



QCDGrid2 Software Review 1

Project Title: QCDGrid2

Document Title: QCDGrid2 Software Review 1

Document Identifier: QCDGRID2-SWRev1

Document Filename: QCDGrid2-Review1.doc

Distribution Classification: Commercial In Confidence

Authorship: Daragh J. Byrne

Approval List: QCDgrid Development Team

Distribution List: UKQCD Collaboration

Document History:

<i>Personnel</i>	<i>Date</i>	<i>Summary</i>	<i>Version</i>
DJB	30 th March 2005	Completed.	1.0

Contents

1	Introduction	2
1.1	The QCDgrid Project	2
1.2	QCDgrid Software Reviews	2
1.3	Purpose	2
2	Software review.....	4
2.1	The QCDgrid Software Suite.....	4
2.2	Components Suitable for Replacement or Upgrade	4
2.2.1	Globus 2.4.....	4
2.2.2	eXist XML Database	5
2.2.3	GridPP Virtual Organisation Server and GSI.....	6
2.3	Upgrades and Replacements.....	6
2.3.1	Globus 4.....	6
2.3.2	gLite	7
2.3.3	Other XML Databases	8
2.4	Other Useful Components.....	8
2.4.1	SRM	8
3	Conclusions.....	9
	References	10

1 Introduction

1.1 The QCDgrid Project

The UKQCD Collaboration aims to “procure and jointly exploit computing facilities for lattice field theory calculations, whose 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 in the form of binary files. The purpose of the QCDgrid project is to provide a software application 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 a software application called *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 six UK sites at: Edinburgh (including the University of Edinburgh Advanced Computing Facility), Liverpool, RAL, Southampton, and Swansea. Glasgow are 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 with Globus Toolkit 2.4, EGEE, and an XML Database Server (XDS)) 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.

QCDgrid allows the user to query and manipulate associated metadata using a *Metadata Catalogue Browser*. The software provides a *Job Submission System* that allows a user to schedule computations on remote HPC systems, from the comfort of their desktop computer. Security is leveraged from the Globus Toolkit, based on digital certificates issued by the UK e-Science Certificate Authority. The result is a reliable, secure data management system.

Looking to the future, the collaboration aims to integrate the UKQCD Grid with similar activities in the International Lattice Data Grid (ILDG) [19], allowing like-minded scientists around the world to share their data and benefit from the scientific progress of other groups. The multi-national data grid will be built on web service technologies. Within the project, a specification will be defined to allow national resources (such as the UKQCD Grid) to be seamlessly integrated into the ILDG. This specification will expose the search and retrieve functionality of the grid but not the job submission aspects.

1.2 QCDgrid Software Reviews

The QCDgrid project work plan allocates effort to carry out two reviews of the QCDgrid software suite. The reviews will consider mechanisms by which the existing QCDgrid software could be enhanced. Specifically, the reviews will consider potential upgrades or replacements to existing components of the QCDgrid software. The conclusions drawn from these two reviews will be used to define a strategy for evolving the QCDgrid software over the duration of the project.

1.3 Purpose

This document contains the results of the first QCDgrid software review, conducted in March 2005. The objective of the review was to define a strategy for upgrading and evolving the QCDgrid software suite by examining a number of potential replacements for its components. At the planning stage of the work package it was decided that the following actions would be carried out:

- Enumerate the software components upon which the QCDgrid software depends;
- Identify components for which replacements or upgrades may be beneficial;

- Describe the potential benefits and problems if replacement or upgrades of each component were to take place;
- Propose a list of software components that might prove useful in extending the functionality of existing QCDgrid software, or in the development of new QCDgrid software.

This software review specifically does not consider solutions that could be considered as replacements for the entire QCDgrid suite.

2 Software review

2.1 The QCDgrid Software Suite

The QCDgrid software suite consists of the following:

- A set of command line tools for storing, searching and retrieving QCD data on the data grid;
- A graphical browser that allows a metadata catalogue to be searched and data to be retrieved;
- A set of servers and executables that facilitate these client tools.

The QCDgrid software suite depends in part upon each of the following third party components:

- The eXist open source XML database (version 0.9) [12], which depends on the Java runtime environment (JRE, version 1.4) and the Apache Tomcat Web Server [7] (currently version 5.0.25 is deployed on the production data grid; this component also relies on the JRE version 1.4);
- Globus 2.4, as deployed by the Virtual Data Toolkit 1.1.14 [8];
- The GridPP Virtual Organisation software [9].

Future work will develop the following components (see the Log Book [20] for details):

- A set of software tools for creating and manipulating QCD metadata files;
- An implementation of the web services necessary to support the activities of the International Lattice Data Grid (ILDG), which will involve the implementation of data access control functionality (See [19] for more information about the ILDG).

2.2 Components Suitable for Replacement or Upgrade

Globus 2.4/VDT, eXist (the XML database server)/Tomcat and the GridPP Virtual Organisation software are considered as potential candidates for replacement or upgrade. The reasons are discussed in the following subsections.

2.2.1 Globus 2.4

Use by QCDgrid software

Globus 2.4 is the foundation on which the core functionality of the QCDgrid software is built. Globus provides the following capabilities that we consider a minimum to support the current architecture of the QCDgrid system:

- Job management;
- Data transfer and replication management;
- Security.

The QCDgrid software uses the following Globus 2.4 components:

- **The replica location services (RLS)** – the RLS is used to maintain a mapping between Logical File Names (LFNs), which identify unique data files stored by the QCDgrid system, and the physical locations of replicas of those data files. This is crucial to the operation of the whole data grid system;
- **The Grid Resource Allocation and Management (GRAM) API**, which is used internally by the QCDgrid system to submit jobs to storage nodes that carry out certain bookkeeping tasks, such as keeping track of the amount of free disk space on each storage node. GRAM is also used by the QCDgrid job submission system;
- **GridFTP**, which is used when replicating data files and by users obtaining access to data file replicas, and is a critical component;
- **The Grid Security Infrastructure (GSI)**, which is used by all of the above components to control access to resources and data;

- Some **Globus threading routines**, which are used to increase the efficiency of certain parts of the QCDgrid software. This is not critical to the functionality of QCDgrid and could be easily removed or replaced.

Any potential replacement would have to provide job management, replica management, data transfer and a robust security infrastructure as a minimum.

Potential issues

Globus is in the process of moving to a new web services based architecture (Globus 4). Globus 4 is scheduled for release in April 2005. Dates for later releases could not be found on the globus.org site. Globus adopt a policy of supporting only the last two stable releases. This has the implication that previous versions of Globus, including 2.4, will become unsupported by the Globus Alliance (in fact, 2.4 is only implicitly supported by virtue of the fact that it is a component of the latest supported release – version 3.2).

There are several disadvantages to working with an unsupported piece of software. The software will not evolve in terms of useful features. Patches to bugs may not be provided by the Globus alliance, except in the case of critical security vulnerabilities (this was revealed in a meeting between Bill Allcock, technology co-ordinator for GridFTP with the Globus Alliance, and the EPCC QCDgrid team).

As discovered in stress testing [5], the QCDgrid suite suffers from a number of scalability issues, which may or may not be related to the components of Globus on which it relies. This could potentially be rectified by use of another middleware suite – for example, gLite (see section 2.3.2) or a newer version of Globus – although further investigation would be required to ascertain this. As described in sections 2.3.1 and 2.3.2, this would be a non-trivial upgrade requiring substantial effort to be expended.

There may be infrastructural advantages to using web service based middleware when the ILDG work it carried out. Each node on the grid would have a web service hosting environment already installed, which could result in a reduction in administrative burden.

Potential for upgrade or replacement

The most obvious way to avoid problems relating to Globus 2.4 becoming unsupported would be to upgrade to version 4.0 of the toolkit when it is released. An examination of the benefits and pitfalls of upgrading is given in section 2.3.1.

A number of projects attempt to produce middleware suitable for grid applications such as QCDgrid. Enabling Grids for E-Science (EGEE) [13] is a European Commission funded project which “aims to build on recent advances in grid technology and develop a service grid infrastructure which is available to scientists 24 hours-a-day”. EGEE is developing a middleware suite called gLite [10]. gLite consists of hardened versions of previously existing components, some of which were originally developed by the European Data Grid project [11]. gLite is examined in more detail in section 2.3.2.

2.2.2 eXist XML Database

Use by QCDgrid software

eXist is used to store metadata about the simulation data stored by the QCDgrid system. The metadata is stored as XML documents conforming to the QCDML schema [21]. These documents are queried using XPath queries, and provide a convenient means of searching for specific but unknown data sets stored on the data grid.

Potential issues

Scalability is an issue with the eXist database. The stress testing activity [5] highlighted the fact that eXist is poor at handling concurrent queries over moderate sized sets of documents. This will become an issue if the metadata catalogue is placed under heavy loads.

Potential for upgrade or replacement

This topic was touched upon in [4]. The version of eXist used on the data grid at present is 0.9. A beta of version 1.0 has been recently released. It is possible that upgrading to version 1.0 might resolve

some of the scalability issues; although further investigation would be needed to assure this would be the case. Also, it was found in the stress testing activity that swapping version 1.0 for version 0.9 would not work without some modification to the client code, due to API changes between the versions. Additionally, although there is a product roadmap on the eXist website, it does not give a clear indication of when the final 1.0 version will be released.

Alternatively, a different mechanism for storing and querying metadata could be considered. It would be essential that this mechanism be XML based, as all metadata will be stored as QCDML documents. Therefore, an XML database that supports either XPath or XQuery would be suitable. Some possibilities are considered in section 2.3.3. A commercial relational database that supports XML, such as Oracle or SQL Server, might also be appropriate, although there are cost implications. It is also possible that the metadata component of gLite could be useful here, as discussed in section 2.3.2, though as stated it is difficult to get concrete technical information about it at this time.

2.2.3 GridPP Virtual Organisation Server and GSI

Use by QCDgrid software

The security features in QCDgrid are based on Globus GSI. At present, all grid users need a certificate from the UK e-Science Certificate Authority. Certificate subjects are stored in the grid map files of every node of the grid. To ease administrative burden, the GridPP virtual organisation server [9] is used. This allows a list of authorised users certificate subjects to be stored in a central place on the grid. Grid map files on each node are periodically rebuilt from this central list.

Potential issues

The present system allows all users access to all data on the grid. It will become necessary to have some sort of access control mechanism, especially in the context of the ILDG.

Potential for upgrade or replacement

There are two realistic mechanisms of upgrading the current security infrastructure to support the desired functionality.

- Design and implement a custom solution, perhaps leveraging the extensibility features of the underlying middleware (e.g. it is possible to add authentication extensions to Globus 2.4 by writing custom code);
- Use a more flexible existing security framework (e.g. the Virtual Organisation Management System (VOMS), a component of the gLite middleware suite);

Developing a custom system carries the risk that writing security code is difficult without specialist knowledge. Using an existing, audited system may therefore be preferred.

The final decision on the best course of action is dependent on:

- Whether a decision is made to upgrade the core middleware to Globus 4 or gLite;
- The exact security requirements for the ILDG middleware, which are not final at this time.

2.3 Upgrades and Replacements

In this section some software components that could be useful in dealing with the issues outlined in section 2.2 are discussed. Advantages and disadvantages are outlined, indicating which of the issues the component might help with.

2.3.1 Globus 4

Discussion

Globus 4 offers essentially the same functionality as Globus 2, such as security, job management, and data management, but is based on a substantially different architecture. In version 4, activities are initiated and monitored by invoking operations on web services (WS). This is in contrast to version 2, where custom protocols operating over a range of different ports are used for communication between different components. Version 4 contains new versions of all major components, including GridFTP

and GRAM. The main differences between each used component of Globus and its newer version are described here.

It appears that the RLS version 4.0 API will be compatible with the RLS version 3.2 API. It is unclear from the Globus website how compatible the new RLS will be with the 2.4 API. A new service, called the Data Replication Service (see [22]), has been introduced to manage data replication activities. This service appears to have a number of features that would make it useful for coordinating replication activities within the QCDgrid software. Of particular note is the ability to monitor the state of replication activities. It is likely that the legacy RLS will be included in the 4.0 distribution, although it is unclear how long it will be supported for.

According to [23], “the GT4 release includes the Pre-WS version of GRAM ... for legacy purposes only”. A new, WS based version of GRAM is included, which consists of a “set of WSRF-compliant Web services to locate, submit, monitor, and cancel jobs on Grid computing resources” (WSRF is the Web Services Resource Framework, a set of specifications outlining a means of accessing and manipulating stateful resources via web services [24]). It is likely that the API will be substantially different from the 2.4 API. WS-GRAM is said to offer better performance, throughput and reliability than Pre-WS GRAM, which could serve to increase these characteristics in the QCDgrid system [23].

Globus 4 contains a complete reimplementaion of the GridFTP server that aims to greatly increase performance. The server has not been altered at the protocol level, so in theory would just plug-and-play in the QCDgrid system. The new server boasts several features that could be useful, such as extensive logging that could be useful when providing usage statistics etc. Globus 4 also provides a Reliable File Transfer (RFT) service, which is a WSRF service that manages multiple GridFTP transfers. This could be especially useful when moving multiple files. The new GridFTP server could be the most straightforward way to add value to the system without significant changes to it.

GSI is still present in Globus 4. However, a new security framework, Web Service Authentication and Authorization (WSAA), is the preferred security mechanism in Globus 4. This provides message level security, for ensuring that communications cannot be intercepted, and an authorisation framework. Almost no technical detail about WSAA is present on the Globus website at the time of writing. It is likely that WSAA would be able to provide the functionality required in QCDgrid, and quite possibly for the ILDG work as well.

Conclusion

Changing to the new GridFTP server on all nodes of the grid is recommended. This would be a low effort way to get an immediate performance benefit.

It is likely that rewriting all or part of the software to use Globus 4 would take some effort. It could be of some benefit, as new features provided by Globus 4 could be leveraged, and we would be using a toolkit that is likely to be supported into the future. However the amount of effort required is likely to be significant due to the large architectural differences between Globus 2 and Globus 4.

2.3.2 gLite

Discussion

gLite consists of a complete set of grid middleware with the following capabilities:

- Job and execution management;
- Data management, including replica and metadata management;
- Accounting, logging & bookkeeping;
- Information and monitoring;
- Security;
- Workload management.

Full details of the suite are available from [25]. It appears that gLite 1.0 is scheduled to be released in March 2005.

It looks likely that gLite could be used to replace Globus as the grid middleware for the QCDgrid system. It offers the basic functionality required – job management, data (and metadata) management,

and security. gLite consists of a set of web services. gLite does not embrace the WSRF standard implemented by Globus 4, but rather concentrates on interoperability by using straightforward web services and the WS-Interoperability standard. This is very much in line with the proposal to use simple web services for the ILDG work.

gLite appears to offer a flexible, standards based security model. The intention is to use WS-Security as soon as it is finalised. This may be an advantage when we come to implement the more complex security requirements of the ILDG. There could be problems when it comes to upgrading to this security system however, and further research would be required. gLite's metadata management subsystem could also be of use, although again it was difficult to find technical detail on the web site.

Conclusion

It is worth considering whether the QCDgrid system should be re-written using version 1.0 software, although by the time a re-write occurred the software should have moved on to later releases. Again, this would cost significant development effort. Further investigation is necessary before a decision can be reached.

2.3.3 Other XML Databases

Discussion

The criteria to consider when specifying a replacement or upgrade for the eXist database software are:

- Will the replacement or upgrade perform acceptably, i.e. will it help solve the scalability problems that have been experienced?
- Will it be difficult to integrate the replacement or upgrade with the existing code?

Alternative XML databases include:

- Xindice [14], which is an open source XML database;
- SleepyCat dbXML [16], which is another open source XML database;
- Tamino, which is a commercial XML database product from Software AG;
- Other commercial relational databases with XML support, such as Oracle or SQL Server.

Xindice has the advantage of supporting the XMLDB API that is used by the QCDgrid software. However, performance has been an issue in the past. SleepyCat produce a relational database which is known for its good performance. Whether this extends to the XML database is not know. SleepyCat dbXML also uses a custom API that is different from XMLDB. Tamino and other commercial offerings may offer good performance, but would have to be paid for. Additionally, Tamino does not support the XMLDB API. In fact, it appears that the XMLDB effort has stagnated somewhat, with the <http://www.xmldb.org> domain registration having lapsed. Comparative performance and scalability figures for XML databases are also difficult to find.

Conclusion

The simplest way to attempt to avoid the scalability problems with eXist is to wait until eXist 1.0 is released and then upgrade. If the problems persist, an attempt to use Xindice will be made. If the problems still persist, the other options listed above may be evaluated.

2.4 Other Useful Components

2.4.1 SRM

The Storage Resource Management working group of the GGF define an interface for accessing and manipulating storage resources in a very direct and flexible manner. The interface specification is available from [18]. SRM facilitates a number of advanced storage management functions, including space management, managing file access permissions, directory manipulation and synchronous and asynchronous file transfer. SRM has been mandated as a component of the ILDG architecture and thus will be used at some point in the QCDgrid2 project.

3 Conclusions

This software review has succeeded in identifying a number of software components that could help to evolve and improve the QCDgrid software suite. The review identified that the QCDgrid software suite could benefit from upgrade or replacement of components in three areas:

- The metadata catalogue;
- The middleware underlying the data grid;
- The security infrastructure.

The component causing most immediate concern is eXist, the database server at the heart of the metadata catalogue. The proposed course of action is to wait until version 1.0 of the server is released and to upgrade. An attempt to port the software to use the Xindice database will be made should problems persist. If problems still occur after porting to Xindice, other components as outlined in section 2.3.3 will be examined.

The middleware underlying the data grid faces long term issues with regard to support. Globus 2.4 is reaching the end of its lifetime and will cease to be officially supported. There are two possible courses of action to address the issue of support: upgrade to Globus 4.0, or use a new suite of middleware such as gLite. The proposed course of action is to defer the decision as to which is most suitable until a more thorough technical assessment can be carried out. At present, Globus 2.4 provides all the necessary functionality required for the data grid, which means that there are no technical pressures to make a quick decision. Additionally, giving the grid community time to assess and feed back on gLite and Globus 4.0 will allow a more informed decision to be made – deferring the decision can be seen as an advantage for that reason. Effort will become available later in the project in order to carry out this process.

The security infrastructure provided by Globus 2.4 and the GridPP Virtual Organisation software is sufficient for present requirements. As the project progresses and the ILDG work is carried out, it will become necessary to review this security infrastructure. It is entirely preferable to use a pre-existing security framework rather than a custom solution. The decision as to which to use is dependent on the choice of middleware and will be carried out at a later date.

References

- [1] UKQCD Collaboration home page <http://www.ph.ed.ac.uk/ukqcd/>
- [2] Setting Up Systems for QCDgrid, James Perry. Available from <http://www.epcc.ed.ac.uk/~jamesp/qcdsetup.pdf>
- [3] UKQCD collaboration, "GridPP Log Book", available on request from qcdgrid-enquiries@epcc.ed.ac.uk (September 2004).
- [4] QCDgrid project, "QCDgrid System Risk Analysis", available on request from qcdgrid-enquiries@epcc.ed.ac.uk (October 2004).
- [5] QCDgrid project, "QCDgrid Load and Stress Testing Results", available on request from qcdgrid-enquiries@epcc.ed.ac.uk (October 2004).
- [6] The Web Services Resource Framework <http://www.globus.org/wsrfl/>
- [7] The Apache Tomcat web server, see <http://jakarta.apache.org/tomcat>
- [8] The Virtual Data Toolkit, available from <http://www.cs.wisc.edu/vdt/index.html>
- [9] The GridPP Virtual organisation server, see <http://www.gridpp.ac.uk/vo/>
- [10] gLite, <http://glite.web.cern.ch/glite/>
- [11] The European Data Grid project, <http://eu-datagrid.web.cern.ch/eu-datagrid/>
- [12] eXist <http://exist.sourceforge.net/>
- [13] The EGEE project, see <http://egee-intranet.web.cern.ch/egee-intranet/gateway.html>
- [14] XIncede, part of the Apache Project, available from <http://xml.apache.org/xindice/>
- [15] Tamino, a product of SoftwareAG. See <http://www2.softwareag.com/corporate/products/tamino/default.asp> for details.
- [16] SleepyCat <http://www.sleepycat.com/>
- [17] Anand Vivek Srivastava, "Comparison and benchmarking of XML databases", <http://www.cse.iitk.ac.in/report-repository/2004/Y1043.pdf>
- [18] SRM Working Group, <http://sdm.lbl.gov/srm-wg/>
- [19] ildgWiki Homepage, <http://www.lqcd.org/ildg/tiki-index.php>
- [20] UKQCD Log Book http://www.gridpp.ac.uk/qcdgrid/documents/UKQCD_LogBook.pdf
- [21] QCDML 1.1 http://www.ph.ed.ac.uk/ukqcd/community/the_grid/QCDml1.1/
- [22] GT 4.0 Component Fact Sheet: Data Replication Service <http://www-unix.globus.org/toolkit/docs/development/4.0-drafts/techpreview/datarep/DataRepFacts.html>
- [23] GT 4.0 Pre-WS GRAM (GRAM2) <http://www-unix.globus.org/toolkit/docs/development/4.0-drafts/execution/prewsgram/index.html>
- [24] WSRF The Web Services Resource Framework <http://www.globus.org/wsrfl/>
- [25] EGEE gLite <http://glite.web.cern.ch/glite/>