eMaterials: Integrating Grid Computation and Data Management Services

Lisa Blanshard l.j.blanshard@dl.ac.uk

Rik Tyer r.p.tyer@dl.ac.uk

Kerstin Kleese van Dam k.kleese-van-dam@dl.ac.uk

@ CCLRC Daresbury Laboratory, Warrington, UK

Abstract

CCLRC is involved in the development of grid and data management tools for the Simulation of Complex Materials e-Science project [1] otherwise known as *eMaterials*. The aim of the project is to bring together computer and computational scientists to exploit new and existing technologies for key current areas of materials science. In particular the scientific aims of the project relate to the development of *combinatorial materials chemistry*, with specific applications to catalysis and ceramics and the *prediction of polymorphs* of organic-pharmaceutical compounds and their properties.

Currently scientific data is distributed across a multitude of sites and scientists have very limited support for accessing, managing and transferring data. In a grid environment that spans numerous sites and organisations, it is essential to ease and automate many of these data management processes.

To this end workflow applications which automate both the simulations processes and the associated data/metadata management have been developed. The scientific models of interest to the project often involve elaborate workflows or extensive parameter space searches. These processes have been automated using Business Process Execution Language (BPEL) and associated tools developed by the UCL node of the project. The scientific FORTRAN codes of interest to the project have been wrapped and exposed as services in order to integrate them into this framework.

The use of this workflow engine has dramatically facilitated and automated the generation of scientific data which then needs to be handled efficiently by the data management infrastructure developed by CCLRC for the project. Given the large number of data files generated by the workflow engine it is essential that they are efficiently stored and annotated with appropriate metadata describing their context and method of generation.

The data/metadata management tools have been integrated with the workflow tools using web services technology. This integration is essential to allow the automatic collection and publishing of metadata relating to the simulations along with efficiently handling the data files themselves.

A typical use case exploiting this integration would be:

1. User logs on to the Workflow Manager and selects a sequence of applications to run and also links the workflow to the study
2. The Workflow Manager sends the jobs to the grid engine
3. When each job ends the Workflow Manager authenticates, creates appropriate directories in SRB and uploads the files.

This paper will detail the integration of the file management functionality within the workflow BPEL framework. The automation of the metadata capture and publishing will be described along with the web service interfaces and functionality required.

Project requirements

Computational requirements

During the last decade, Professor Sally Price's group at UCL Chemistry department has developed a computational approach for predicting the crystal structures of small, rigid, organic molecules. [3] However each search involves running large iterations of computationally expensive calculations and currently takes a few months to perform. Studies on larger molecules, which are more typical of manufactured organic materials, are not feasible using the existing computing infrastructure.

The project scientists use a number of simulation programs. These are a combination of commercial codes (Gaussian 98, Cerius2) and open-source (Molden, Molpak, Dmarel) plus others. For example, one of the steps to predict polymorphs from a molecular formula is to use Gaussian 98 to calculate the molecular properties such as density and population.

The applications can take a number of days to run and many runs are completed in sequence to ascertain scientific results.

Data management requirements

The scientists generate a large amount of files while running simulations and typically they have been stored on individual's machines or on the machines that are used for simulations. This has made access to the data within and outside the group very difficult. In addition, securing access to the data is complex due to the number of machines involved. There is a high risk of data loss as few backups are taken and this is disorganized.

To improve the situation a number of requirements have been established as a result of consultation with the scientists.

The three main aspects for data management concern management of *files*, *metadata* and *data.*

File Management includes storage of simulation input and output on a number of servers, provision of *interface(s)* for the scientists to manage their own files, *transfer* tools, functionality to share their files with scientists at different locations and publishing to the wider community.

Progress so far

The project has been running for a year and we have made significant progress in the areas of computation and data storage. These tools are currently in use by the project scientists.

Computational chemistry on the grid

Making use of early implementations of the OGSA specification the UCL team have wrapped the Fortran binaries into OGSI-compliant service interfaces to expose the existing scientific application as a set of loosely coupled web services. The OGSA implementation facilitates the distribution of such applications across a large network, radically improving performance of the system through parallel CPU capacity, coordinated resource management and automation of the computational process. A computational workflow service enables users to distribute and manage parts of the computational process across different clusters and administrative domains. The web service coordination language, Business Process Execution Language (BPEL) makes such workflow services easily configurable. The aim is to provide services for running applications across a grid that scientists can configure themselves using relatively user-friendly languages such as XSLT and BPEL.

The reason for using BPEL for workflow management is that we do not necessarily know how the computational process will develop in the future and we are trying to give the scientists freedom to alter their methodology. Our solution prevents software from dictating their agenda.

Managing scientific data

A number of middleware tools have been developed or installed in the areas of file

management and metadata management. These are the Storage Resource Broker (SRB) [7] for file management developed by SDSC and the Data Portal [5] developed by CCLRC. A relational database [6] housed at CCLRC is used to store metadata and a Metadata Editor has been developed.

Uploading files automatically from computation to storage

Currently the scientist must manually collate the results of computation and then upload them manually to the SRB. Obviously this is not ideal. The next step is to integrate this process so that the data files are stored automatically in some predefined directory structure in the users home directory in the SRB. The following requirements have arisen from discussions:

- the workflow manager will upload files at the end of each calculation
- new directories will be created in the relevant user's home directory on SRB to store the files
- a security mechanism must be in place so that only designated programs/users may upload files
- the above functionality must be accessible from a number of different heterogeneous machines in the various clusters on the grid as this is where the applications would run.

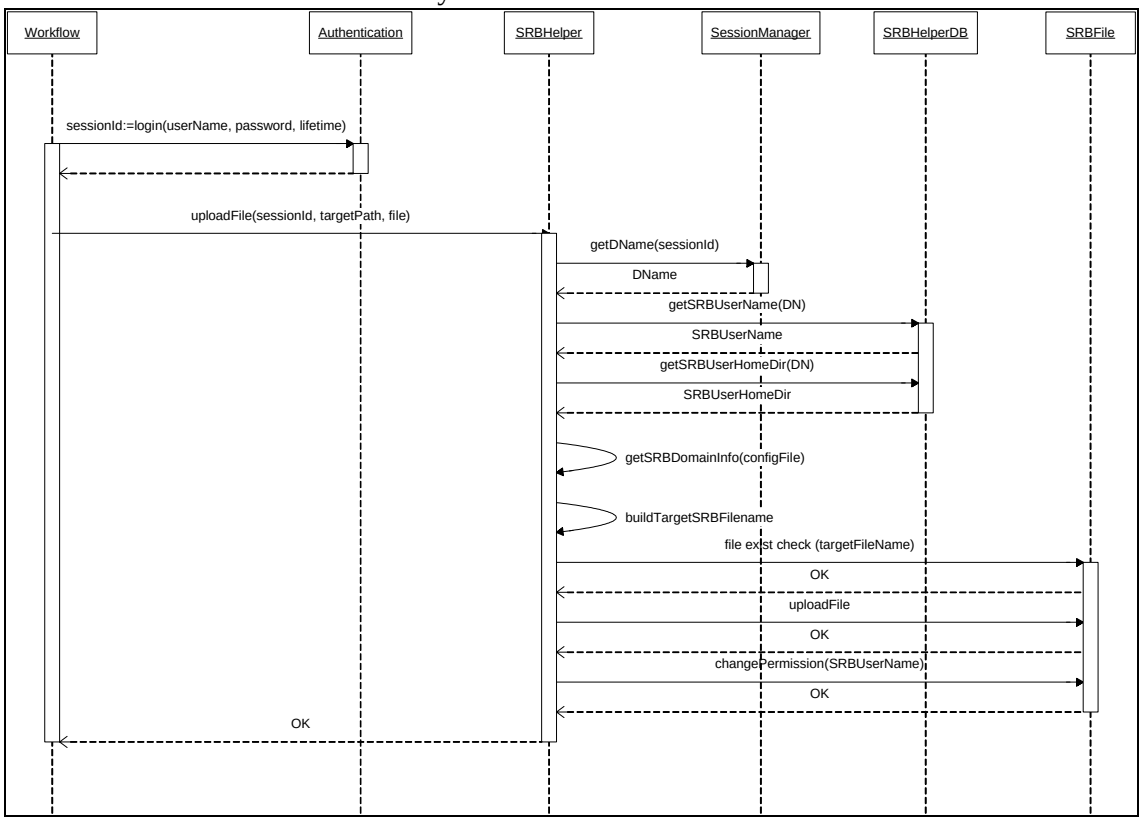

Figure 1 "Upload file to SRB" sequence diagram

To meet these requirements an SRB Helper web service will be provided by CCLRC:

- *SRB helper web service* - used to upload a single file to the SRB to a chosen directory in the user's home; also to create new directories in SRB
- *SRB Helper local database* – to store the name of the user's home directory in SRB

Since the compute services do not maintain conversational (session) state, the workflow would need to provide authentication details as part of each request to upload files or create a directory and hence degrade

performance. To alleviate this we will use two other services that have been developed as part of the Data Portal:

- *authentication web service* – uses x.509 certificates issued by the UK e-Science Certificate Authority to verify the identify of a user and returns a session identifier
- *session manager web service* – has access to a local database of sessions and associated *distinguished names* i.e. the user name and *lifetime left* i.e. the number of hours left on the certificate

The process of uploading files is as follows:

1. User logs on to the Workflow Manager and selects a sequence of applications to run
2. The Workflow Manager sends the jobs to the grid engine
3. When each job ends the Workflow Manager authenticates, creates appropriate directories in SRB and uploads the files.

Figure 1 shows the interaction between the Workflow Manager and the various web services during a file upload. Note that further files may be uploaded until the session has expired. The Workflow Manager then needs to re-authenticate with the user's details.

References

[1] Simulation of Complex Materials e-science project http://www.e-science.clrc.ac.uk/web/projects/complexmaterials

[2] Overview of Polymorph Prediction http://www.accelrys.com/cerius2/polymorph.html

[3] Prof Sally Price (UCL) - Research Pages http://www.ucl.ac.uk/~ucca17p/home_all.html

[4] CCLRC Scientific Metadata Format http://www-dienst.rl.ac.uk/library/2002/tr/dltr-2002001.pdf

[5] CCLRC Data Portal http://www.e-science.clrc.ac.uk/web/projects/dataportal

[6] CCLRC Database Services http://www.e-science.clrc.ac.uk/web/projects/database_service

[7] Storage Resource Broker http://www.npaci.edu/DICE/SRB/

[8] Implementing and using SRB, Proc UK e-Science All Hands Meeting 2003, © EPSRC Sept 2003, ISBN 1-904425-11-9 http://www.nesc.ac.uk/events/ahm2003/AHMCD/ahm_proceedings_2003.pdf

[9] Chemical Markup Language http://www.xml-cml.org/