

A Grid infrastructure for managing workflows in bioinformatics applications

Falzone Alberto**, Melato Maurizio**, Porro Ivan*, Ratto Stefano*, Schenone Andrea*, Torterolo Livia*

** NICE s.r.l. Cortanze (AT), Italy

* Department of Communication, Computer and System Sciences (DIST) – University of Genoa, Viale Causa 13, 16145 Genova, Italy
livia@bio.dist.unige.it

Abstract. Many tools have been developed for the composition and enactment of bioinformatics workflows for the science research community. Most of these tools provide a full featured and friendly user interface relying on computationally demanding client applications running on local workstations. This approach is not only limited by the lack of computing resources but also it is not able to provide functionalities for fault tolerance, pervasive access and monitoring of the workflow status and results. In this paper we propose a Grid infrastructure supporting workflow design with grid services as building blocks, workflow activation and monitoring. A bioinformatics prototype application exploiting the described infrastructure is also presented.

Keywords: Grid, workflow enactment, Web services, bioinformatics.

1 Introduction

There is a considerable development in the workflow managing area both in the e-Science and in the e-business communities, but the state of the art still shows a large number of competing proposals using different tools and no accepted standards.

In bioinformatics applications, large scale sequence analysis is a complex task involving the integration of results from numerous computational tools and information repositories that must be tied together in a coordinated system in order to automate the execution of a set of analyses in sequence or in parallel [1].

Bioinformatics scientists need to orchestrate these services in workflows as part of their analyses. With the increasing number of bioinformatics databases and processing tools exposed as Web services, a workflow managing system able to operate according to the SOA standards is an essential tool for e-Scientists for taking full advantage of such resources [2].

In a scenario where data sources and services are distributed and there is an increasing request for computational power, Grid infrastructures plays a very important role [3].

Moreover, since in a Grid environment almost everything can be regarded as a Grid service, special attention has been given to workflows based on Grid services which bring many attractive characteristics such as great efficiency, load balancing, fault tolerance and reliability [4].

In this paper we propose a Grid infrastructure able to provide the basic building blocks for workflow construction and to support workflow enactment and life cycle management. In particular we will show how the EnginFrame Grid gateway can be easily employed to expose Grid services that can interact with workflow designer applications (e.g. Taverna) and that can be used to submit and control workflows. In the proposed scenario workflows are activated by the batch workflow enactor Moteur. It is important to note the double role of EnginFrame at two different levels of the proposed architecture:

1. Grid services provider. Grid services exposed as Web services by EnginFrame can be used in the workflow as standard nodes.
2. Web interface for managing workflow submission and management on the Grid.

In both of its roles EnginFrame exploits the computational power and data access capabilities provided by the backend Grid infrastructure.

In particular, in section 2, we will first introduce the tools we used to build the proposed architecture. Then, in section 3, we will describe the double role of EnginFrame as Grid services producer and as a tool for workflows submission and monitoring on the Grid. Finally, section 4 briefly describes the implementation of a bioinformatics application workflow exploiting the proposed architecture.

2 System and Methods

2.1 EnginFrame and Genius Grid Portals

EnginFrame is a Web-based innovative technology, by the Italian company Nice S.r.l.¹, that enables the access and the exploitation of Grid-enabled applications and infrastructures.

It allows organizations to provide application-oriented computing and data services to both users (via Web browsers) and in-house or ISV applications (via SOAP/WSDL based Web services), hiding all the complexity of the underlying Grid infrastructure and actually providing an abstraction layer on the Grid technologies.

Since developed services are automatically exposed as Web pages, Web services and portlets, EnginFrame is also called a Grid *gateway*.

In particular, EnginFrame greatly simplifies the development of Web portals exposing computing services that can run on a broad range of different computational Grid

¹ <http://www.nice-italy.com>

middlewares (including Platform LSF², Sun Grid Engine³, Altair PBS⁴, Globus⁵, LCG⁶ and EGEE gLite⁷).

EnginFrame supports several open and vendor-neutral standards and seamlessly integrates with JSR168 compliant enterprise portals, distributed file systems, GUI virtualization tools and different kinds of authentication systems (including Globus GSI).

Because EnginFrame greatly simplifies the use of Grid-enabled applications and services, it has already been adopted by numerous important companies all over the world. EnginFrame is also well known in the Grid research world for being the technology which GENIUS [5], the official Grid portal of the European project EGEE⁸, is based on.

Thanks to the specific grant provided by the INFN Grid Project⁹ and to the agreement between INFN and Nice S.r.l., the GENIUS code is open source and the EnginFrame license is free of charges for the academic and research organizations.

2.2 Taverna Workflow Manager

Taverna¹⁰ is an open-source workflow tool which provides a workflow language (Scufl - Simple Conceptual Unified Flow Language) and graphical user interface to facilitate the easy building, running and editing of workflows allowing the integration of resources that are published as Web services [6].

The tool includes a workbench application which allows users to construct complex analysis workflows from components located on both remote and local machines, run these workflows on their own data and visualize the results.

Workflows can be executed through a workflow enactment engine called Freefluo. It is a Java workflow orchestration tool for Web services that supports a subset of the Web Services Flow Language as well as Scufl.

Since Taverna is widely adopted by the bioinformatics community and the Scufl language it uses is becoming a standard within the e-Science community, we choose it as the tool for designing and creating workflows in our infrastructure.

² <http://www.platform.com/Products/Platform.LSF.Family/>

³ <http://gridengine.sunsource.net/>

⁴ <http://www.altair.com/software/pbspro.htm>

⁵ <http://www.globus.org/>

⁶ <http://lcg.web.cern.ch/LCG/>

⁷ <http://glite.web.cern.ch/glite/>

⁸ <http://public.eu-egee.org/>

⁹ <http://grid.infn.it/>

¹⁰ <http://taverna.sourceforge.net/>

2.3 Moteur Workflow Enactor

Moteur (hoMe-made OpTimisEd scUfl enactoR) is a service based workflow engine developed in the AGIR project¹¹ optimized for dealing with data intensive applications¹². The current prototype is able to enact a workflow of standard Web-Services described with the ScufI language used by Taverna workbench.

Apart from describing the data links between the services, this language supports a synchronization strategy providing coordination constraints; it also specifies iteration strategies to handle input data of the services and to efficiently reduce complex data-intensive applications graphs into much simpler ones.

Moteur has been selected as workflow enactor for our prototype because it can benefit from computation parallelization at different levels: intrinsic workflow parallelism, data parallelism (multithread) and services parallelism (pipeline), in order to better exploit the resources available on a grid infrastructure [7].

Compared to Taverna that imposes a fixed number of threads to 10 and doesn't support pipelined execution of a workflow, Moteur allows to dynamically determine the thread number during the execution and can better support multiple parallel iterations of a single workflow on many input data sets, conditions often required by bioinformatics algorithms.

3 Proposed Grid infrastructure

In the prototype architecture we designed and realized we gave a double role to the EnginFrame Grid gateway.

The first role sees EnginFrame as Grid services provider. Defined Grid services are exposed as WSDL/SOAP standards Web services and made available to consumer applications like ScufI workflow enactors.

The important result achieved by this functionality is that Web/Grid services exposed by EnginFrame can be used as nodes of the workflow graph by workflow designer applications and can be triggered for running during the workflow execution.

In the proposed infrastructure, we use Taverna as workflow designer.

The second EnginFrame role is as Web interface for handling workflow enactment and management on the Grid.

The Grid portal Web interface provides the user with a service to submit workflows as Grid jobs and monitor/control them as standard jobs.

In the proposed scenario workflows are executed by the batch workflow enactor Moteur.

The service interface implemented provides the following functionalities:

1. *Upload of ScufI workflow and related inputs*: users can upload their own workflows in ScufI language and insert input data for execution

¹¹ <http://www.aci-agir.org/>

¹² <http://www.i3s.unice.fr/~glatard/software.html>

2. *Submission of Moteur as Grid job*: the workflow is executed by Moteur on the Grid infrastructure as a standard Grid job
3. *Monitoring of workflow*: it is possible to check both Moteur submission and the workflow processing status together with data produced by intermediate results.
4. *Results visualization*: when jobs are terminated, workflow results are staged in the EnginFrame spooler area and made available through the portal for visualization, post-processing or download.

With this service bioinformaticians can pervasively run and monitor their own workflows simply using a standard Web browser, exploiting the computational power and data access capabilities provided by the backend Grid infrastructure and taking all advantages typical of server-side services (reliability, fault tolerance, load balancing, etc).

4 A bioinformatics application

In order to evaluate and test the described infrastructure, we have implemented a workflow version of an innovative bioinformatics application developed by the Bio-Lab team of the University of Genoa and based on DChip [8], one of the most complete and diffuse free software for the microarray data analysis.

The application is quite innovative and suitable for a workflow implementation due to its architecture designed into different parallel modules:

1. data set opening and normalization
2. model based gene expression
3. extraction of differentially expressed genes
4. clustering

The first module carries out the opening of the microarray images and their normalization based on invariant set method; the second one performs the gene expression calculation; the third module works out the gene expression extraction and the last one performs the clustering.

For each application module a new EnginFrame service has been defined.

Referring to the first role of EnginFrame in the infrastructure, Taverna has been used in order to build the application workflow. The workflow uses the Grid services corresponding to the application modules exposed by EnginFrame as processor nodes. A further service has been developed and published in EnginFrame in order to submit ScufI workflows to the Grid and execute them by the batch workflow enactor Moteur. This service accepts as input the ScufI workflow description and the input parameters which the workflow will be instantiated with. This implements the second role of the EnginFrame Grid gateway as explained in the infrastructure section.

5 Conclusions

In this paper we have proposed a Grid infrastructure supporting workflow design and managing with Grid services as building blocks, workflow enactment and life cycle management.

With the described approach we have gains and benefits about the following aspects: access to the infrastructure through a common browser independently from the user workstation, de-serialization of atomic jobs thanks to their submission on the grid, logic abstraction of services used to perform the submission on the Grid, shorter running time given from the grid distributed environment.

Today available workflow management systems like Taverna have strong limitations due to their intrinsic client only nature. Introducing a server side component to execute workflow on a grid infrastructure away from user workstation could provide greater reliability and fault tolerance, remote access to the work environment and a significant reduction of workflow processing times.

Even if Moteur appears as a performing enactor tool, it imposes some limitations about processors and data inputs that must be string typed. As a consequence, it causes many standard incompatibilities with most common used bioinformatics services proposed by Taverna and other workflow management systems.

In the future, a more robust service layer will be developed. Improvement in workflow compatibility with currently available bioinformatics resources will be performed taking particular attention to Moteur-Scufl support.

Acknowledgments.

This work is funded by the Italian FIRB project LITBIO (Laboratory for Interdisciplinary Technologies in BIOinformatics).

References

- [1] Merelli I., Morra G., D'Agostino D., Clematis A., Milanesi L., "High performance workflow implementation for protein surface characterization using grid technology", *BMC Bioinformatics*, 2005 Dec 1;6 Suppl 4:S19
- [2] Altintas I., Berkley C., Jaeger E., Jones M., Ludscher B., Mock S., "Kepler: Towards a Grid-Enabled system for scientific workflows", *Workflow in Grid Systems Workshop in GGF10*, Berlin, March 2004.
- [3] Deelman E., Blythe J., Gil Y., Kesselman C., Mehta G., Patil S., Su M., Vahi K., Livny M., "Pegasus : Mapping scientific workflows onto the Grid", *Across Grids Conference*, Nicosia, Cyprus, 2004.
- [4] Gil Y., Deelman E., Blythe J., Kessleman C., Tangmunarunkit H., "Artificial Intelligence and Grids: Workflow Planning and Beyond", *IEEE Intelligent Systems special issue on e-science*, 19(1):26-33, 2004.
- [5] Andronico A., Barbera R., Falzone A., Kunszt P., Re G. L., Pulvirenti A., Rodolico A., "Genius: a simple and easy way to access computational and data grids," *Future Gener. Comput. Syst.*, vol. 19, no. 6 (2003), 805–813.

- [6] Oinn T., et al. "Taverna: Lessons in creating a workflow environment for the life sciences", *Concurrency Computat.: Pract. Exper.* 2000; 00:1-36
- [7] Glatard T., Montagnat J., Pennac X., "Grid-enabled workflows for data intensive applications", *Proc. 18th IEEE Symp. on Computer Based Medical Systems (CBMS'05)*, Dublin, Ireland, June 23-24, 2005. IEEE.
- [8] Beltrame F, Corradi L, Milanesi L, Papadimitropoulos A, Porro I, Scaglione S, Schenone A, Torterolo L, Viti F "A Grid Based Solution for Management and Analysis of Microarrays in Distributed Bone Marrow Stem Cells experiments" *Proc Bioinformatics Italian Society*, April 28-29, Bologna, Italy, 2006