QCDgrid Software Review 2

Project Title: QCDgrid

Document Title: QCDgrid Software Review 2

Document Identifier: QCDGRID2-D1-4-2-SWR

Document Filename: D1.4.2_Software_Review_2.doc

Distribution Classification: Public

Authorship: Mark G. Beckett, James T. Perry

Approval List: QCDgrid Project Management Board

Distribution List: Public

Document History:

Personnel	*Date*	*Summary*	*Version*
MGB/JTP	19/SEP/2006	First release	1.0

Contents

1 Introduction

1.1 The QCDgrid Project

The QCDgrid project [14] is a core activity of UKQCD [13], a collaboration of UK academics and researchers that aims to procure and jointly exploit computing facilities for lattice field theory (commonly referred to as Lattice QCD) calculations. The primary aim is to increase the predictive power of the *Standard Model of elementary particle interactions* through numerical simulation of *Quantum Chromodynamics*. Such numerical simulations produce significant amounts of data and the purpose of the QCDgrid project is to provide software and supporting infrastructure that simplifies the management, storage, and manipulation of this data.

In the first three years of the project (2002—2004), software engineers at EPCC developed QCDgrid—a data management system that combines the distributed resources of the collaborators into a robust facility called the *UKQCD Grid*. The result is a multi-terabyte storage facility over seven sites at: University of Columbia, University of Edinburgh (including the UoE Advanced Computing Facility), University of Liverpool, Rutherford Appleton Laboratories (RAL), University of Southampton, and University of Swansea. The University of Glasgow is also a member of the consortium.

The facility is based on commodity hardware and open-source software. The hardware consists primarily of high specification, PC-based servers running the Linux operating system and managing large RAID storage arrays. On top of this infrastructure, the QCDgrid software (built using components from the Globus Toolkit [9], EGEE application stack [6], and an XML database) provides *Datagrid* management and user functionality – furnishing a simple and intuitive environment that hides the complexities of the underlying grid and presents a standard file system to the user. It incorporates a robustness metric that automatically disperses datasets across the grid, providing a resilience that ensures data is not affected by the loss of one (or possibly more) storage nodes. Security is leveraged from the Globus Toolkit, based on X.509 digital certificates issued by an approved Certificate Authority. The result is a reliable, secure data management system.

UKQCD is an important contributor to the International Lattice Data Grid (ILDG) [12], a group of like-minded scientists, working around the world, who aim to share their data to accelerate scientific progress in the field of Lattice QCD. The ILDG was initiated in 2002 by UKQCD and, at the time of writing, has significant representation from research groups in Australia, France, Germany, Italy, Japan, UK and USA.

The ILDG infrastructure is being assembled as a web services layer that will aggregate the particular resources of contributing collaborations (for example, the UKQCD Grid) for the benefit of the wider community. To achieve its objectives, ILDG has established working groups to:

- Facilitate data sharing, through the standardisation of the form and content for Lattice QCD scientific data and associated metadata.
- Produce a set of specifications that define an architecture for an international *Grid of Grids* for Lattice QCD.

1.2 QCDgrid Software Reviews

This report is the second software review conducted as part of the QCDgrid project. This review and the previous one [5] consider opportunities to enhance or improve the UKQCD Grid, focusing on both the QCDgrid software and external applications/services resourced by the project. The conclusions drawn from the reviews helps to define the strategy for evolving the UKQCD Grid over the duration of the project. We do note, however, that this review does not consider solutions that could be considered as replacements for the entire QCDgrid suite.

We begin, in Section 2, with an overview of the UKQCD Grid infrastructure, highlighting the functionality requirements and software dependencies of the infrastructure. Then, in Section 3, we specify the third-party software that is employed within the production grid. This is complemented by

Section 4, which provides an appraisal of the software deployed in relation to other comparable solutions that might be available. Based on the investigations that have been performed, we draw together conclusions and proposals for further work in Section 5.

2 Overview of UKQCD Grid

The UKQCD Grid is a *data grid* that combines the disparate, commodity storage resources of the UKQCD collaboration into a unified data resource, which provides:

- multi-Terabyte, distributed storage.
- automatic data replication across multiple sites.
- continuous data validation and consistency checking.
- data provenance with application-specific metadata.
- simple and intuitive client tools.
- interface between managed storage and external compute resources.

2.1 Architecture

The UKQCD Grid has a deployment architecture illustrated in Figure 1. It is a distributed computing installation that we consider as containing three types of node, as follows:

- Client
- Storage node
- Control node

These three types of node are described, in turn, below.

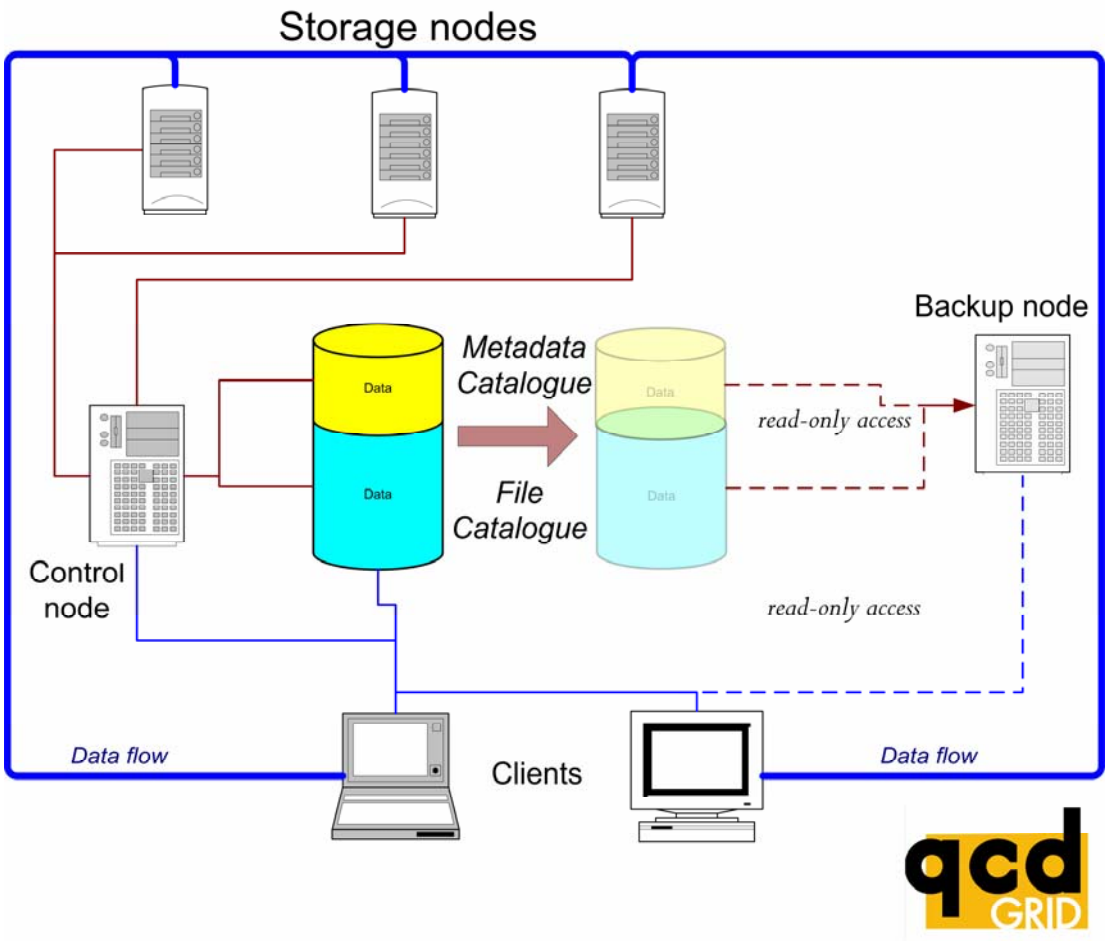

Figure 1: Overview of the QCDgrid architecture showing flow of data and control between clients, storage elements, control node and the various catalogue services.

2.1.1 Client

A user accesses the UKQCD Grid using a client node. This can be any computer on which: the QCDgrid client tools are installed, plus the user has a user account and valid X.509 certificate. From a UKQCD Grid client node, a user can:

- Initialise a proxy certificate that allows them to *authenticate* to UKQCD Grid services.
- Confirm that the UKQCD Grid is operational.
- List and retrieve files from the data grid.
- Query the metadata catalogue for ensemble or gauge configuration metadata.
- Submit new data/metadata to the data grid.
- Submit a compute job to a remote compute resource[1].

All of these interactions are supported by the QCDgrid Command Line Interface (CLI). In addition, a subset of the above functionality is furnished by the QCDgrid Graphical User Interface (GUI).

2.1.2 Storage element

A storage element provides disk-based storage space to the UKQCD Grid for holding copies of user data files. The data files held on a storage element are assumed to be immediately available (a property often referred to as *on-line* when compared to tape-based storage space which is referred to as *off-line*) via the GSI-secure FTP (GSIFTP) protocol. Support for GSIFTP is provided by the Globus Toolkit GridFTP server. At the time of writing, a mix of server versions (both Version 2.4 and Version 4.0) are deployed on the UKQCD Grid.

User file transfers to and from Storage elements are normally initiated from the QCDgrid CLI and GUI clients. Therefore, a user need never access the storage element directly.

Other file transfers are also implemented by the Control Node as part of its data replication and validation functionality (see Section 2.1.3).

At present all client and storage nodes run Linux, however any flavour of UNIX which supports pre-web service Globus will run the QCDgrid software.

2.1.3 Control node

The control node hosts a persistent agent (commonly referred to as the *Control Thread*) that continuously tests and validates the integrity and availability of the contents of the data grid. It performs a number of checks, including:

- checks that storage nodes are responding (based on response to a Globus GRAM job requests).
- checks if new data is awaiting insertion into the data grid.
- checks that sufficient copies of each file exist on the data grid, and performs a replication operation if a file is detected with insufficient copies.
- compares the state of storage node file systems with a record that it maintains locally in the *file catalogue*.
- checks that each copy of a file is consistent with its peers – that is, all replicas of a file are identical.
- computes the free capacity that remains in the data grid and takes appropriate action if the capacity is deemed to be too low.

The control node is autonomous and typically orchestrates management of the data grid without intervention from a user/administrator.

[1] At present, the UKQCD Grid does not provide any compute resources. All current compute resources are furnished and managed outwith the QCDgrid project.

The control node is a critical component of the QCDgrid system architecture. If the control node becomes unavailable, then the system switches to a backup control node (called the Backup Node) that provides read-only access to the data grid.

2.1.4 Backup node

The Backup Node provides a subset of the functionality of the Control Node, permitting users to retrieve data from the grid, in the event of a failure of the Control Node. Specifically, the Backup Node hosts a copy of the file catalogue allowing the QCDgrid client tools to locate and retrieve data. from the Storage Elements The Backup Node is intended as a temporary substitute for the Control Node: it cannot add new data to the grid, nor can it replicate or validate existing data on the grid.

2.2 Data management

The QCDgrid software supports the storage of file-based datasets on distributed storage space. It also allows the user to bind metadata (marked up using an XML schema) to data-files, providing a data provenance function.

Within the UKQCD Grid, the majority of data-files represent lattice gauge configurations. The form in which gauge configurations are stored has been standardised, by the ILDG [12], to facilitate the sharing of datasets between collaborations (at the time of writing, the ILDG file format is at Version 1.0). This file format is complemented by metadata defined within an XML application called QCDML (at the time of writing, QCDML is at Version 1.3). The specifications of both the file format and the metadata schema are available from the ILDG [12].

The QCDgrid software stores metadata separately from data – in an XML database. This facilitates query and search operations, which are implemented by the QCDgrid GUI Browser (see figure 2).

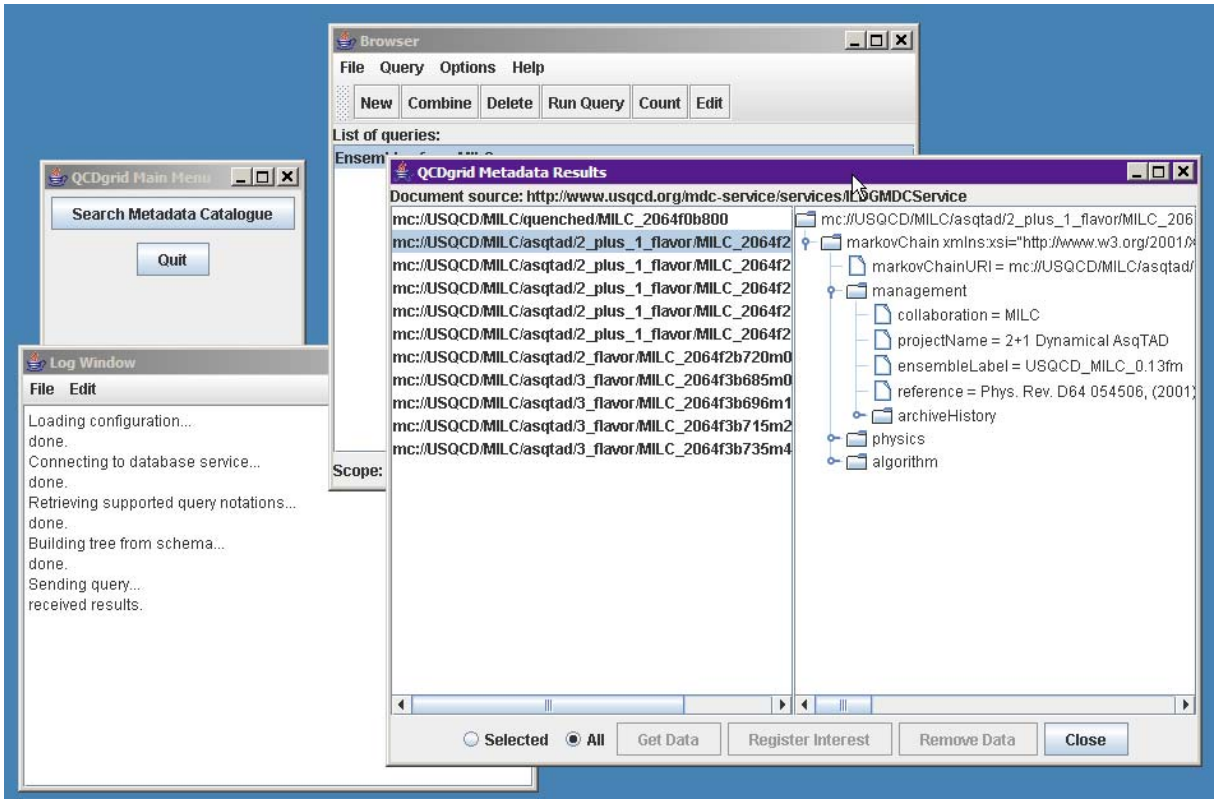

Figure 2: QCDgrid GUI Browser displaying query results from a search of gauge configuration metadata (running in ILDG mode).

2.3 ILDG Service Layer

As noted in Section 1.1, the UKQCD is a key contributor to the International Lattice Data Grid (ILDG). As part of the commitment to ILDG, UKQCD are endeavouring to make the data held within the UKQCD Grid available to scientific communities around the world. This is being achieved with a standardised, web service layer that is being assembled by the ILDG. The UKQCD contribution to this activity is, in part, captured within work packages 3 and 4 of the QCDgrid project [14].

At the time of writing, the UKQCD Grid supports unauthenticated read-access to the metadata catalogue through a web service hosted on the Control Node. A modified version of the QCDgrid GUI has also been produced, called the ILDG Browser Client (see Figure 2). This is a lightweight version of the standard QCDgrid GUI with data grid functionalities disabled.

Preparations for authenticated read-access to the file catalogue and data grid are on-going. Once implemented, the ILDG Browser Client will be extended to support this additional functionality.

2.4 Security

Almost all operations on the UKQCD Grid (both user- and agent-initiated), require authentication. This is achieved using the GSI architecture, which requires a user/agent to provide a valid X.509 (proxy) certificate with each operation request. Access control is achieved using the standard "Gridmap" mechanism that furnishes a binding between the distinguished names (part of the X.590 certificate) of recognised UKQCD Grid members and guest user accounts on UKQCD Grid resources.

Read operations to the Metadata Catalogue do not require authentication. This is in line with the ILDG policy that all collaborations make their scientific metadata public. Write/modify operations on the Metadata Catalogue are authorised by username/password, in line with the security implementation for the chosen XML database.

3 Software components

To fulfil the functionality of the UKQCD Grid, a number of third-party web and grid service components[2] are utilised. These are described, in turn, below.

3.1 Globus Toolkit

QCDgrid is a *Grid Application* that is built on top of the Globus Toolkit (at the time of writing, the QCDgrid team recommend the use of Version 4.0[3]). QCDgrid specifically utilises a number of the (non-web service) components of the Globus Toolkit, as follows:

- **Replica Location Service (RLS)** – RLS is a *catalogue* that maintains the mapping between each Logical File Names (LFN) – the unique identifier for a data file – and the physical locations of replicas of the data file.
- **Grid Resource Allocation and Management (GRAM) –** GRAM is used internally by the QCDgrid system to submit jobs to storage nodes that carry out certain bookkeeping tasks, such as monitoring the amount of free disk space on each storage node. GRAM is also implicitly accessed by the user, when invoking the QCDgrid job submission system.
- **GridFTP** – GridFTP is the primary data transport mechanism within the data grid. It is used by the control thread to replicate/relocate data files and (implicitly) by users to download/upload data from/to storage elements via QCDgrid client tools.

  In addition to the native transfer protocol (GSIFTP), GridFTP also supports a number of other standard protocols: HTTP; HTTPS; FTP and SFTP (although only GSIFTP is used by QCDgrid at present). However, GridFTP does not – at the time of writing – support the SRM interface (see Section 4.5.1).
- **Grid Security Infrastructure (GSI)** – GSI is used by all of the above components to authenticate service requests and, hence, control access to resources and data.

3.2 EGEE Application Stack

As noted above, user authentication within QCDgrid is implemented within the GSI architecture. Users authenticate using a short-lived, X.509 proxy certificate and gain authorisation for operations using a coarse-grain access control mapping called a Gridmap file, which must be defined on each QCDgrid node[4].

Within the UKQCD Grid, the Gridmap file is automatically generated from a central database that fulfils the role of a simple Virtual Organisation (VO) manager. This functionality is provided by a VO management database and software within the EGEE application stack.

3.3 Apache Tomcat / Axis Container

Apache Tomcat [2] (Version 5.0) is the web application container utilised by the UKQCD Grid. It hosts the metadata catalogue and, in conjunction with Apache Axis [1] (Version 1.3), it provides a web service hosting environment for the ILDG access layer.

In the current deployment, a single Tomcat/Axis instance is deployed on the control node. However, it is perfectly acceptable to host the metadata catalogue and ILDG web services on distinct servers, not necessarily coincident with the control node or backup node.

---

[2] All third-party applications leveraged by QCDgrid are freely available under an Open Source-type licence. The interested reader should consult licensing information provided with the specific software package.

[3] The QCDgrid project team recommend Globus Toolkit Version 4.0, though the software has a minimum requirement for Globus 2.4.

[4] All GSI-enabled grid services support Gridmap-based access control.

3.4 eXist XML Database

As noted above, the Metadata Catalogue is held in an XML database. QCDgrid requires a database that implements the XMLDB interface and supports the Xpath query syntax. At the time of writing, these requirements are satisfied by the eXist XML database [8].

3.5 Java Runtime Environment

A number of components of QCDgrid, and the UKQCD Grid more generally, require a Java Runtime Environment (JRE) with minimum version number 1.4.

4 Options for upgrade of software components

In line with the aims of this review, we now consider the key software components, in turn, and evaluate possible replacements/upgrades.

4.1 Globus Toolkit

The Globus Toolkit grid application environment is widely used across the international, academic community and beyond. The QCDgrid development team have significant experience with its usage, which is complemented by on-going expert support from the Globus development team[5].

Since the previous review, QCDgrid has been successfully updated to support the latest stable release of the toolkit (Version 4.0), though has not been modified to utilise any of the new or enhanced functionalities. Specific components that may enhance UKQCD Grid functionalities, include:

- **Reliable File Transfer (RFT) Service** – a WSRF [10] service that provides automated, asynchronous management of file movement operations. A client specifies a list of source and destination URLs to the service, which then creates a job description for the request that is archived into a database. Once the service has taken a request, interactions with it are similar to those with a job scheduler. Service methods are provided for querying the transfer status and subscribing to WSRF notifications of state change events.

  RFT has the potential to dramatically reduce the management function of the control thread code – which is currently required to monitor and maintain all data transfer operations through to completion. Furthermore, file transfers that are managed by RFT benefit from RFT Fault Tolerance and Recovery allowing specific transfer operations to recover in the event of transient errors. This further reduces the manual involvement in file transfer operations.

  RFT utilises GridFTP to perform the actual data transfer (see next item). Therefore, it supports the same transfer protocols as GridFTP.

- **GridFTP** – the implementation of GridFTP distributed with Version 4 of the toolkit represents a complete re-write of that distributed in previous versions. However, both the server and client provide interfaces that are backward compatible with previous versions – a fact that allows the UKQCD Grid to continue to run nodes with both Version 2 and Version 4 installations.

  GridFTP Version 4 incorporates extensions to the basic API that furnish a number of performance enhancements. One extension in particular is immediately useful to UKQCD: this is parallel transfer mode, in which the data to be transferred is shared between multiple network streams. This has the potential to significantly reduce the transfer time for a transaction.

- **Data Replication Service** (DRS) – the primary function of DRS is to automate the replication of multiple files, plus register the replication results in the RLS. In this way, the DRS overlaps with a subset of the QCDgrid Control Thread functionality. Currently, the DRS is classed as a technology preview and is unsuitable for production use.

- **Monitoring and Discovery Service (MDS) –** MDS is a suite of components that monitor and discover resources and services on a grid. MDS provides query and subscription interfaces to arbitrarily detailed resource data and a trigger interface that can be configured to take action when pre-configured incident conditions are detected. MDS has the potential to dramatically reduce diagnostic time for UKQCD Grid-related incidents by providing a rich presentation of status information for many different aspects of the UKQCD Grid operations.

[5] We have a fruitful relationship with a number of experts from the Globus team including: Lisa Childers (Globus Technical Product Manager, Argonne National Laboratory), Bill Allcock (GridFTP Technical Director, Argonne National Laboratories), Ann Chervenak (Technical Co-ordinator for Globus Replica Location Service, Information Sciences Institute, University of Southern California), Jennifer Schopf (Technical Co-ordinator for Globus Monitoring and Discovery Services, e-Science Institute, University of Edinburgh).

4.2 gLite

The gLite middleware [7], which is a substantial component of the EGEE application stack, is an analogous toolkit to Globus that provides a number of data management services potentially of relevance to QCDgrid.

Since the previous review [5], LCG and gLite have been refined and enhanced. The current version (Version 3.0) contains a rich set of components that has been hardened within the infrastructure of the World-wide LHC Computing Grid Project (WLCG). Here, we review the specific gLite components that are comparable to those provided by Globus and utilised by QCDgrid (Section 3.1):

- **LCG File Catalogue** – this service maintains mappings between LFNs and storage URLs that allow actual instances of datasets to be located in storage elements. In this way, it is similar to the Globus RLS. However, where it differs from Globus RLS is in its specific focus on files. Unlike Globus RLS, which provides a generic service to map unique identifiers (LFNs) to replicas, the LCG File Catalogue encapsulates logic that is specific to files. For example, it maintains a prescribed set of metadata including file size, creation data, and checksum information. This kind of information is very useful for a data grid. It can be crafted into the Globus RLS, which supports the inclusion of arbitrary metadata with catalogue entries, though requires some modification to the QCDgrid software layer.
- **File Transfer Service (FTS)** – this service is responsible for file transfers between storage elements. It is an analogue of the Globus RFT service though, at the time of writing, does not support automatic recovery from failed transfer operations. Unlike Globus RFT, FTS does support the Storage Resource Manager (SRM) interface (see 4.5.1).
- **Workload Management (WMS)** – this service exposes the gLite job submission functionalities. Version 3 of gLite provides a mature job control environment that is more than sufficient for the requirements of the QCDgrid software – which only requires simple GRAM job submission.

With regard to security, the gLite middleware utilises the same architecture as Globus – specifically, the Grid Security Infrastructure. The majority of gLite services comply with the GSI specification.

gLite is expected to be deployed across the GridPP infrastructure as part of its contribution to both WLCG and other GridPP-member experiments. Should UKQCD wish to utilise these storage and compute resources more widely, it would be necessary to modify the QCDgrid software such that it can interface with gLite-powered services. One option would be to re-engineer the software to be based on gLite services. This would necessitate an overhaul of the current design and would represent a significant volume of development work.

4.3 Virtual Organisation Management

The VO Management software that is currently used by UKQCD Grid is somewhat out of date. It has been superseded by the EGEE Virtual Organisation Management Service (VOMS), which provides a number of new functionalities for managing authentication and authorisation over and above the original VO software. UKQCD are working in collaboration with their ILDG collaborators to establish an ILDG-wide Virtual Organisation (registered within the VO pool). The actual service is already deployed and hosted by LDG on a server in DESY Zeuthen [11], it is supported by an ILDG VO policy statement, and is in the process of being populated with membership information.

Once this process is complete, it is anticipated that the UKQCD Grid security infrastructure will be modified to source authorisation information from the ILDG VO. The modifications will involve some modifications to the UKQCD Grid infrastructure, though more importantly will require a review of the security implications of the change.

4.4 XML databases

As noted in Section 3.4, QCDgrid uses an XML database in which to record metadata for managed datasets. QCDgrid requires that the chosen XML database implements the XMLDB interface specification and supports the XPath query syntax.

Alternative XML databases that could possibly be used instead of eXist include: Xindice [2], which is an open source database; SleepyCat dbXML, another open source database; and Tamino, which is a commercial XML database product from Software AG.

Xindice was the original choice of database for the UKQCD Grid. However, it was replaced with eXist as it exhibited poor performance for XPath queries. It is possible that a more recent version of Xindice has improved performance for such queries.

SleepyCat dbXML stores data internally in a relational form. Such databases are typically expected to provide better query performance. LDG – another contributor to the ILDG – uses a relational database to manage the metadata catalogue for the LATFOR collaboration. Their experience is that, while relational databases provide better performance than hierarchical databases (for example, an XML database), this performance is nullified by the complex nature of some XPath queries, when translated into a form suitable for a relational database.

At the time of writing, there are no known significant issues with the eXist database deployed for the UKQCD Grid. For this reason, adoption to an alternative database is considered a low priority.

4.5 Other possible components

4.5.1 Storage Resource Managers

The Storage Resource Management working group [15] are defining a specification for middleware components whose function is to provide dynamic space allocation and file management of shared storage components on grid infrastructure. Of greatest interest to UKQCD is the support for both off-line (for example, tape storage) and on-line (for example, disk storage) storage facilities. At present, all UKQCD storage elements are on-line. This may not always be the case in the future, in which case modifications would need to be made to the QCDgrid client tools to support off-line storage.

The SRM specification is an attractive direction to follow for two reasons:

1. SRM is the chosen interface for the WLHC (and hence the GridPP) infrastructure.
2. SRM-compliant storage elements are already deployed by several collaborations within ILDG – specifically, USQCD and LDG.

4.5.2 GITS

The UK Grid Integration Test Script (GITS) is a simple Perl application that can be used to validate the functionality of grid services. It provides good coverage of Globus services and, with a little effort on initial setup, furnishes a simple enhancement to reduce the diagnosis time for typical infrastructure problems within the UKQCD Grid. The intention is to invest a small amount of support effort to investigate the viability and impact of using GITS within the UKQCD Grid infrastructure.

The activity is actually simplified by the fact that the author of GITS, David Baker, is the local administrator for the Southampton Storage Element on the UKQCD Grid.

5 Conclusions

Based on the information gathered within this report, we conclude that the UKQCD Grid infrastructure generally makes appropriate use of the web and grid service tools that are available. Most of the third-party dependencies are realistically up-to-date and provided by live projects, reducing the risk that the UKQCD Grid will be isolated from future improvement. Furthermore, it is apparent that the UKQCD Grid integrates third-party solutions in a modular manner, improving the viability of future replacement for components that are deemed to be no longer suitable.

The report has highlighted several tasks that would be of benefit to UKQCD, as follows:

- The VO management software is that originally developed within the EDG project. Provision has already been made to replace this dependence with the EGEE VOMS alternative.
- The eXist database that manages the metadata catalogue is a pre-production release. A version 1.0 release candidate has been published by the eXist development team, which it would be appropriate to evaluate.
- The additional functionalities of Globus Toolkit Version 4 offer a number of potential enhancements to the UKQCD Grid. Once all of the resource providers have upgraded to Version 4 of the toolkit, it would be pertinent to evaluate these enhancements.
- The GITS application offers a simple route by which to reduce the time taken to diagnose problems and thus improve availability of UKQCD services. It would be worthwhile to invest a little effort in the implementation of GITS within the infrastructure.
- Switching to the gLite middleware would require a significant amount of development effort to re-engineer the QCDgrid software
- Looking into the future, it is likely that some provision will be needed for SRM protocol support within UKQCD. An investigation of the impact of such provision could realistically be undertaken with the cooperation of our ILDG collaborators.

6 References

[1] Apache Project, *Axis SOAP Environment*. Project homepage at http://jakarta.apache.org/tomcat.

[2] Apache Project, *Tomcat Web Application Container*. Project homepage at http://jakarta.apache.org/tomcat.

[3] Apache Project, *XIncide XML Database*. Project homepage at http://xml.apache.org/xindice/.

[4] D. Baker, *The UK Grid Integration Test Script – GITS*. Project homepage at http://www.soton.ac.uk/~djb1/gits.html .

[5] D.J. Byrne, *QCDGrid2 Software Review 1*. QCDgrid project deliverable (March 2005).

[6] EGEE, *Enabling Grid for E-Science*. Project homepage at http://public.eu-egee.org/ (2006).

[7] EGEE, *gLite – Lightweight Middleware for Grid Computing*. Project homepage at http://glite.web.cern.ch/glite/ (2006).

[8] eXist, *eXist – An Open Source Native XML Database*. Project homepage at http://exist-db.org/ (2006).

[9] Globus Alliance, *Globus Toolkit*. Project homepage at http://www.globus.org/ (2006).

[10] Globus Alliance, *The WS-Resource Framework*. Available from http://www.globus.org/wsrf/ (April 2006).

[11] International Lattice Datagrid, *ILDG Virtual Organisation*. Service hosted at https://grid-voms.desy.de:8443/voms/ildg/ (2006).

[12] International Lattice Datagrid. Project homepage at http://www.lqcd.org/ildg/tiki-index.php.

[13] UKQCD Collaboration. Homepage at http://www.ph.ed.ac.uk/ukqcd/ (2006).

[14] UKQCD Collaboration, *QCDgrid Project*. Homepage at http://www.gridpp.ac.uk/qcdgrid.

[15] SRM Working Group, http://sdm.lbl.gov/srm-wg/ (2006).