CSTgrid: a high performance environment for searching "Conserved Sequence Tags"

Castrignano` T.(1), Talamo I.G.(1), Grillo G.(2), Licciulli F.(2), Gisel A.(2), Liuni S.(2), Mignone F.(3), Pesole G.(2,3)

(1)Consorzio Interuniversitario per le Applicazioni di Supercalcolo per Universita` e Ricerca, CASPUR, Rome, via dei Tizii 6b, 00185, Italy (2) Istituto Tecnologie Biomediche - Sezione di Bari, C.N.R., Bari, Italy (3) University of Milan, Dipartimento di Scienze Biomolecolari e Biotecnologie, via Celoria 26, Milan 20133, Italy

Motivation

The explosive growth of the biological data, stimulated by genome projects, has generated a parallel development of efficient computational approaches suitable for several biological research projects. In this area the need of high performance computing is growing, though usually not affordable by computational resources of a single research laboratory. Grid computing addresses this problem by coordinating and unificating several computational resources. To face the problem of searching "Conserved Sequence Tags" (CSTs) between an input DNA sequence, and several whole model genomes a grid framework can provide high performance, high availability and can fairly handle hundreds of concurrent request. Because the size of several whole genomes now exceed the memory capacity of a single machine, it is necessary to spread the search across multiple distributed working hosts to achieve high performance. This also improves the high availability, since the redundancy of the services increases the tolerance to both machine and network failures. This system also guarantees that the same services can be completed by many machines, reaching the ability to perform more requests that a single machine can handle.

Methods

For monitoring the system and balancing the load of its components it has been developed an ad-hoc solution that makes use of Nagios, a monitoring tool. Nagios has been used exclusively for the underneath monitoring functions (integrated with ad-hoc developed plug-ins, required to monitor specific services such BLAT and CSTminer servers). The system is able to reveal which hosts are down/up, which servers are down/up/heavy-loaded, and to inform system administrators when predefined events occur by sending email or sms (if an sms gateway is available). A PHP library (PHPagios) has been developed to query the status of the system. The PHPagios functions can then be used to show on web pages the status of the system. It follows a diagram of the system: PHP Web Page --> PHPagios --> Nagios --> Network system

Results

Some ad-hoc web pages have been created to monitor the system, and it is possible to explore it in a tree-like way, or view a geographical map with the status of the different hosts that are part of the system and the services they provide. Also a new library is under development, that will help to obtain information about the system status through an XML stream offering new opportunities to extend the system capability. CSTgrid provides a grid-based framework allowing the users to search CSTs in the whole genomes with grants of high performances and high

availability. CSTgrid is also enough generic to implement different bioinformatics applications, maintaining both high performance and high availability requisites.

Contact email: tiziana.castrignano@caspur.it
URL: http://www.caspur.it/CSTgrid