Genomics and biomedical data management system

Gnocchi M[1,2],Milanesi L[2]

[1]Institute of Biomedical Technologies ITB-CNR, Milano, Italy
[2]Lombard inter-university consortium for automatic computation CILEA,
Segrate (MI), Italy

Motivation

Genomics and linkage analysis have recently demonstrated an efficient approach for the genetic epidemiology and population studies. These are statistical methods for the individuation of the link between different loci in a genetic disease, and markers inside different families by following their transmission through the generation mutation. Actually it doesn't exist an exhaustive solution, in fact the relative algorithm has a NP-Hard complexity and the various approximate methods has a very high computational and memory cost. This means that can be difficult or impossible to use a single CPU for the implementation and execution of appropriate algorithms on a large sets of data. The aim of the present work, is the systems implementation that is able to use high performance computing infrastructures (for example the EGEE-III Grid or dedicated cluster), for the different execution of genomics and linkage analysis program in a bioinformatics scope.

Methods

The developed system can be divided into three parts. The first part is composed by CLIENT SIDE, which offers an easy and intuitive interface. Through it, the users can choose, configure and submit theirs analysis. The second part is the SERVER SIDE, which uses user's data to submit and monitor the analysis on the specified infrastructure. The last part is composed by HPC (High Performance Computing) SIDE, which executes the server directive and returns the obtained results. For the web interface creation has been used client side languages, such as Javascript, HTML and CSS. They provide an environment intuitive and easy to use, through which it is possible configure the different analysis parameter. This configuration is sent to the server, through a XML formatted string, that is dynamically generated with an javascript script. For the communication between client and server, the system uses AJAX language, which implements the application asynchronous structure and it is compatible with major standard browsers. The server side is developed by using technologies as java, web services and XML, in order to promote an easy portability and usability on the different systems and architectures. This part can be divided into four micro areas: 1.Servlet: execute the users request (jobs state, upload input files, etc...) 2.Parser XML: Creates and copy all files necessary for the analysis execution through the uses of user XML configuration file. 3.Web Services: submits, monitors and manages the results of the different analysis 4.Workflow: orchestrates Web services to automate the execution of analysis For the cluster job submissions, have been developed web services able to interact with the queue scheduler software called PBS; this software permits the execution of parallel jobs on a cluster platform. To increase the possibility of distributing the computation on a wider platform based on Grid technology, have been developed web services able to interact with VNAS, that is the framework developed specifically to facilitate the submission and monitoring of the job in a Grid environment. Through VNAS, the system divides the workload and process between the different distributed resources. It's an advanced system for the submission and monitoring of the jobs, through the integration of advanced strategies for the individuation and automatic resubmission of the failed or hang-up jobs. All the web services developed are then orchestrated by a BPEL based workflow. This allows the whole process of submission, monitoring and jobs output manage to be automatic and software independent.

Results

Various attempts have been made in order to evaluate the computational efficiency offered by this type of approach. The first attempt is to test the execution performance of the infrastructure. To get the estimate of the Grid environment performance, has been made the comparison between a single CPU with 2 GHz and dedicated cluster able to distribute the load until 280 core. The second attempt has been made to test the system overhead, during the submission, monitoring and result management, execution. Also in this case has been used a dedicate Cluster and 2 GHz sinlge CPU to estimate the Grid environment performance. The attempts show that the system adds a variable time between three and five minutes. This overhead is due to several reasons: Time for parsing the XML file,time for analysis submission, time for monitoring the analysis execution and time for the results management. For a limited number of a parallel submission, this overhead affects the system performance; but with ten or more parallel submissions it doesn't have a great influence.

Contact : matteo.gnocchi@itb.cnr.it