Integrative Biology VRE

Work Package 5: IB Integration Report

Lakshmi Sastry

Intended Audience

Anyone.

Issued: 31 Mar 2007
1 Introduction

This is a final report on the implementation and delivery of a robust job submission service that unifies the way simulation and analysis tasks are composed and submitted to the National grid Service (NGS) including HPCx and local clusters. The job submission includes a simple workflow and job monitoring together with data management through the Storage Resource Broker (SRB) on the NGS.

During this short project, CCLRC is merged with PPARC and a new Council Science & Technology Facilities Council was formed. This is for the information of the readers who needs to refer to the original proposal which was submitted in the former name.

1.1 Background

Over the past year, the main EPSRC funded Integrative Biology (IB) project and the JISC funded IBVRE project both have developed job submission mechanisms for their software that focussed on serving their intended audience as first priority. As a result, the IB job submission used Globus protocols to submit and manage jobs on the grid and used SRB as the data management infrastructure directly. IBVRE focussed on portals technology and more importantly submitting jobs to local clusters and used PostgreSQL database for data management. The technical teams identified that a unified robust mechanism could be adopted by both projects for job submission and management to support the wider user community and to encourage the use of metadata.

2 Architecture

Figure 1 below describes the overview architecture of the IB/IBVRE job submission service.

The above architecture is best explained by a step by step usecase of a scientist using the service:
1. User sets up the experiment on IB VRE
2. The experiment setup files are uploaded onto SRB
3. Gaptk job submission and monitoring services are invoked
4. A Proxy Grid Certificate is downloaded from MyProxyServer
5. Job is submitted onto NGS or HPCx
6. Setup files are downloaded onto compute resource and jobs are executed
7. Results are uploaded onto SRB
8. The results can be accessed by The GAPtk-IB visualisation services to create visualizations of results which can be streamed to the user client.
9. Alternatively, IB VRE server downloads the result files.

The next Section describes the various software modules that implement this architecture.

3 Software description

3.1 IBRun script
IBRun is a Linux script developed by STFC as part of the Integrative Biology project. It takes a string which contains information such as the application executable name, pointers to input and output files and directories in the SRB, the compute resource and the number of processors to be used for the application. The string may contain as many or as few parameters beyond a set of essential ones. The script can be directly invoked by a user from a command line or through portals, portlets and other standalone Graphical user Interfaces. The IBRun on its own required Globus based communication protocols to execute and hence posed a formidable difficulty for general use, as the users will require their systems to have particular ports to be opened.

The Applications Group have been developing various different ways of submitting jobs to the Grid, including Web Service wrappers to encase scripts and applications that used Globus protocols. Under the Integrative Biology project, we combined the GODIVA grid job submission service with IBRun to create a robust server side mechanism for generic job submission. This methodology is discussed in detail in the next Subsection.

3.2 Job submission service
Job submission is a Web Service. It can be used only for GT2jobType. When invoked it returns a string either a jobID to indicate successful job submission or an error message. Each invocation of job submission is associated with a new job instance, described by user data such as username and password and other job specific details contained in the job submission schema.

Once a successful job submission is made, a fresh job record is created on the job monitoring database with a unique record ID which is returned to the user. This unique handle is used to query or cancel the job.

3.3 Job monitoring service
Job monitoring is a daemon process, which continuously polls the job monitoring database to retrieve a list of jobs. The status of any job is one of submitted, pending, active, unknown and unsubmitted. The job monitor method loops over a list of job records by record identifier. It updates the job monitoring schema table in the database.

3.4 Job query and/or cancel
The users requests to query the status of one or more jobs or cancel these are also implemented as service. The query is a straightforward enquiry handle by the database.
3.5 Security

The Grid Security Infrastructure is used via the grid certificate and the use of myProxy server (http://grid.ncsa.uiuc.edu/myproxy/license.html). If and wherever the password is stored encryption is used.

3.6 IBVRE client

The IB job submission services are interfaced to the IBVRE infrastructure through the GridClusterProxy class. This class extended the existing IBVRE infrastructure to enable more structured data management through the use of Storage Resource Broker (http://www.sdsc.edu/srb/index.php/Main_Page) on the National Grid Service. It enables the user to upload input files and create directories to store output data in SRB.

The IBVRE configuration was amended to include MyProxyServer details, SRB configuration and details and URIs for the GAPtk services. The IBVRE database is extended to include details of NGS clusters including HPCx.

For the IBVRE pilot, MemfemGridExecution class implements an interface to the GAPtk services to submit a set of jobs to NGS to run the parallel version of memfem and update the IBVRE database with results.

The IBVRE client classes are invoked by the user from the IBVRE portal interface.

4 Conclusion

The STFC e-Science Centre undertook to interface the Integrative Biology job management service to the IBVRE infrastructure so that the IBVRE users could exploit the resources available on the UK National Grid Service, including HPCx, including Storage Resource Broker. The job management services comprises of the job submission and job monitoring components. A client to the services have also been developed and delivered. A parallel version of an application code memfem was compiled on NGS cluster to demonstrate the extended functionality. The infrastructure is capable of supporting other applications as long as they are available for the NGS cluster architectures. The integrated services have been delivered in time for the IB Showcase held on 27th March 2007.