A Middleware Agnostic Infrastructure for Neuro-imaging Analysis

Yasir Mehmood, Irfan Habib, Peter Bloodsworth, Ashiq Anjum, Tom Lansdale, Richard McClatchey & The neuGRID Consortium
Center for Complex Cooperative Systems (CCCS), Univeristy of the West of England, Bristol
Yasir.mehmood@uwe.ac.uk,{ Irfan.Habib, Peter.Bloodsworth, Ashiq.Anjum, Tom.Lansdale, Richard.McClatchey} @ cern.ch
(Please see the acknowledgements section for further details regarding neuGRID partners).

Abstract

Large-scale neuroscience research projects are necessary in order to make significant progress in the study of degenerative brain diseases. At present the effectiveness of such efforts is being somewhat restricted by the absence of specifically tailored computing infrastructures. The neuGRID project aims to address this through the provision of a high-level service oriented infrastructure that enables complex neuro-science research. One of the principle aims of this work is to develop portable services that can be re-used in a larger set of related medical applications to access distributed computing resources. These services will provide high-level functionality that will support workflow authoring and planning, provenance storage and retrieval, querying against heterogeneous data sources as well as security and data anonymization amongst others. This paper introduces the neuGRID service architecture and outlines the design of two specific services, namely the Pipeline Service and the Glueing Service. A proof of concept implementation to evaluate the neuGRID design approach has been developed.

1. Introduction

In recent years medical science has benefited from a huge increase in both the quantity and quality of research data. This has increased the need for the synthesis and processing of large amounts of information. In order to adhere to such requirements the medical domain is increasingly moving towards large-scale distributed computing infrastructures. These infrastructures, including Grids, enable scientists to find more efficient ways of accessing information from diverse data sources and to carry out coordinated research efforts that span multiple institutions. Numerous service frameworks in the domain have been built to facilitate the integration of data and the analysis, querying and collaboration over such infrastructures. Most of these service frameworks[1][2] are Grid based and have been developed for a particular community of medical users. However, the services built by one community cannot be easily shared and re-used in other medical domains due to architecture, interface and platform dependencies.

An aim of the neuGRID project is to provide a user-friendly grid-enabled e-infrastructure, which will enable the European neuroscience community to carry out research that is necessary for the study of degenerative brain diseases. One of its principal goals of the neuGRID infrastructure is to provide a set of generic services, specified in consultation with its clinician user communities. The services should thereby be tailored for medical informatics that can be reusable both across Grid-based neurological data and for wider medical analyses. Services include, but will not be limited to, query, pipeline, provenance, glueing and abstraction, anonymization and portal provisions. In this work we focus on two services that are used to demonstrate how SOA principles and the neuGRID design approach have been applied. The services provide workflow authoring and enactment as well as grid middleware abstraction. Results from a prototype implementation are presented and discussed.

2. The neuGRID Services Infrastructure

The neuGRID infrastructure is based on a Service Oriented Architecture (SOA). We aim to develop a standards compliant service infrastructure making it easier to port services to other medical domains. A SOA enables developers to separate base functionality from the user facing frontend services. Services are published and globally accessible via service interfaces. This principle will not only allow different user communities to interact with neuGRID, but will also help infrastructure services to communicate with each other through standardized interfaces. SOAs have been leveraged in multiple biomedical projects where computational procedures are usually decoupled from client applications.

978-1-4244-4878-4/09/$25.00 ©2009 IEEE

Taverna[3], caGrid[2], @neurIST[4] and many other bio-informatics projects are based on SOA design concepts, enabling wider user communities to benefit from algorithmic and computational features. The neuGRID generic medical services are designed on these principles and an interaction of the subset of services is shown in Figure 1.

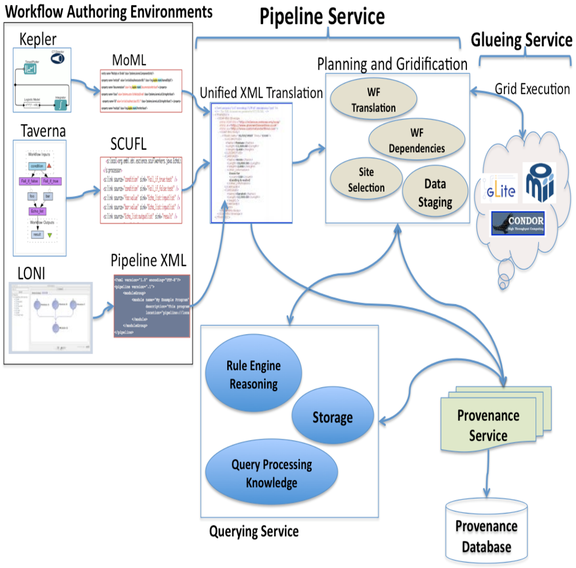

Figure 1: Service Interaction

Unlike state of the art projects, the neuGRID infrastructure focuses on portable and middleware agnostic services. At the backend, services are designed to be middleware agnostic to support deployment on numerous platforms and grid middleware. At the frontend services are designed to support numerous frontends. The neuGRID Pipeline Service for instance, supports common scientific workflow authoring environments and transforms the abstract workflow representations used by these frontends into a common grid executable format. This concrete workflow format is enacted via the Glueing Service. The Glueing Service provides a middleware agnostic platform for submitting and executing workflows. It extends OGF SAGA[7] to SOA environments. The workflow specifications along with the execution information are logged in a provenance database through the provenance service. A querying service will be able to interact with the provenance service in order to bring information out of provenance database.

3. The neuGRID Pipeline Service

Neuro-imaging pipelines allow neuroscientists and clinicians to apply series of automated transformations and processes on brain images for decision support purposes using complex and nested workflows. Often these processes are very compute intensive and deal with large amounts of data. Grid enabled neuro-imaging pipeline services are either proprietary or limited in capability and neuroscientists have to rely on command line scripts to design and execute the pipelines. The primary functionality provided by the Pipeline Service is the enabling of pipeline authoring in a user-friendly environment, using neuro-imaging executables as actors. This component will parallelize and Grid-enable the abstract user defined pipeline for optimal execution over a grid. It will also handle the submission and enactment of the pipeline over a Grid resource. Results of the execution as well as intermediary provenance data will be provided to the end-user

The principles of SOA have been applied during the design of the neuGRID Pipeline Service, where the user facing workflow authoring, workflow planning and gridification and the enactment are decoupled from each other. This enables the service to support numerous user-facing frontend authoring environments including Kepler and LONI. On the service side the Pipeline Service extends existing workflow planners like Pegasus[10] to translate abstract workflow descriptions into concrete grid executable workflows. Unlike related projects, the workflow enactment is decoupled form the underlying Grid middleware. This enabled the Pipeline Service, like other neuGRID services to be middleware agnostic and are portable across Grid middleware. The workflows are enacted via the Glueing Service, which encapsulates the complexities of the Grid Middleware.

3.2. The Glueing Service

The Glueing Service extends the OGF SAGA standard to SOA. SAGA (simple API for Grid Applications) is an open standard defined and maintained by the Open Grid Forum (OGF), which describes a high-level interface for easy programming of Grid applications. SAGA is not aimed at a single middleware but rather it provides APIs for developing applications using different Grid environments. SAGA APIs give a provision of running a job on a particular middleware by runtime loading of a middleware adaptor. SAGA adaptors pass jobs to the middleware as its clients and are responsible for the low-level communication with it. This removes the complexity of using middleware specific APIs for interacting with different Grid middleware.

neuGRID implements a SAGA SOAP Adaptor, which translates SAGA API functions into standard web service calls. Client services and applications are developed against the standard SAGA API while on the backend the SAGA invocations are transformed into SOAP messages handled by the Glueing Service. This enables SAGA-based clients to interact with the Glueing Service in a seamless way, hence enabling the development of grid middleware agnostic services. This shields users and applications from writing complex Grid specific functionality. It will provide a simplified approach, which enables clients to gridify their applications without installing too many Grid specific libraries.

4. Proof of Concept Implementation

A prototype using a subset of services has been developed to build an experimental setup for performing analysis irrespective of middleware dependencies. The primary aim of the implementation was to evaluate the practicality of the design approach. In this prototype two primary services, Pipeline and Glueing, have been deployed. A neuro-imaging pipeline may for example be constructed to identify the cortical surface of raw brain imaging data. In this case the workflow is composed of different algorithms from the MNI neuro-imaging toolkit[11]. It performed a sequence of morphological pruning on brain scans and a 3D cortical surface map.

The implementation is depicted in Figure 2. Kepler specifications are represented by MoML [12], which is translated into a generic XML format by the Pipeline Service. This abstract workflow description is then described by an object oriented workflow API. Pipeline service uses the *Workflow* object to eventually enact each task in the workflow with all its dependencies. The Glueing Service performs the execution of each task, based on the task definition, input literals or data files and middleware information. The middleware that used was the OMII Grid Middleware[13].

5. Related Work

There are numerous state of the art efforts, which aim to create high-level grid service infrastructures for the medical domain. caBIG[2] is a Grid services infrastructure which aims at providing standard applications, common data models and tools to enable more efficient access and sharing of distributed computation resources in cancer research. caGrid is the services infrastructure used in caBIG. It is based on the Globus Grid toolkit [7]. caGrid provides a number of services ranging from analytical, data services and service advertisement and discovery as well as security.

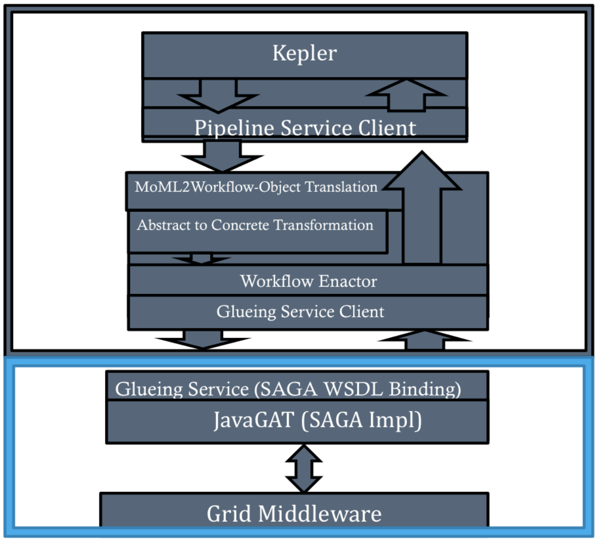
Kepler
Pipeline Service Client
MoML2Workflow-Object Translation
Abstract to Concrete Transformation
Workflow Enactor
Glueing Service Client
Glueing Service (SAGA WSDL Binding)
JavaGAT (SAGA Impl)
Grid Middleware

Figure 2: Implementation Architecture

A related effort is NeuroGrid[14]. Unlike other projects, NeuroGrid is not built on a specific Grid middleware, but uses web-services. It provides pre-developed workflows for stroke analysis, dementia and psychosis. Users select data sets and the workflows and retrieve results from a portal. NeuroLOG[15] is a service workflow-based infrastructure. It uses Taverna[3] and MOTEUR for authoring service based workflows, gridifying and enacting them. Taverna authors SCUFL workflows which are web-service based. neuGRID pipelines mostly comprise neuro-imaging algorithms, which are executable tasks. Deploying a NeuroLOG like architecture would be expensive and at the same time would constrain scheduling. Therefore, this approach would not help in optimizing the use of Grid resources for rapid and efficient processing of neuro-imaging pipelines on a Grid infrastructure.

The Biomedical Informatics Research Network (BIRN)[1] is a federated and distributed infrastructure for the storage, retrieval, analysis and documentation of biomedical data using state of the art open source toolkits built on Grid technology. The BIRN infrastructure is based on the Globus Toolkit. Condor cluster middleware is the primary means of executing workflows in the BIRN infrastructure. Various BIRN partners have developed numerous workflow authoring environments to aid authoring and enactment on the Grid. These include the LONI Pipeline and Kepler.

Both BIRN and caBIG, aim to develop domain specific Grid infrastructures based on a specific Grid middleware, while NeuroGrid provides its own middleware framework. The NeuroLOG infrastructure is based on service-based workflows, which would constrain scheduling decisions in the Grid. neuGRID aims to develop an infrastructure of services which is both portable and middleware agnostic enabling its application to a number of related medical applications. Services will also be designed to support a number of interfaces. As in the case of BIRN, both Kepler and Loni Pipeline are supported while adding the capability of executing authored pipeline against any available Grid Middleware.

6. Conclusions

In this paper we have presented an overview of the services infrastructure in the neuGRID project and highlighted the significance of an SOA based framework in component-based development environments. The main focus of this work is to demonstrate the provision of a middleware agnostic paradigm in biomedical analysis frameworks. For this purpose much emphasis has been given to the seamless interaction of client applications with the underlying Grid middleware. The importance of the Pipeline and Glueing Service, in the context of decoupling client applications from distributed Grid frameworks, has been elaborated. We have also presented a proof-of-concept implementation, which provides a middleware abstraction layer to isolate user interactions from complex Grid functionalities. This has motivated us to test our experimental setup with different Grid middleware and exploiting their computational and storage features simultaneously. The Glueing Service is a step towards an *Inter-Grid Services Framework.* It can also enable the scheduling of workflows to virtual organizations, which constitute of multiple Grids created with different middleware. At the same time, it allows the deployment of neuGRID Services against middleware that is fit for purpose.

Acknowledgements The authors acknowledge the financial support of the Framework Programme 7 of the EC through the Grant Agreement number 211714. In addition they thank the partners in neuGRID for their contributions to this paper from Fatebenefratelli (Brescia, Italy), UWE (Bristol, UK), Maat Gknowledge (Archamps, France), VUmc (Amsterdam, Netherlands), HealthGrid (Clermont, France), Prodema (Bronschhofen, Switzerland), CFconsulting (Milan, Italy) and the Karolinska Institute (Stockholm, Sweden).

References

[1]. Jovicich, J., et al., “Biomedical Informatics Research Network: integrating multi-site neuroimaging data acquisition, data sharing and brain morphometric processing,” *Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems, 2005.* pp. 288 - 293.

[2]. Saltz, J., et al., “caGrid: design and implementation of the core architecture of the cancer biomedical informatics grid,” *Bioinformatics*, Vol. 22, 2006, pp. 1910-1916.

[3]. Oinn, T., et al., “Taverna: Lessons in creating a workflow environment for the life sciences,” *Concurrency and Computation: Practice and Experience*, Vol. 18, 2006, pp. 1067-1100.

[4]. Rajasekaran, H., et al., “@neurIST - Towards a System Architecture for Advanced Disease Management through Integration of Heterogeneous Data, Computing, and Complex Processing Services,” *Proceedings of the 21st IEEE International Symposium on Computer Based Medical Systems*, 2008, pp. 361-366.

[5]. Foster, I., and Kesselman,C., “The Globus toolkit,” *The grid: blueprint for a new computing infrastructure*, 1998, pp. 259-278.

[6]. Wolstencroft, K., et al., “Panoply of utilities in Taverna,” *First International Conference on e-Science and Grid Computing, 2005*, Vol 7. pp. 162.

[7]. Goodale, T. et al., “SAGA: A Simple API for Grid Applications. High-level application programming on the Grid,” *Computational Methods in Science and Technology*, 2006,

[8]. Altintas, I. et al., “Kepler: an extensible system for design and execution of scientific workflows,” *Proceedings of the 16th International Conference on Scientific and Statistical Database Management, 200*4, pp. 423-424.

[9]. Rex., D., Ma, J., Toga, A., " The LONI Pipeline Processing Environment", *NeuroImage*, 2003, Vol 9 (3), pp. 1033-1048

[10]. Deelman, E., “Pegasus: A framework for mapping complex scientific workflows onto distributed systems,” *Scientific Programming*, 13, 2005, pp. 219-237.

[11]. “The MNI-BIC Software Distribution, http://www.bic.mni.mcgill.ca/software/distribution/,”

[12]. Lee, E., and Neuendorffer, S., “MoML — A Modeling Markup Language in XML — Version 0.4,” *Technical report, University of California at Berkeley, March, 2000.*

[13]. Bradley, J., et al., “The OMII Software Distribution,” *Proceedings of the UK e-Science All Hands Meeting*, 2006.

[14]. Geddes, J., et al., “NeuroGrid: using grid technology to advance neuroscience,” *Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems*, 2005, pp. 570-572.

[15]. Balderrama, J., et al., “NeuroLOG: Neuroscience Application Workflows Execution on the EGEE Grid,” *EGEE conference*, Sept. 2008, Istanbul Turkey.

[16]. Glatard, T. et al., “Flexible and efficient workflow deployement of data-intensive applications on grids with MOTEUR,” *International Journal of High Performance Computing and Applications (IJHPCA),* 2007.