Particle Physics Grid Deployment in the Czech Republic

L. Fiala[*], J. Chudoba[*], J. Kosina[*], J. Krásová[*], M. Lokajíček[*], J. Švec[*], J. Kmuníček[**], D. Kouřil[**], L. Matyska[**], M. Ruda[**], Z. Salvet[**], M. Mulač[***]

[*] *IP AS CR and CESNET*

[**] *ICS, MU Brno and CESNET*

[***] *CIV ZČU Pilsen and CESNET*

Abstract

As an extension of the national Grid project METACentrum, run by CESNET, two Grid farms are to large extent dedicated to the particle physics applications. In this article we will describe the hardware installed on both farms, how the farms are connected into the large Grid infrastructure (Grid projects LCG and EGEE) and provide an overview how the computing facility is shared between different projects, so that concurrent interoperability of all the installed software and infrastructure can be assured.

1 Overview

The project METACentrum [14] covers the majority of activities concerning Grids, super-, cluster- and Grid computing and/or high performance computing in general in the Czech Republic. The aim of the METACentrum project is the maintenance of current computational resources and expansion of available computational capacity of the largest academic centers in the Czech Republic in future. Two Grid farms – Golias and Skurut - in the Czech Republic providing their resources to the particle physics community are run as one of extensions of the METACentrum project. In addition to the main use for the particle physics applications, the resources of the Skurut farm are provided both for educational and dissemination activities in the scope of the GILDA project [16], and also for the provision of the distributed Grid facilities to the wide scientific community of Central Europe. This activity is acronymed VOCE, which stands for Virtual Organization for Central Europe[1], and besides Czech resources, also Austrian, Hungarian, Polish, Slovak and Slovenian computational resources are available through it. VOCE infrastructure was designed and implemented on resources of CESNET [2] and Institute of Physics AS CR [15] on Skurut and Golias farms, respectively. The Czech Republic provides full support for the VO (including configuration and maintenance of all needed service nodes, replica catalogues, etc.) for other participating entities.

2 Hardware Resources

2.1 Golias Farm

The configuration of the nodes belonging to the Golias [9] farm are listed in the following table.

[1] Virtual Organization is, in the Grid terminology, an abstract institution, grouping together the users and resources that these users are allowed to use, within the Grid environment.

Table 1: Golias farm hardware configuration

Hardware	Number of nodes
HP LP1000r 2x PIII 1,13 GHz, 1GB RAM, 18GB SCSI HDD	34
HP LP4100TC PIII 1,13 GHz, 2GB RAM, 40GB ATA HDD	2
HP ProLiant DL 140 2x XEON 3,06 GHz, 2GB RAM, 40GB ATA HDD	53
HP ProLiant DL 360 2x XEON 2,8 GHz, 2GB RAM, 40GB ATA HDD (RAID1)	2
HP ProLiant DL 145 2x Opteron 244, 2GB RAM, 40GB ATA HDD	3
HP ProLiant DL 140 2x XEON 3,06 GHz, 4GB RAM, 40GB ATA HDD	14
Total computing power in SpecInt2000	**150000**
Total raw shared storage capacity	**41 TB**

The nodes are divided into the different hardware groups, as they were purchased in different points of time. Disk storage is provided by three disk arrays, each of them having different size and also hardware design.

The smallest 1 TB array uses 15 SCSI disks connected together into RAID 5. This disk area is provided by a separate server through NFS to other nodes.

The second array comprises three EasySTOR 1400RP boxes, each of them containing 14 ATA disks, 250 GB each. Disks within every single box are connected together into the RAID 5. The first EasySTOR box differs from the other two - the mini-server is a part of the box. The server has Linux installed and provides the disk space through NFS (other two EasySTOR boxes are connected into this box by UltraSCSI 160 bus). The total disk size provided by this array is approximately 10TB.

The third array consists of six EasySTOR 1600RP boxes. Each box contains 16 ATA disks, 300GB each. Disks within each box are interconnected together to form RAID6, which is quite a new technology. A RAID 6 array is essentially an extension of a RAID 5 array with a second independent distributed parity scheme. Data and parity are striped on a block level across multiple array members, just like in RAID 5, and a second set of parity is calculated and written across all the drives. This schema is illustrated on the following figure.

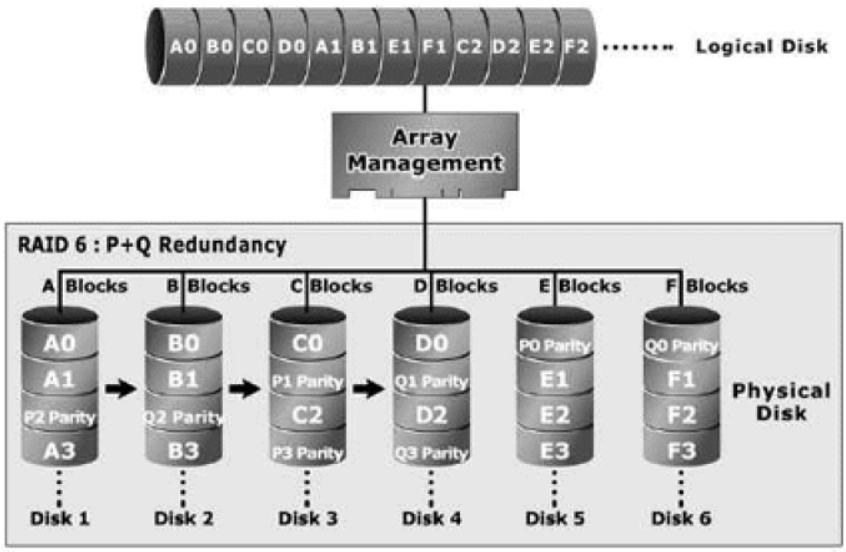

Fig. 1. RAID6 schema

All these boxes are connected via UltraSCSI 160 bus to the HP DL145 server, which provides the disk space (of total raw size approximately 30 TB) via NFS.

2.2 Skurut farm

This farm consists of 32 identical dual Intel Pentium III 700 MHz nodes with 1GB of RAM and 8GB of disk space. This farm is using the disk space provided by the Golias farm, through NFS. This is possible even though the farms are physically located on distant city places, because of a good quality of network interconnection provided.

3 Software Configuration

In this section, we will describe how the software (Grid middleware and experiment specific software) is installed, shared and configured on the farms.

3.1 Golias farm

The Grid middleware configuration of the farm is described by the following figure.

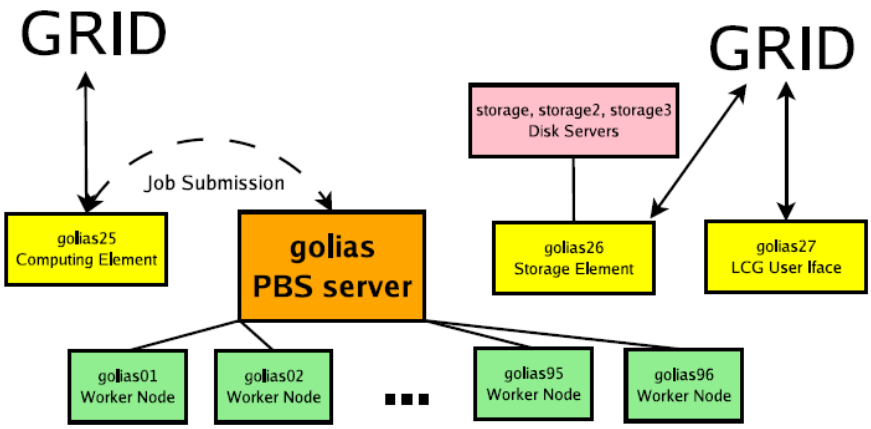

Fig. 2. Golias farm architecture

The server node golias serves as a main PBS (local batch system) server for the whole farm. PBSPro [1] version of the PBS software is used. This caused that the installation and configuration of the LCG/EGEE [13] middleware was non-trivial, as the middleware is written in a way to support Torque version of the PBS, which is slightly different from the PBSPro version. We have described all the adjustments that were needed to be performed (against a 2_4_0 version of the LCG/EGEE middleware) in [12]. Golias server is also used as an entry-point to the farm for the users that are not running Grid-aware jobs, but they are willing to submit jobs manually to the local queues that are available for such purpose.

The golias25 node serves as a Computing Element in the terms of the LCG/EGEE [13] middleware and submits all the Grid jobs that are obtained from the Grid to the main PBS server at golias. The LCG/EGEE queues are separated from the local queues, so that the capacity (or number of CPUs) can be easily adjusted when needed.

The golias26 serves as a Storage Element, providing the Grid-aware storage entrypoint for the storage at the site. The data that Grid users (or jobs) are storing to this Storage Element are immediately stored on one of our disk servers through NFS mounts.

The golias27 machine is User Interface, allowing local users to submit jobs to the LCG/EGEE Grid.

All application projects (or groups, in terms of Virtual Organizations) have their software installed on the shared area, physically stored on the disk servers and shared among all the Worker Nodes, so that jobs can access the software installation easily, no matter the Worker Node they are executed on.

The Golias farm supports the following HEP experiments: ATLAS [5], ALICE [4], D0 [7] (these three experiments are supported through the LCG/EGEE and SAMGRID [17] distributed Grid infrastructures), STAR [8], AUGER [6] (these are supported in terms of locally submitted jobs).

3.2 Skurut farm

The architecture of the farm is described by the Figure 3. From the „Grid point of view“ this farm has more elaborate configuration than the one described above for the Golias farm. This is because the resources of the Skurut farm are shared between different Grid projects and the individual nodes provide a variety of different tasks.

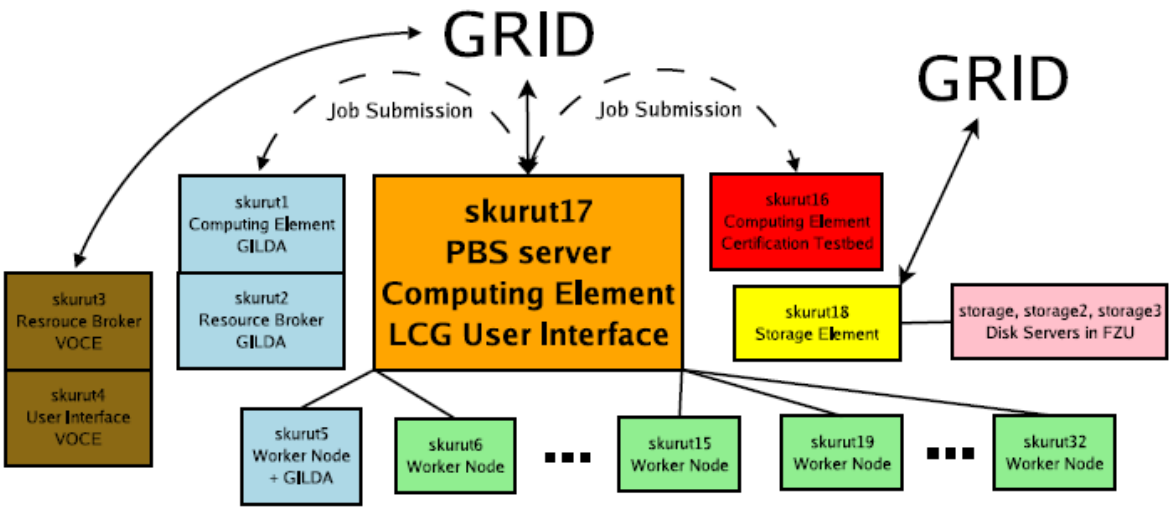

Fig. 3. Skurut farm architecture

The skurut17 server is a „central“ node of the farm - it is a Computing Element for the LCG/EGEE Grid and also hosts a local batch system (Torque) server for the farm. User Interface is also installed on this server. As can be seen, this Torque server is used also by skurut1 and skurut16, which are both LCG/EGEE computing elements, but with no batch system installed. They „forward“ all the Grid jobs that they receive to proper queues of the PBS server skurut17. The skurut1 is a Computing Element for GILDA, the system that is used for dissemination and training activities within the of EGEE project [3], called GILDA. GILDA is a dissemination Grid network that operates on top of the standard LCG/EGEE middleware, but in addition to that provides applications that are designed to teach new users of Grid how to use the infrastructure. Many tutorials, held worldwide, are often using the GILDA based Grid.

The skurut16 is a Computing Element dedicated to the Certification Testbed activity [18]. The Certification Testbed is an EGEE CE (Central Europe) region joint initiative to facilitate the development of Grid software by providing people from CE community with access to EGEE-compliant Grid infrastructure for learning, testing and software validation purposes. The infrastructure is independent from EGEE production infrastructure, it is composed of Grid resources including all necessary central services and a minimal set of computing nodes.

skurut3 and skurut4 machines provide workload management and User Interface servers for the VOCE activity [19]. VOCE is a virtual organization provided for all EGEE Grid users within the Central Europe region willing to utilize computational resources available in the CE and are not members of any already existing VO. VOCE directly supports CE researchers by providing a computing service. This service consists of shared data resources and computational capacities available within the CE and the installed Grid middleware and other software to solve various types of computational jobs. skurut3 is both Resource Broker for the VOCE-enabled resources, and also provides a RLS/RMC catalogue (database serving a data storage catalogue within a VO) for the VOCE needs. This RLS/RMC catalogue is published in the main LCG BDII located at CERN, so that it is easily reachable. skurut4 is a User Interface. We are running patched version of openssh server on this machine, so that users can be authenticated in a way that is usual in the Grid environment - by their X.509 Grid certificate. Also the creation of the user accounts on the User Interface is automatic, and the user account is created as soon as its certificate is registered within VOCE. This is managed by Perun [11] system, developed at Institute of Computer Science at Masaryk University in Brno.

4 Network Connectivity

The network connectivity is provided by CESNET, via its CESNET2 high speed academic network to the pan European Geant and Geant2 networks. The farms have further access to dedicated 1 Gbps external optical lines, which makes the integration into international Grid an easy task.

As the HEP computing often involves the need to transfer large amounts of data needed for analysis, we have also separate optical link, in the scope of the CzechLight project by CESNET to FermiLab [10] in Chicago. This dedicated link is used to transfer all the data for the computations for the D0 experiment, that are run on the Golias farm. This link is operated on top of BGP routing protocol implementation, so that in case of a failure, the traffic is automatically rerouted to „standard“ way through Internet.

5 Conclusion, Results

Both Golias and Skurut farms are used for the computing tasks of the experiments, and Skurut farm is also used for educational and dissemination activities of the EGEE project, with active participation in the EGEE activities (VOCE setup and management).

Both farms participated in several large scale tests of the Grid infrastructure performance executed by Atlas experiment. The so called Rome production took place during the first half of 2005. All participating sites with LCG middleware processed altogether about 40000 successful jobs, Golias and Skurut farms contributed at the level of 4% respectively 1%.

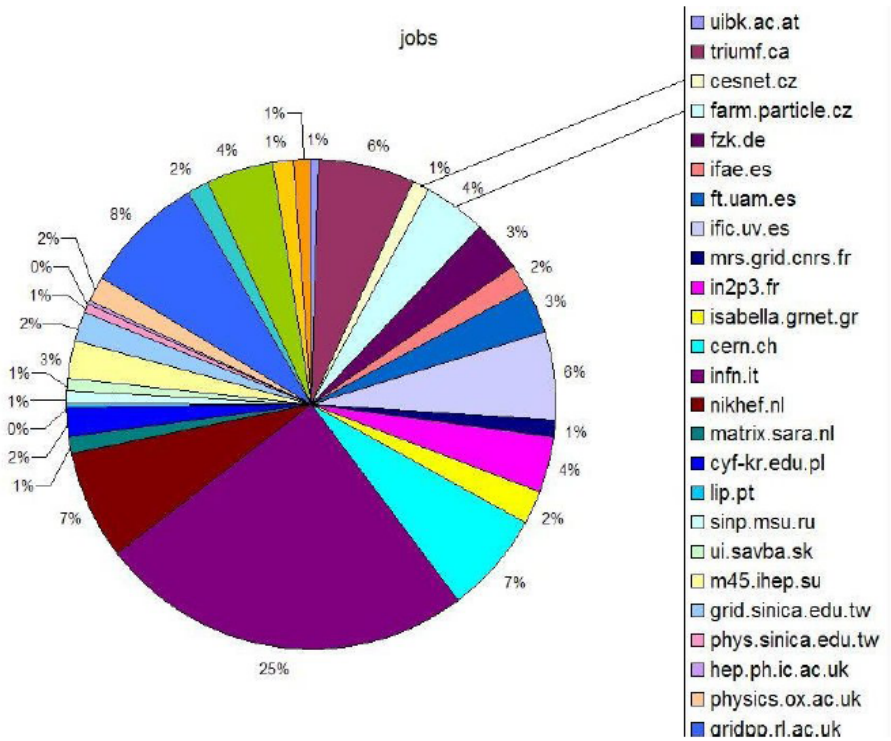

Fig. 4. Job distribution on LCG sites during ATLAS Rome production

For the D0 experiment production, only the Golias farm was used. The farm was used to reprocess about 28000000 events (which equals 5.57% of the total production amount), ranking the farm 4th when counting the number of processed events, as can be seen on Figure 5. The processing was done using the Grid middleware in the scope of the SAMGRID project.

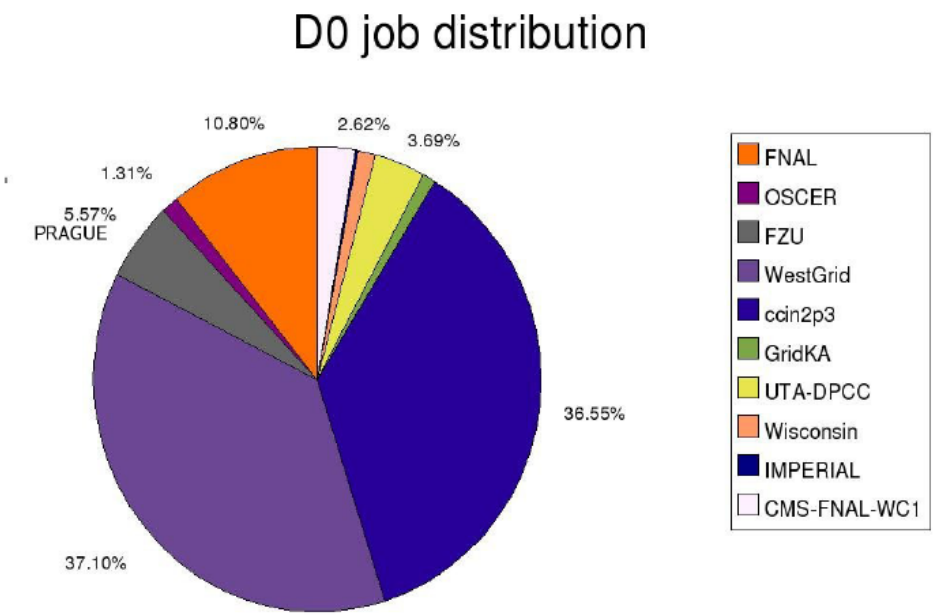

Fig. 5. Computed portion of D0 jobs

References

[1] Altair, PBSPro, http://www.altair.com/software/pbspro.htm
[2] CESNET, Homepage, http://www.cesnet.cz/
[3] EGEE, Homepage, http://cern.ch/egee
[4] Alice Experiment, Homepage, http://cern.ch/alice
[5] Atlas Experiment, Homepage, http://cern.ch/atlas
[6] Auger Experiment, Homepage, http://www.auger.org/
[7] D0 Experiment, Homepage, http://www-d0.fnal.gov/
[8] Star Experiment, Homepage, http://www.star.bnl.gov/
[9] Golias Farm, Homepage, http://www.particle.cz/farm
[10] FermiLab, Homepage, http://www.fnal.gov/
[11] Aleš Křenek and Zora Sebastianová, Perun - fault tolerant management of grid resources, Proceedings of Krakow Grid Workshop, Krakow, Academic Computer Center Cyfronet, pages 133-140, 2005.
[12] Jiří Kosina, Running LCG on top of PBSPro, http://www.gridpp.ac.uk/tb-support/faq/LCG2_PBSPRO.txt
[13] LCG, Homepage, http://grid-deployment.web.cern.ch/
[14] METAcentrum, Homepage, http://meta.cesnet.cz/
[15] Institute of Physics Academy of Sciences of The Czech Republic, Homepage, http://www.fzu.cz/
[16] Gilda Project, Homepage, https://gilda.ct.infn.it/
[17] SAMGRID, Homepage, http://projects.fnal.gov/samgrid/
[18] Certifcation Testbed, Homepage, http://grid.cyfronet.pl/egee/tiki-index.php?page=Certification+Testbed+Main+Page
[19] VOCE, Homepage, http://egee.cesnet.cz/en/voce/index.html