

# GridPP

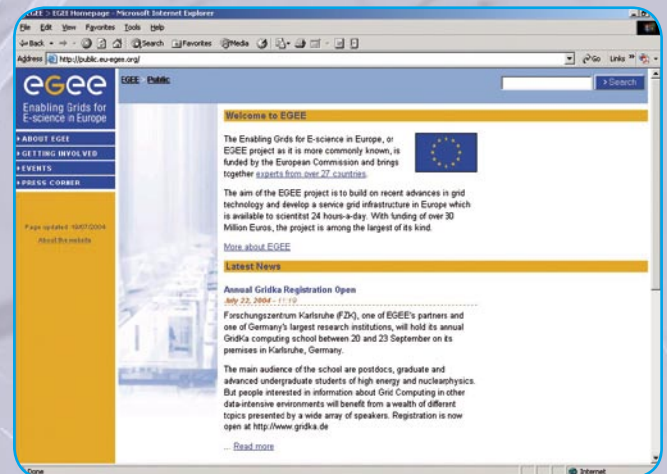
## Middleware - running a computing grid

'Middleware' is the key to a successful Grid. Middleware allows the software being used by scientists to talk to the Grid's hardware, distributing computing jobs efficiently around the network. It also deals with issues such as security, ensuring that only authorised users can access the Grid. Essentially it provides an interface between the user and the Grid, so that users don't need to make decisions themselves on the best place to send their job or retrieve their data, or have accounts on many different machines.

GridPP uses middleware developed as part of the European DataGrid (EDG) project. In association with the CERN's LHC Computing Grid project (LCG, [lcg.web.cern.ch](http://lcg.web.cern.ch)) we have tested and hardened these and other components from the Virtual Data Toolkit ([www.cs.wisc.edu/vdt](http://www.cs.wisc.edu/vdt)). Middleware development continues as part of the Enabling Grids for E-science in Europe project (EGEE, [www.eu-egee.org](http://www.eu-egee.org)). Some of the key components are described below.

### • Workload Management

The first step in using the Grid is to describe a job in a way that the Grid middleware can understand. This is done using a high level Job Description Language (JDL), through which a user is able to specify: the characteristics of the job itself (the executable name, the parameters required, the number of instances to consider, standard input/output/error files etc.); its required resources (CPU, storage, data etc.); and how the execution sites should be ranked (e.g. by fastest CPU, greatest number of free CPUs etc.).



Enabling Grids for E-science in Europe (EGEE)

### • Information and Monitoring

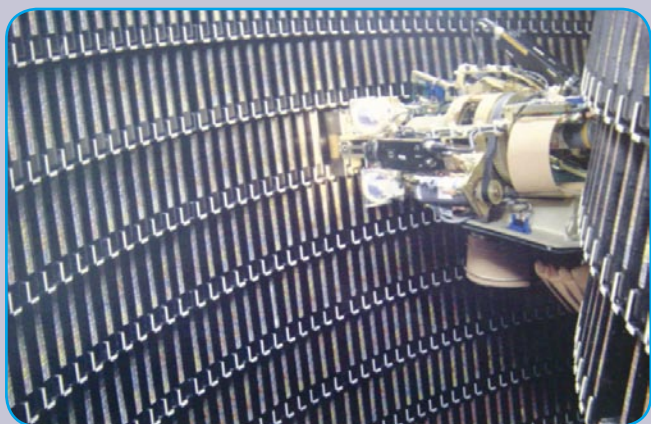
To operate a Grid, information about its resources (computing, storage and networking) and applications must be shared and monitored. As part of EDG, GridPP researchers have developed and implemented the Relational Grid Monitoring Architecture (R-GMA, [www.r-gma.org](http://www.r-gma.org)), which uses the SQL query language to publish and access distributed information. R-GMA transparently combines information from multiple sources and is migrating to WS-I compliant Web Services.

### • Security

For security, GridPP uses X.509 digital certificates issued by the UK e-Science Certificate Authority, and other European equivalents. The Certificate Authority checks the identity of users, providing them and computer hosts with electronic certificates that can authenticate them to the Grid. Based on this authentication, users can then be authorised (or not) for a range of tasks through their membership of a virtual organisation. GridPP has also used this concept to produce a web facility called GridSite ([www.gridsite.org](http://www.gridsite.org)), which lets users edit a website remotely, using their certificates for authentication.

## • Storage Element

Storage middleware provides a uniform interface to storage systems in general, and to mass storage systems in particular. Within EDG, the storage middleware used by GridPP was deployed as an interface to the Rutherford Appleton Laboratory Atlas DataStore, as well as to other storage at various sites throughout Europe.



The tape store at Rutherford Appleton Laboratory

## • Data management

In a Grid, data is one of the most important resources, and so must be managed effectively and efficiently. GridPP has contributed a large effort towards developing a set of integrated data management services. It also played a key part in the development of Spitfire, which permits access to relational databases from the Grid, and OptorSim, a Grid simulation for studying file replication algorithms.

## • Fabric management

A Grid site's computing fabric consists of a collection of nodes (for example, computers or routers), which need to be set up correctly - configured - in order to work optimally with each other, with application software, and with other nodes in the Grid. GridPP has capitalised on the UK development of the LCFG tool ([www.lcfg.org](http://www.lcfg.org)), which provides a sophisticated and highly automated approach to fabric management.

## • Network Monitoring Services

The network monitoring suite used by GridPP examines the network connections between sites, and publishes this information into the Grid information system. The information is used for two purposes (i) to allow the status of the entire Grid to be viewed through a graphical tool (MapCentre) and (ii) to publish connectivity and throughput rate information which in future can be used by the middleware in order to make scheduling decisions with respect to data transport.

## • The user point of view

So, what does a user need to do to submit a job to the Grid? Key steps are:

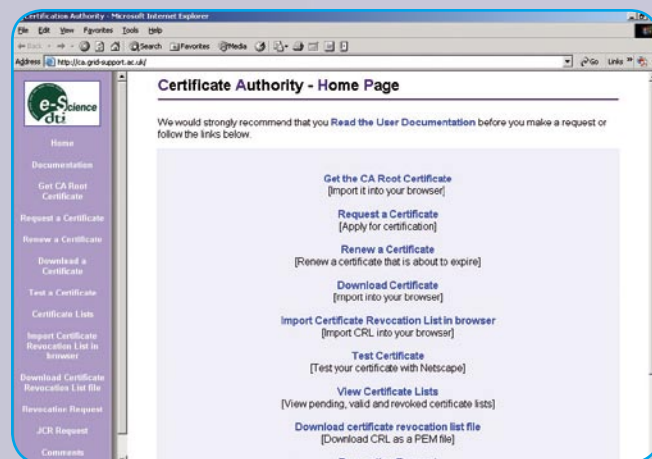
**Preparation** - A Grid user must have a valid X.509 digital certificate. They must also be a member of a supported virtual organisation and have access to a user interface, which acts as a gateway to the Grid.

**Getting started** - The user creates a proxy certificate. This is a copy of the real certificate, but with a limited lifetime (typically twelve hours). It avoids the user having to use their real certificate in Grid jobs that could compromise the certificate's safety.

**Submitting a job** - The user copies their precompiled executable or script, additional libraries and input files to the user interface computer. They then write a text file containing a description of the job in the Job Description Language (JDL), and submit the job using the 'job submit' command. A unique job ID is returned to the user in the form of a URL that can be used to identify the job later on.



**Retrieving the results** - Once the job has successfully completed, the user can retrieve the output using the job ID. They can also get detailed logging information about the progress of the job from submission to retrieval of the output.



The UK e-science certificate authority