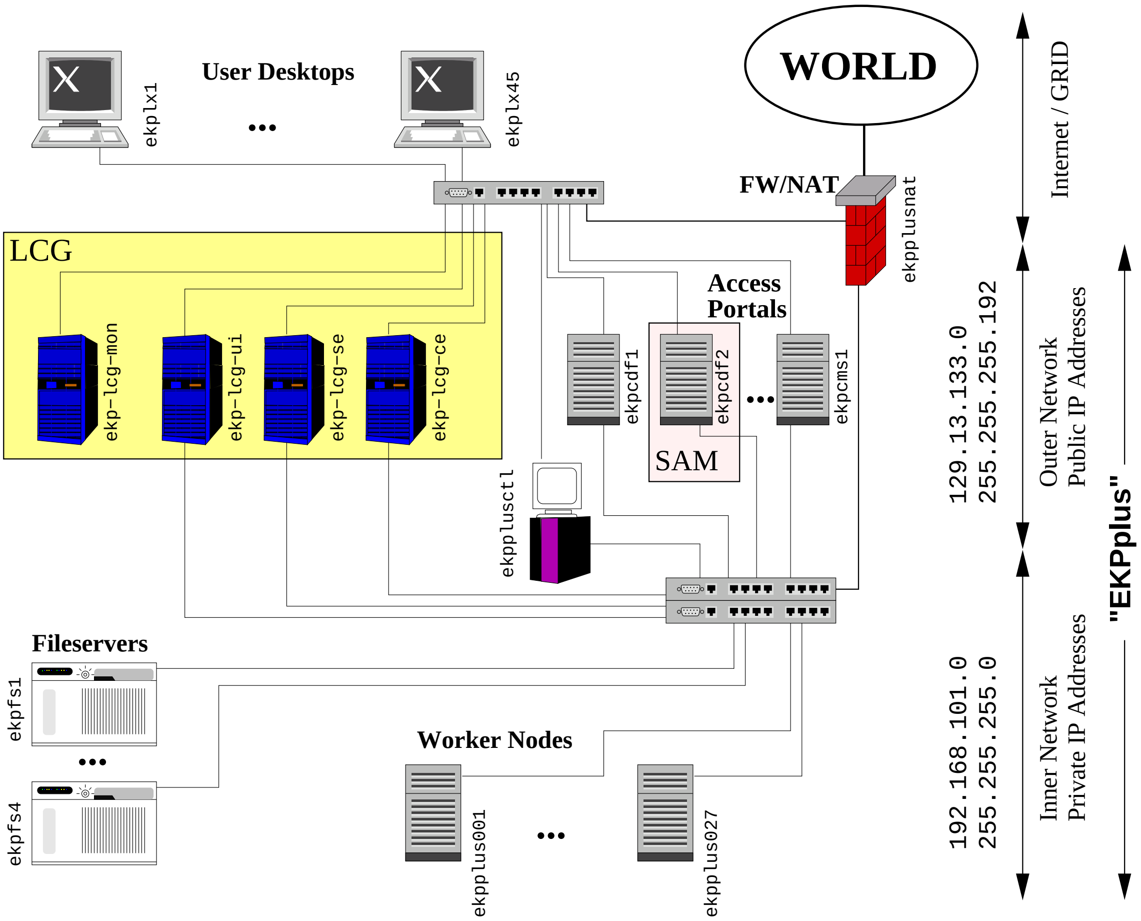

Figure 1: A schematic overview of the architecture of the IEKP Linux cluster “EKPplus” and the integration of the SAMGrid and LCG components.

Site specific grid services

Site specific grid services are services each site has to provide in order to offer computing power and disk space to a Virtual Organisation. Since the IEKP Linux cluster is integrated into SAMGrid and LCG, the mandatory grid services for these grids have been implemented in the cluster. In detail, these services are:

SAMGrid Only one dedicated machine per cluster, the SAMGrid station, is needed in order to integrate the cluster into the SAMGrid. At the IEKP, the SAMGrid station is one of the CDF portals, which offers user interaction, mass storage and file import, export and delivery to analysis programs. The file exchange runs via the GridFTP protocol, and all activity is written to a central database. A detailed overview of SAMGrid can be found [8].

LCG The four obligatory LCG site specific services and their interactions are listed below.

- The Computing Element (CE) is the gateway to the local computing resources, i.e. to the worker nodes via the local batch system. A Globus gatekeeper running on the CE controls the access to the local resources. The CE offers a layer of abstraction, i.e. the peculiarities of the local computing resources are irrelevant to the user since he/she does not have to deal with different batch systems on different grid sites.
- The Storage Element (SE) is the gateway to local storage (disk and/or tape). Like the CE, it can be a gateway to the local storage system, but it also can offer disk space itself. Since the SE is a Globus Gridftp service, it supports by default the gsiftp protocol. Other protocols like SRM can be added by the local site manager.
- The Monitoring Box (MON) publishes information about the functionality of all affiliated grid site nodes.
- The User Interface (UI) is the user’s access point to the grid. Several client programs are installed which enables the use of grid services. The User Interface is also the portal for (grid) file access.

THE IEKP LCG SITE

Within the LCG, the IEKP site realises the concept of a Tier-2/3 prototype centre. With its dedicated configuration, it offers the full grid functionality such as

- grid based physics analyses;
- software installation for grid and local users;
- data storage.

Since there is no Virtual Organisation Membership Service (VOMS) yet, the need for a prioritisation of certain user groups is satisfied by supporting different Virtual Organisations. In the present workaround, the treatment of grid users depending on their affiliation is organised by mapping them to different accounts and user groups and by the configuration of the corresponding queues they can use. According to their affiliation, the grid users can use the IEKP capacities with different queue priorities which are managed by fair share targets.

- Local CMS users are mapped to their local accounts and are members of the local user group cms and the user groups dcms and cmsgrid.
- German CMS users are mapped to generic accounts dcms001 – dcms050 which are members of both, the user groups dcms and cmsgrid.
- Other members of the CMS collaboration are mapped to generic accounts cms001 – cms050 which are members of the user group cmsgrid only.

PRESENT CMS GRID STRUCTURE IN GERMANY

DCMS is the collaboration of the German CMS institutes at Aachen, Hamburg and Karlsruhe. The main goals of DCMS are the exchange of experience in analyses (on the grid) and the sharing of the available resources. This is realised by the prioritisation of dcms users at the DCMS sites, Since January 2005, an infrastructure for the VO dcms is setup at DESY. This VO comprises and supports all CMS members registered at the German grid sites. Thus, all members of the VO dcms are also members of the VO cms. The VO dcms will hopefully become obsolete with the VOMS.

CONCLUSION AND OUTLOOK

Computer clusters at universities are often shared among different working groups and are situated in a very heterogeneous environment. The IEKP cluster has successfully been integrated into the SAMGrid and the LCG without compromising local user groups. The virtualisation of the LCG components is foreseen in the near future.

A successful and beneficial integration of existing university clusters into computing grids is possible, but there is still a long wishlist. As a precondition, university groups must be motivated to share their free computing resources within grid environments. For this purpose, mechanisms for priorisation, accounting and charge-back or “billing” are desperately needed. Some workarounds for these issues have been suggested above. Moreover, the following needs of a university cluster should be fulfilled:

- Experiment specific and grid software, in particular the LCG middleware, should be independent of the operating system.
- To ease the LCG installation on existing computing environments, a lightweight and non invasive installation procedure is needed, in particular concerning the worker nodes.
- The LCG environment/hosts should be installed on a single powerful server/node. This virtualisation can improve the utilisation of resources as well as security aspects.
- The virtualisation of worker nodes may cope with significant differences of the Linux flavours required by certain user groups.

So far, the differences between computing environments, e.g. Linux flavours, were small enough to run them in parallel. Still, mechanisms for priorisation and accounting are a precondition for a fruitful integration of university clusters into computing grids.

ACKNOWLEDGEMENTS

We are grateful to the IEKP system administrators for their work in constructing and maintaining the IEKP Linux cluster, namely Yves Kemp, Michael Milnik, Christian Sander and Patrick Schemitz. Furthermore, we thank the dcms LCG administrators Andreas Gellrich, Mathias de Riese, Matthias Kirsch and Andreas Nowack for their support during the LCG middleware installation procedure. We wish to thank the German Bundesministerium für Bildung und Forschung BMBF for financial support, and we are grateful for the technical support by the Universitäts-Rechenzentrum Karlsruhe.

REFERENCES

[1] I. Bird *et al.*, “LHC computing Grid. Technical design report”, CERN-LHCC-2005-024

[2] Sequential data Access via Meta-data:
http://projects.fnal.gov/samgrid/

[3] LHC Computing Grid: http://lcg.web.cern.ch/LCG/

[4] V. Büge *et al.*, “Integrating the IEKP Linux cluster as a Tier-2/3 prototype centre into the LHC Computing Grid”, IEKP-KA/2006-3,
http://www-ekp.physik.uni-karlsruhe.de/pub/web/thesis/2006-03.pdf

[5] The Storage Resource Manager Collaboration:
http://www-isd.fnal.gov/srm/

[6] Tera-scale Open-source Resource and QUEue manager TORQUE:
http://www.clusterresources.com/products/torque/

[7] Maui Cluster Scheduler:
http://www.clusterresources.com/products/maui/

[8] U. Kerzel *et al.*, “Experiences with operating SAMGrid at the GermanGrid centre ”GridKa”, Prepared for CHEP’06: Computing in High-Energy and Nuclear Physics, Mumbai, India, 13-17 Feb 2006