EGI User Forum 2011

EUROPEAN MIDDLEWARE INITIATIVE

Contribution ID: **116** Type: **Demonstration**

g-INFO portal for monitoring Influenza A on the Grid

Wednesday, 13 April 2011 06:30 (7h 30m)

Overview

In this paper, we present a portal for monitoring Influenza A based on the g-INFO system. g-INFO (Grid-based International Network for Flu Observation) project aims at running and connecting various bioinformatics programs, recognized for their accuracy and speed, to continuously reconstruct a robust phylogenetic tree from a set of sequences publicly available and daily updated. We implemented a dynamic bioinformatics workflow in g-INFO so that an expert can choose which components he wants as well as the order of the components in the workflow for their specific analysis. Workflows are deployed on grid resources to take advantages of its high security, heterogeneity and large-scale computation. Finally, a portal was developed on the top of g-INFO to help normal user without knowledge of grid can build and run bioinformatics workflows easily.

Impact

We have made some tests with the workflows created by g-INFO portal. One test studies the difference of HA / NA segment of H5N1 strains between year 2009 and 2010 and it shows that most of 2010 virus sequences are clustered together. The other test studies the capacity of running several instances of a g-INFO workflow in parallel. With each sequence in the input file, BLAST will find n sequences in a database that are most similar to that sequence. The rest of the workflow will construct a phylogenetic tree from these n + 1 sequences. In g-INFO system, each job can handle a task. If and only if the task has done, the job is free to handle another task. In consequence, we can conclude that we need at most N jobs to run the workflow in parallel. This test runs with a workflow on 12 sequences of H5N1 (HA segment, 2010, human host). Time to run 12 instances of the workflow is 897s while the time maximum of an instance is 866s. We submitted only 20 jobs for this test. In the real production running mode, WPE has approximate 3000 jobs available on the Grid.

Description of the work

g-INFO is implemented and deployed on the EGEE (Enabling Grids for E-sciencE) infrastructure, which is based on a Grid Middleware stack called gLite. Besides gLite, a large-scale deployment of the phylogenetic pipeline requires the use of an environment for job submission and output data collection: the WISDOM Production Environment (WPE). The WPE is composed of 4 principal components

- The Task Manager interacts with the client and hosts the tasks to be done;
- The Job Manager submits the jobs to the Computing Elements (CEs) where the tasks managed by the Task Manager will be executed;
- The Data Manager interacts with the client to handle data in batch mode;
- The WISDOM Information System uses AMGA (ARDA Metadata Grid Application) to store all meta-data needed by the Data Manager and the Job Manager.

The TaskManager of g-INFO hosts 4 general tasks for phylogenetic analysis: BLAST, Muscle, Gblocks and PhyML. With these tasks available in WPE, we can implement static phylogenetic pipelines. However, to enable workflows in g-INFO, we use the MOTEUR workflow engine. MOTEUR is a workflow designer and enactor developed by I3S and CREATIS laboratories that is interfaced with gLite grid middleware and handles application services asynchronously. For this reason it is perfectly suited to handle long makespan workflows such as g-INFO. MOTEUR provides a very flexible framework to run g-INFO as the workflow can be built from a set of independent services, and can be modified interactively through a graphical interface. Furthermore, it provides advanced data parallelism constructs well adapted to exploit distributed grid resources.
The g-INFO portal was developed with several web technologies: web services, JSF 2.0, ajax, etc. The portal interacts with g-INFO system via web services. User can create workflow template, run workflow and visualize results with a user-friendly interface on a web browser.

URL

http://g-info.healthgrid.org/

Conclusions

Analysis of the influenza virus genome is of utmost importance to understand its pathogenicity, origin and capacity for human-to-human transmission, and anticipate a potential pandemic. H1N1 received lately great attention from public health authorities and media, but the H5N1 virus has also continued to evolve and cause outbreaks, requiring relevant tracking. Therefore, we presented in this paper a portal that can help expert to create dynamic bioinformatics workflow implemented in g-INFO to monitor avian flu. The current phylogenetic workflow is just a starting point. The work in perspective includes the access to more influenza databases. Although having the possibility of using non-public data, a security framework must be developed to allow the data owner to keep privileges on his own data. More bioinformatics tools will be added to g-INFO system.

Primary author: Mr TUNG, Doan (Institut de la Francophonie pour l'Informatique)

Co-authors: Mr NGUYEN, Hong-Quang (Institut de la Francophonie pour l'Informatique); Dr BRETON, Vincent (CNRS)

Presenter: Mr TUNG, Doan (Institut de la Francophonie pour l'Informatique)

Session Classification: Demonstrations

Track Classification: Demonstration - Technology/Service