Experiences in Setting up a Pan-European Datagrid using QCDgrid technology as part of the ENACTS Demonstrator Project

Chris Johnson, James Perry, Lorna Smith and Jean-Christophe Desplat

June 18, 2004

Abstract

The recent QCDgrid software, funded by the GridPP (PPARC/EC) project, will be discussed, along with our experience of constructing a pan-European datagrid using QCDgrid during the EC-funded ENACTS Demonstrator activity. QCDgrid was written to handle QCD data, however, we have found QCDgrid to be an excellent tool for managing data/metadata in general, across the grid, once the highly non-trivial task of setting up the underlying technologies (*eg* Globus, firewalls, etc.) has been achieved.

1 Introduction

The QCDgrid software [1] was developed to satisfy the high data storage and access requirements of UKQCD, a group of geographically dispersed theoretical Quantum Chromodynamics (QCD) scientists in the UK.

QCD is an application area that requires access to large supercomputing resources and generates huge amounts of raw data. UKQCD currently stores and requires access to around five terabytes of data, a figure that is expected to grow dramatically as the collaboration's purpose built HPC system, QCDOC, comes on line later in 2004. This data is stored on QCDgrid, a data grid currently composed of six storage elements at four separate UK sites: Edinburgh, Liverpool, Swansea and RAL. Southampton will soon join this grid.

QCDgrid is part of the GridPP project [2], a collaboration of Particle Physicists and Computing Scientists from the UK and CERN, who are building a Grid for Particle Physics.

UKQCD is a well established collaboration, with a high level of trust and with similar system administration policy requirements between different sites. In addition many of the storage elements have been purchased with input from the collaboration, and are thus fairly homogeneous in nature. The UKQCD software is however designed to be platform independent and to be applicable across a wide variety of sites. In this paper we focus on the recent deployment of the QCDgrid software across a series of European sites, as part of the EC-funded ENACTS (European Network for Advanced Computing Technology for Science) [3] demonstrator project: "Demonstrating a European Metacentre". This involved setting up a pan-European data grid and represents an ideal test of the software across a series of heterogeneous sites. This led to some interesting issues regarding policy requirements and certificate usage across countries which will be discussed later.

2 ENACTS

ENACTS is a Co-operation Network in the 'Improving Human Potential Access to Research Infrastructures' Programme and brings together many of the key players from around Europe who offer a rich diversity of High Performance Computing (HPC) systems and ser-

vices. The project consists of 14 partners across Europe involving close co-operation at a pan-European level to review service provision and distil best-practice, to monitor users' changing requirements for value-added services, and to track technological advances. Back in 1999, when the ENACTS programme was set up, the key developments in HPC were in the area of Grid computing and driven by large US programmes. In Europe we needed to evaluate the status and likely impacts of these technologies in order to move us towards our goal of European Grid computing, a 'virtual infrastructure' - where each researcher, regardless of nationality or geographical location, has access to the best resources and can conduct collaborative research with top quality scientific and technological support. One of the mailn goals of the ENACTS project was to perform these evaluations and assements in the European environment.

The ENACTS project itself was split into two phases. Phase I mainly comprised of reports and surveys where current trends in HPC and grid technology were identified. Our 'Demonstrator' activity was part of Phase II. The objective of the 'Demonstrator' was

"*to draw together the results of the Phase I technology studies and evaluate their practical consequences for operating a pan-European Metacentre and constructing a best-practice model for collaborative working amongst facilities.*"

A number of the technologies identified in Phase I were inherent to the QCDgrid software described in the next section. Our deployment of QCDgrid technology enabled the 'practical consequences' of setting up a pan-European to be evaluated. The conclusions of this evaluation are described here.

3 The QCD Data Grid (QCDgrid)

The aim of the QCD data grid is to distribute the data across the sites:

- **Robustly** Each file must be replicated at least two sites;
- **Efficiently** Where possible, files should be stored close to where they are needed most often;
- **Transparently** End users should not need to be concerned with how the data grid is implemented.

3.1 Hardware and Software

The QCDgrid software builds on the Globus toolkit [4]. This toolkit is used for basic grid operations such as data transfer, security and remote job execution. It also uses the Globus Replica Catalogue to maintain a directory of the whole grid, listing where each file is currently stored. Custom written QCDgrid software is built on Globus to implement various QCDgrid client tools and the control thread (see later). The European Data Grid (EDG/EGEE) software [5] is used for virtual organisation management and security. Fig. (1) shows the basic structure of the data grid, and how the different software packages interact.

3.2 Data Replication

The data is of great value, not only in terms of its intrinsic scientific worth, but also in terms of the cost of the CPU cycles required to create or replace it. Therefore, data security and recovery are of utmost importance. To this end, the data is replicated across the sites that form the QCDgrid so that even if an entire site is lost, all the data can still be recovered.

This system has a central control thread running on one of the storage elements which con-

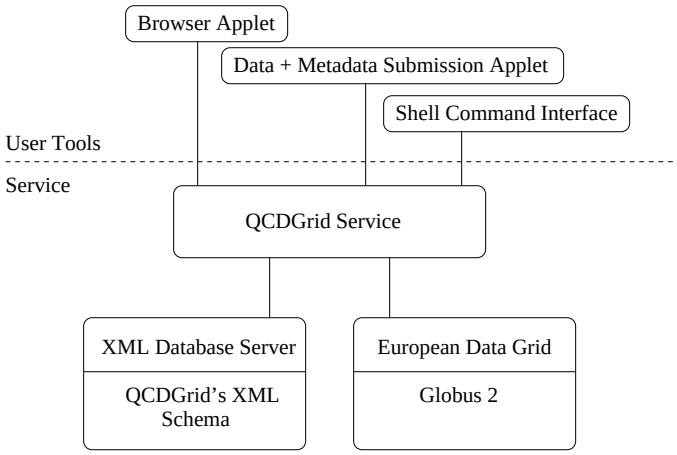

Figure 1: Schematic representation of QCDgrid, showing how the different software packages interact.

stantly scans the grid, making sure all the files are stored in at least two suitable locations. Hence when a new file is added to any storage node, it is rapidly replicated across the grid onto two or more geographically separate sites.

3.3 Fault Tolerance

The control threads also scans the grid to ensure that all the storage elements are working. When a storage element is lost from the system unexpectedly, the data grid software e-mails the system administrator and begins to replicate the files that were held there on to the other storage nodes automatically. Nodes can be temporarily disabled if they have to be shut down or rebooted, to prevent the grid moving data around unnecessarily.

A secondary node is constantly monitoring the central node, backing up the replica catalogue and configuration files. The grid can also still be accessed (albeit read-only) if the central node goes down.

3.4 File Access

The software has been designed to allow users to access files easily and efficiently. For example, it generally takes longer to transfer a file from Swansea to Edinburgh than it would to transfer it from another machine at Edinburgh. Therefore, when a user requests a file, the software will automatically return a copy of the replica of that file which is nearest to the user. Additionally, a user can register interest in having a particular file stored on a particular storage element, such as the one located physically closest to them. The grid software will then take this request into account when deciding where to store the file.

4 The MetaData Catalogue

In addition to storing the raw physical data, the project aims to provide an efficient and simple mechanism for accessing and retrieving this data. This is achieved by generating metadata, structured data which describes the characteristics of the raw data. The metadata is in the form of XML documents and is stored in an XML Database server (XDS). The XML database used is eXist, an open source database that can be searched using the XPath query language.

The XML files are submitted to the data grid, to ensure that there is a backup copy of the metadata. Hence the metadata catalogue can be reconstructed from the data grid if neces-

sary.

UKQCD's metadata contains information about how each configuration was generated and from which physical parameters. The collaboration has developed an XML schema, which defines the structure and content of this metadata in an extensible and scientifically meaningful manner. The schema can be applied to various different data types, and it is likely to form the basis of the standard schema for describing QCD metadata, being developed by the International Lattice DataGrid (ILDG) collaboration. ILDG is a collaboration of scientists involved in lattice QCD from all over the world (UK, Japan, USA, France, Germany, Australia and other countries), who are working on standards to allow national data grids to inter operate, for easier data sharing.

Data submitted to the grid must be accompanied by a valid metadata file. This can be enforced by checking it against the schema. A submission tool (graphical or command line) takes care of sending the data and metadata to the right places (see Fig. (2)).

5 MetaData and Data Grid Browser

The system also consists of a set of graphical and command-line tools by which researchers may store, query and retrieve the data held on the grid. The browser was originally developed by OGSA-DAI and has been extended to suit QCDgrid requirements. It is written in Java and provides a user-friendly interface to the XML database. The browser is also integrated with the lower level data grid software through the Java Native Interface and data can be fetched from the grid easily through the GUI. A simple interface for data/metadata submission and grid administration is currently under development.

6 The ENACTS Demonstrator

6.1 Deployment of QCDgrid

As previously described, QCDgrid was originally deployed on sites around the UK which were similar both in terms of the architectures used and the operating policies and grid environments present. Our task within the ENACTS project was to break away from these constraints and attempt to deploy QCDgrid across sites using the facilities already present. To summarise, this meant that instead of deploying QCDgrid over a set of Linux machines all running GT (Globus Toolkit) 2.0 and all using the same CA (Certificate Authority), we were provided with the facilities shown in Table. (6.1).

As can be seen, although the version of GT used is the same in all cases, it is not the same version as was deployed by UKQCD.

6.2 Scientific Scenario

In order for the Demonstrator activity to mimic the activities of a large collaboration, we required a "scientific scenario". In this case we appealed again to QCD. The reason QCD was chosen was simply due to the time constraints, *ie* the time associated with finding a new discipline. In principle there is no reason why a different subject could not have been chosen. It was also an advantage that we were able to adapt the XML schema already well-developed by the QCD community. Our scientific scenario involved the running of several Markov chains, a common application in QCD. Each of these chains was to run on a different node with the resulting data stored on the data grid. We were grateful to two members of the QCD community who acted as scientific users in addition to those working on the Demonstrator project itself, who could also act as users.

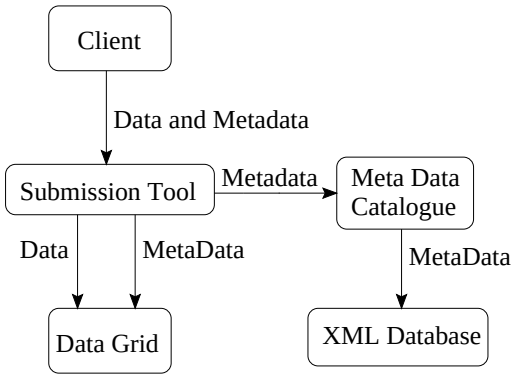

Figure 2: Schematic representation of data being added to the data grid and metadata catalogue.

Centre	Location	CA	Architecture OS	Gridware
EPCC	Edinburgh, UK	UK e-Science	Sunfire E15K Redhat 9	GT 2.4
Parallab	Bergen, Norway	NorGrid	Linux Cluster Solaris 2.9	GT 2.4 & Replica catalogue
TCD	Dublin, Ireland	Grid-Ireland	Linux Cluster Redhat 9	GT 2.4 & Replica catalogue

Table 1: The Facilities available to the ENACTS Demonstrator

6.3 Deployment problems

Most of the problems faced when deploying QCDgrid in this fashion were actually related to the Globus Toolkit itself and this will be commented on first. The issues relating to the pan-European deployment of this software along with using differing CAs and heterogenuos environments were actually of less significance, but are also worth commenting on.

The intention here is to describe the kinds of problems that are faced when setting up a pan-European grid, based on what we encountered during the ENACTS Demonstrator activity.

6.3.1 Globus and the Globus Replica Catalogue

There were really two separate issues facing us when dealing with Globus. Firstly there was the issue of setting up Globus to work between the sites in question, and secondly the issue of getting QCDgrid to work with the particular version of GT we were using.

At each of the three sites, installations of Globus had already been attempted and at two of the sites Globus was already in use. Our first task was to perform the most basic operation of running simple Globus jobs between the sites with each of the sites able to act as either client or server to any of the other sites. In our case this was simply a bi-directional, all-to-all task involving three sites, but it was still highly non-trivial. The reasons for this was due to both the fact that multiple CAs were involved (see later) and due to firewall issues which will not be discussed here.

It was surprising to find how many problems we were faced with when simply moving from GT 2.0 to GT 2.4. Some functionality previously found in libraries had disappeared and it turned out that the replica catalogue provided in GT 2.4 was not self consistent. The problems associated with missing libraries were

overcome with some simple re-writing of the QCDgrid code (the missing libraries were simply error handling routines). To solve the problems associated with the replica catalogue required us to completely rewrite the replica schema.

The installation of Globus on Solaris is not straightforward and for this reason the "information services" bundle was not installed. The implications of this were that the Solaris node could not act as the “control thread” and could not be used to submit data from. However, it was still able to act as a storage node.

6.4 Solaris

Moving away from the Globus installation issues, Solaris presented a few relatively minor problems with system specific functions used by QCDgrid and the usual issues which appear when a code is ported from one system to another. We were fortunate to have the GNU ‘gcc’ compiler installed on this system; code which calls Globus routines can be difficult to compile on native compilers.

6.5 Certificate Authority (CA)

In order for our grid to operate, each user was required to be able to run Globus jobs from any machine to any other on the grid. The control thread itself was also required to have the same ability. This creates a number of issues to do with trust and security as well as the logistical problems of acquiring CAs and getting them installed.

It was decided that each user would use their present certificate or would get hold of one from their respective countries. The length of time taken to acquire a certificate can vary considerably, from a few days in the case of the UK E-Science and NorGrid CA to several weeks in the case of Grid-Ireland. This is something which may have a big impact on systems with many users.

It was also necessary to acquire a certificate for the control-thread. This can present problems. Certificates are only generally available to real users who, in the case of UK E-Science, are expected to show photographic identification before being allowed to own a certificate. This is not really appropriate in the case of the control-thread which is simply a machine.

The UKQCD collaboration overcame this issue by allowing the control-thread to use the machine certificate to gain access to each of the machines involved. This was perfectly reasonable as UKQCD was using equipment exclusively for QCD. However, in the case of the ENACTS Demonstrator this was not the case.

Sharing of a machine certificate requires a high level of trust between systems which is not generally acceptable. We were fortunate in the situation as one of the CAs was able to provide a certificate for the use of the control-thread. This could in general be a problem if such an acquisition was not achievable.

6.6 The Pan-European Aspect

An interesting question is:

“How well does the QCDgrid system perform in a pan-European system?”

The answer to this question is that once the system is set up, the actual performance is perfectly acceptable. A few “timeout” tolerances needed to be changed to cope with the longer delays across Europe - typically from a few seconds to around 3 or 4 minutes for a simple Globus job and the frequency of failure was higher than with the UK system due to nodes being temporarily unavailable, but nothing which indicated a problem with the performance of submitting, querying or retrieving data.

7 Feasibility

7.1 User Feedback

7.1.1 Collecting feedback

This section describes the feedback we have received from users, both the scientific users and those involved in the ENACTS project itself. The feedback received was a mixture of email and face-to-face interviews conducted with the users.

7.1.2 Using the system

Users found that acquiring a certificate was in some cases an extremely frustrating activity resulting in one user dropping out of the activity before even using the system thus not allowing for much exposure of the system to users. However, although this was not a disaster for this relatively short project which was simply a "proof of concept", it does highlight the difficulties involved in getting a grid project off the ground. In general, users appeared to be happy with the functionality offered by QCDgrid, although would have liked to see it integrated with a job-submission tool (work is presently being undertaken on this facility).

7.1.3 Metadata

While the benefits of using metadata are obvious to some, it is not always easy to convey this to users. We found that even after explanation of the merits of using metadata systematically to describe your data, our users were not convinced that this was necessary and thought the creation of metadata an unnecessary overhead, even when the task of creating metadata was straightforward. The whole concept of metadata, and converting users to it, is something which would need to be undertaken at the collaboration level, as is already happening within some scientific communities (for example, see [6, 7]).

8 Conclusions: Setting up a pan-European Datagrid

We have shown that it is possible to set up a small pan-European datagrid using the QCDgrid system given a set of suitable resources (hardware and software) and have highlighted some of the problems which would need to be solved if the system were to be used for a full-scale pan-European collaboration. It is worth adding that most of the problems encountered were simply in the setting up of Globus and acquiring the necessary certificates for users. We conclude that subject to a successful installation of Globus and associated software and the resolving of any firewall and certificate problems between participating centres, QCDgrid provides a perfectly good environment in which to manage data across Europe. The system contains nothing which is specific to QCD and could easily be used within other disciplines.

9 Acknowledgements

We would like to thank Mike Peardon and Jimmy Juge for their assistance with the scientific scenario and providing us with very useful feedback on the system. We would also like to thank Craig McNeile and Bálint Joó from the UKQCD collaboration for several useful discussions.

References

[1] *QCDGrid*, http://www.epcc.ed.ac.uk/computing/research_activities/grid/qcdgrid/.

[2] *GridPP*, http://www.gridpp.ac.uk.

[3] *ENACTS* http://www.enacts.org.

[4] *The Globus Toolkit*, http://www.globus.org.

[5] *EGEE: Enabling Grids for E-science in Europe*, http://www.eu-egee.org.

[6] *ILDG: International Lattice DataGrid*, http://www.lqcd.org/ildg/

[7] *CML: The Chemical Markup Language (CML™)*, http://www.xml-cml.org/