e-HTPX – HPC, Grid and Web-Portal Technologies in High Throughput Protein Crystallography

Rob Allan (r.j.allan@dl.ac.uk), Ronan Keegan (r.m.keegan@dl.ac.uk), **David Meredith (d.j.meredith@dl.ac.uk)**, Martyn Winn (m.d.winn@dl.ac.uk), Graeme Winter (g.winter@dl.ac.uk), CCLRC Daresbury Laboratory
Jonathan Diprose (jon@ strubi.ox.ac.uk), Chris Mayo (chris.mayo@strubi.ox.ac.uk), University of Oxford, The Welcome Trust Centre of Human Genetics, Oxford
Ludovic Launer (launer@embl-grenoble.fr), MRC France, ESRF, Grenoble
Joel Fillon (fillon@ebi.ac.uk), European Bioinformatics Institute, Cambridge
Paul Young (pyoung@ysbl.york.ac.uk), York Structural Biology Laboratory

Abstract

We present details of work being carried out to increase throughput of protein crystallographic structure determination through the use of Grid enabled web-portal technologies, from which users can remotely plan and direct experiments. The project integrates a number of key services provided by leading UK e-Science, protein manufacture and synchrotron laboratories. A comprehensive data model has been developed in the project to a) allow information exchange and communication between multiple sites using a combination of web-service and Grid technologies, and b) to facilitate independent development and implementation of remote services. The project is particularly well suited to the web-portal model, where laboratory specific services are made accessible via the main e-HTPX web portal interface. Of particular importance to the project is the automation, speedup and monitoring of data collection and processing as rapid feedback and the ability to make 'on the fly' decisions regarding the quality of data collected on a synchrotron beam-line is essential for the efficient operation of high-throughput systems. These requirements are well suited for the application of High Performance Computing, Grid and portal technologies. The project provides additional e-Science related challenges including interfacing with robotic hardware and transferal and monitoring of physical protein samples between different sites.

1.0 Introduction

The e-HTPX project provides the distributed computing infrastructure required for the workflow illustrated in Figure 1. The project enables users to remotely plan experiments and manage the flow of data, from the initial stages of on-line protein-crystal target selection (stage 1), to the automated deposition of the final refined (digital) protein model into public databases (stage 6). Predictably, implementation of this work flow depends heavily upon a range of e-Science technologies. The e-HTPX web-service hub / Grid-portal (referred to singularly as ‘portal’ from here on in), provides a single point of access for the remote user to access this workflow and to the remote sites involved in the project. Implementation of the work flow requires a range of e-Science disciplines including, provision of service portals for both ‘generic’ Grid services and application-specific portals for protein crystallography, HPC services and expert systems. This paper focuses upon the current efforts made regarding implementation of this workflow, and upon the e-Science related hardware/software issues and challenges met.

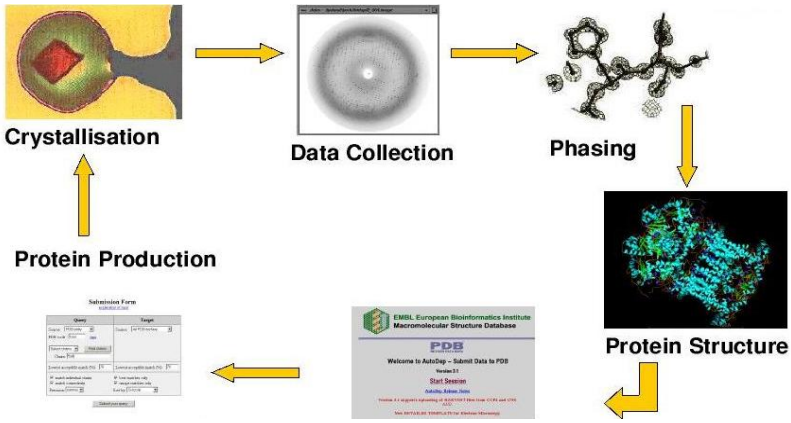

Figure 1. Protein crystallography workflow. Stage 1 – Protein Production (target crystal selection), Stage 2 – Crystallisation, Stage 3 – Data Collection, Stage 4 – Phasing (data processing), Stage 5 – Solution of digital protein structure model, Stage 6 – Submission of protein model into public database.

2.0 Workflow Overview

The initial stages of the work-flow (stages 1 and 2), are centered on project planning and initiation. This involves on-line completion and submission of requests to protein production facilities for the growth of specific protein crystals. This stage also includes remote monitoring of the progress of crystal-growth and of the delivery of crystals to a specified synchrotron. Inevitably, these stages require numerous communications between the remote portal user and the various e-HTPX laboratories involved in the project. In order to simplify these complex communications, the portal has been designed to centralize the requests and responses made between the portal web-service hub and the various e-HTPX laboratories. The web-services communicate via a comprehensive data model which standardizes the format of each communication (refer to section 3.2).

Following crystal delivery to a specified synchrotron, stage 3 involves remote access via the portal to automated data collection systems (e.g. DNA expert system [1]). When data collection is completed, data transfer to Grid storage resources (e.g. National Grid Service cluster machines [2]) is implemented using Grid middleware. Data processing by HPC services are also provided (stage 4) with the use of job-submission portals. These stages utilize a combination of web-services and Grid middleware technologies.

After post-data collection processing, information from academic projects will eventually be deposited into public databases (stage 5 and 6), such as the Macromolecular Structure Database provided by the European Bioinformatics Institute (EBI).

3.0 e-Science in e-HTPX

3.1 e-HTPX Portal / Hub (The Software Stack)

The portal consists of a software stack commonly implemented by Grid enabled web-portals. This includes a user interface implemented with Java JSP and Servlet technologies [3], hosted by the Tomcat application web server [4] available from the Jakarta Apache project [5], and a combination of Grid and web-service (SOAP) middleware protocols. The chosen version of Grid middleware is currently Globus version 2.4 [6]. This has been chosen mainly for its stability that is necessary for production services, but also due to the currently changing standards in Grid middleware. Despite this, moves to GT4 and

newly developing technologies such as the Web Service Resource Framework [7] (WSRF) will be undertaken in the future. The java Commodity Grid (CoG) toolkit [8] provides the functionality for Grid operations. A set of re-usable e-HTPX Java Beans have been developed using the CoG kit to encapsulate both 'generic' Grid operations (e.g. resource monitoring, Grid-FTP, job submission and monitoring), and operations specific to e-HTPX. Java Beans technology is particularly useful for portal development since they may be configured to have different 'scope' within an application. Examples include the 'ehtpxUser' Bean, which is held in 'session' scope and used to store the users credentials, and a job monitoring bean, which is held in 'application' scope, enabling continuous job-monitoring even after the user has logged off from the portal.

Security and data-confidentiality is of significant importance, as the services will be made available to both industrial users and to academia. Authentication and authorization is handled through the use of personal X.509 v3 digital certificates issued by the UK Certification Authority (CA) [9]. For international users, additional CA's in which the UK e-Science CA has agreements may also be used. The portal supports the Grid Security Infrastructure (GSI) [10] and a MyProxy [11] login mechanism. Use of a MyProxy server provides a secure online repository in-which users can store their temporary credentials, created and delegated from their personal workstations. When logging onto the portal, and when other software components linked to the portal require authentication to Grid resources, the users temporary credentials can be retrieved as required from the server. Figure 2 shows the users home page which provides information on the temporary Grid-proxy certificate delegated from a MyProxy server.

A modified version of the Java CoG kit 'proxy creation and delegation tool' has been provided in the form of a Java Web-Start application [12]. This enables users to simply run the tool by clicking on the link on the portals home page and provides a means to delegate their credentials to the MyProxy server. Following proxy delegation, the user can benefit from single-sign-on to multiple Grid resources on which the user has accounts. This includes e-HTPX specific services and 'generic' Grid services. Figure 3 shows a screen-capture of the interface used to monitor the current status of Globus GRAM job-submission managers and GSI FTP servers on which the user has accounts.

Figure 2. User home page showing Grid-proxy details. Temporary Grid proxy certificates are delegated from a MyProxy server when logging onto the portal.

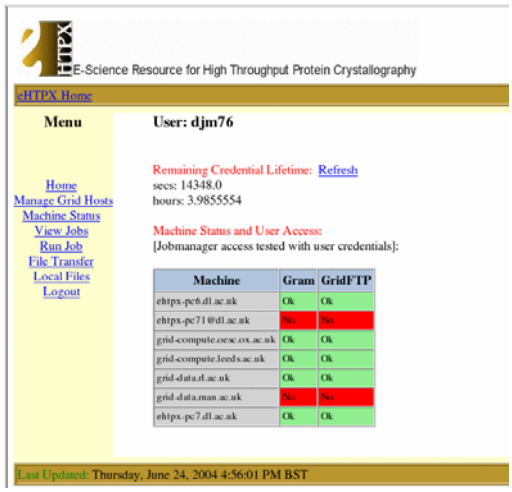

Figure 3. Users can test the status of Grid resources on which they have accounts. Here, the status of GRAM job managers and Grid FTP hosts provided by the National Grid Service [2] (NGS) and e-HTPX are tested.

Client-side software requirements include, a) an HTTP(S) enabled web-browser with cookies to allow session data to be transferred between the client and the application server, b) possession of a personal X.509 v3 digital certificate, and c) Java run-time environment, required to launch Java Web Start applications (provided as standard with Java version 1.4+).

3.2 Remote Project Initiation, Protein Crystallization, Tracking Crystals to the Synchrotron (Web-services and Data Model)

As previously commented upon, the portal centralizes numerous web-service requests and responses made between the user and the various e-HTPX laboratories. This is especially relevant to stages 1 and 2 of the workflow. This includes on-line completion of safety forms, specification of crystal growth conditions, remote authorization of necessary permissions and allocation of sufficient beam-time. Figure 4 shows how the web-service communication call-stack is presented to the user, while Figure 5 illustrates the complete sequence of calls implemented in the process.

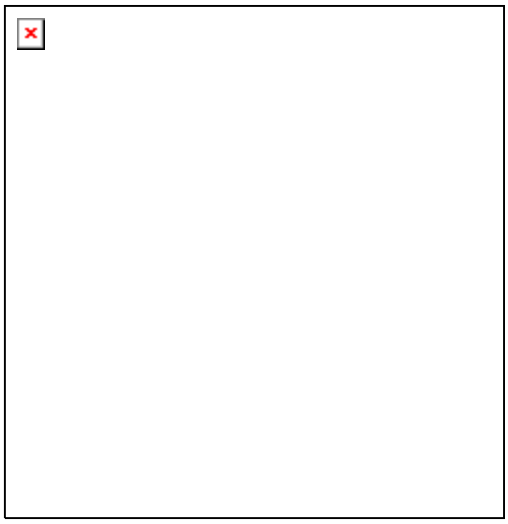

Figure 4. Interface to show the web-service communication call stack between the different laboratories involved in the project. The status of each call is automatically updated and presented to the user.

The web-service calls communicate via a comprehensive data model which standardizes the format of each communication, and provides a common data inter-change language between the remote services. The protein production data model (PPDM) is defined in UML and implemented in XML schema. The use of web-services also facilitates the independent development and implementation of each remote service.

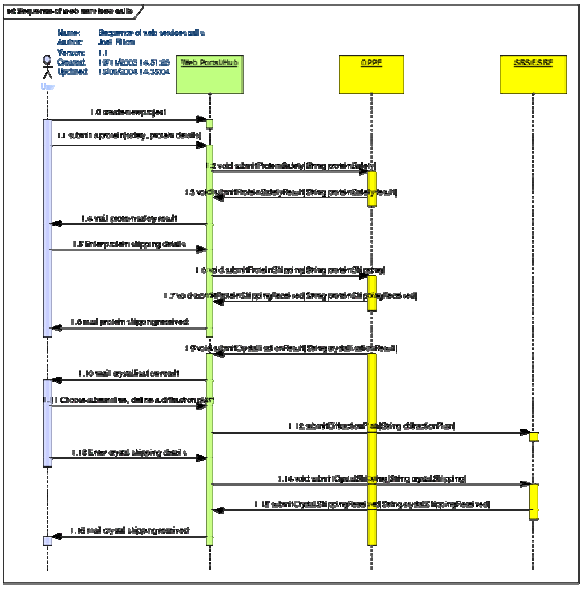

Figure 5. Complete sequence of web-service calls implemented in workflow stage 1. Protein manufacture and crystallization are undertaken at the Oxford Protein Production Facility (OPPF), where web-services have been developed for users to request images from the automatic crystallization facilities to view remotely. The shipping of crystals and transfer of database-information to synchrotron (X-ray) beam-line facilities for data collection (European Synchrotron Radiation Facility, CCLRC Daresbury Laboratory), is remotely monitored and tracked (e.g. with bar codes) from the e-HTPX portal.

3.3 Remote Data Collection and Transfer (Expert Systems and Grid Middleware)

Upon delivery of the crystal to the synchrotron, automated data collection systems (e.g. DNA [1]) are used to collect X-Ray diffraction data. Remote access will be provided via the portal, where the user may specify experimental requirements, objectives and the amount of processing required. A key requirement of the portal that is catered for by Grid middleware technology, relates to the remote management and transfer of large volumes of data produced during data collection. A single experiment typically produces in excess of 5 GB of data consisting of multiple high resolution image files (~20-50 MB each). These data need to be securely transferred from the data collection machine located on a synchrotron beam-line, to either the e-HTPX HPC data processing cluster (refer to section 3.4), or a user's external Grid data-storage resource. The e-HTPX portal will implement three interfaces for transferal and management of data depending on user requirements.

a) Grid-FTP Enable Data Collection Software
The automated data collection systems may be configured to transfer data via Grid-FTP as it is being generated 'on-the-fly' to either the user's external Grid-data storage resource, or more typically, to the e-HTPX HPC data processing cluster. In order to initiate this remote 3rd party data transfer, the user's temporary Grid-proxy-certificate is automatically delegated to the data collection software from either the portal-server using session-scope web-services, or from the MyProxy server. Consequently, this gives the data collection system the necessary credentials to authenticate to the users Grid-data storage resources or HPC cluster on behalf of the user.

b) Provision of Manual Grid-FTP GUI Tools
In some scenarios, the user may need to be present when collecting data on the beam-line, providing the option to manually choose and transfer specific data. A modified version of the Java CoG kit [8] file transfer GUI tool has therefore been made available as a Java Web-Start application. The link to the file transfer application may only be accessed by authenticated on-line portal users. The application is supplied to the remote machine in digitally signed jar files. This grants the application user-access to the disk-drive of the machine running the file-transfer GUI. On startup of the file-transfer software, the users Grid-proxy certificate is delegated to the application allowing the user to authenticate to other resources and transfer data.

c) Web-Interface to Grid-FTP
A web-based portal interface to 3rd party Grid-FTP data transfer has also been implemented in the portal, from which the user can modify and navigate the file structures of remote Grid hosts. This is illustrated in Figure 6. Data may also be downloaded from, and uploaded to any of the users Grid-host accounts via HTTP(S) (and by the Web Start file transfer tool using Grid FTP protocols provided the user has no restrictions on port 2811).

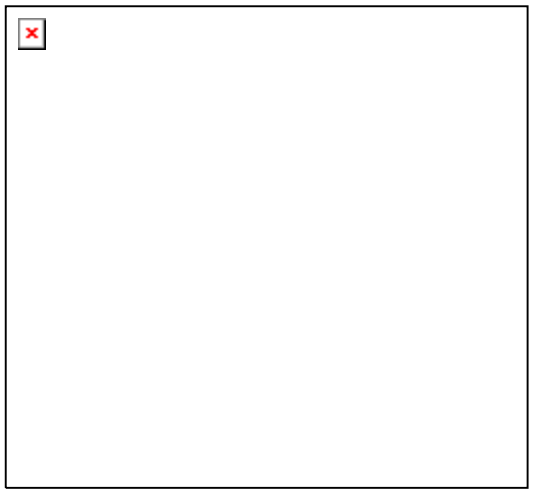

Figure 6. Web-Interface to Grid-FTP. Users can navigate and modify the file hierarchy of remote Grid hosts and make 3rd party file transfers.

3.4 Data Processing (HPC and Grid Middleware)

A dedicated compute cluster, consisting of 9 dual Xeon processors (18 cpu's) and a Gigabit link interconnect has been made available for data processing. The cluster will host a popular suite of protein crystallography codes provided by the CCP4 [13] distribution. To increase high-throughput, we have developed parallelised versions of some of the key codes, including a parallel version of the molecular replacement code (Beast). The cluster facilitates job submission via the Globus GRAM job-submission manager which interfaces with Condor [14] job farming tools.

In many cases, submission of jobs using data collected on the beam-line will be automated by the data collection software, which may act on the users behalf due to Grid proxy delegation (refer to section 3.3). In addition, the portal also provides an interface for remote job submission and status monitoring, enabling the user to upload and transfer required data to the job manager, and to review the job queue.

4.0 Challenges Met

To date, the main challenge encountered in the project relates to a method of secure access from the portal web-server to e-HTPX specific Grid service machines that are required to be situated behind an institutions internal firewall (as dictated by site-policy, e.g. access to beam-line machines to monitor data gathering). These Grid hosts require both freely outgoing and incoming connections via the main Grid-Service ports and an ephemeral port range. A solution to this problem that is currently under investigation is to implement IP recognition by firewalls, thus allowing access to the Grid host machines located behind an institutions firewall to be from the portal server alone. In doing this, security and authentication can be focused and maximized onto the single 'entry-point' provided by the portal web server.

5.0 Conclusions

The e-HTPX web-service hub / Grid portal integrates a number of e-Science technologies needed to realize the high-throughput protein crystallography workflow in Figure 1. Authentication and authorization is facilitated by the Globus GSI security mechanism which implements the use of personal X.509 v3 digital certificates. A MyProxy server is used to log onto the portal. The complex sequence of web-service calls is simplified by the portal which centralizes the requests and responses between the remote user and the various laboratories associated with the e-HTPX project. A comprehensive data model (PPDM) is implemented in XML schema and provides a common data-interchange language for the web-services. Remote data collection is facilitated by automated expert systems, which implement Grid-proxy delegation and can therefore act on the users behalf. This enables data collection software to transfer large volumes of data via Grid middleware protocols to Grid data storage resources and/or HPC data processing nodes. The user may remotely monitor data collection and data processing via the portal interface. The main challenged faced by the e-HTPX project concerns firewall restriction policies of the laboratories involved with the project.

6.0 References

[1] DNA Project, http://www.dna.ac.uk/
[2] NGS: The UK's core production computational and data grid, http://www.ngs.ac.uk
[3] Java Servlet 2.3 and Java Server Pages 1.2 Specifications, http://java.sun.com/products/servlets
[4] Tomcat Open-Source Servlet Container, http://jakarta.apache.org/tomcat
[5] Jakarta Apache Project, http://jakarta.apache.org
[6] Foster, I. and Kesselman, C. (1997) Globus: a metacomputing infrastructure toolkit. International Journal of Supercomputer Applications.
[7] Globus Alliance Web Service Resource Framework, http://www.globus.org/wsrf/
[8] Laszewski, G., Foster, I. and Gawor, J. (2000) CoG kits: a bridge between commodity distributed computing and high-performance grids. Proceedings of the ACM java Grande Conference, 2000
[9] Grid Support Centre: http://www.Grid-support.ac.uk/
[10] GSI Software Information, http://www.globus.org
[11] Novotny, J., Tuecke, S. and Welch, V. (2001) An online credential repository for the grid: MyProxy. Proc. 10th IEEE Symposium On High Performance Distributed Computing
[12] Java Web Start Technology, http://java.sun.com/products/javawebstart/.
[13] CCP4 – Collaborative Computational Project Number 4, http://www.ccp4.ac.uk/main.html
[14] The Condor Project, http://www.cs.wisc.edu/condor/