HealthGrid 2003
Lyon, January 16-17, 2003

GénoGRID
an experimental grid for genomic applications

D. Lavenier[1] - H. Leroy[1] - M. Macwing[1] - R. Andonov[1] - M. Hurfin[1] - P. Raipin-Parvedy[1]
L. Mouchard[2] - F. Guinand[3]

[1]IRISA - campus de Beaulieu - 35042 Rennes - France
[2]ABISS, Univ. Rouen, 76821 Mont Saint Aignan Cedex
[3]LIH, 5 rue Philippe Lebon, BP 540, 76058 Le Havre

The GénoGRID project aims to experiment with a grid of parallel computers for time-consuming genomic computations. The computing and data resources belong to genomic or bioinformatics centers split over the western part of France, and are interconnected through the Renater1 and the Megalis2 high speed French networks. The access to the grid is secured and restricted to authentified users.

The project mainly includes 3 different aspects:

1. *A secure and interactive access to the grid*. The idea is that a biologist can access the grid as simply as he can access a standard WEB site. The only difference is that he must be recognized from the system. A connection is thus established through a secured portal by the means of a Certificate Authority protocol. Once connected, a list of applications is proposed according to the user identification. Running an application is done by filling up one or several forms to tune the application parameters and to provide access path to the data. A job control panel allows the biologist to follow interactively the progress of the jobs.

2. *A transparent use of the resources*. The grid is composed of several parallel computers (nodes) geographically dispatched in the western part of France. Since they are located into genomic or bioinformatics centers, the main genomic banks and software are available on the different nodes. On the GénoGRID system, running an application across the grid consists of: (1) splitting the application into independent batches, (2) selecting the nodes having the right resources (data and/or software), (3) broadcasting the request to these nodes, (4) allocating the batches to the nodes according to their loads. Actually, the last operation is performed dynamically: every node runs a distributed algorithm based on a consensus protocol mechanism. From the user side, the allocation of the grid resources is entirely transparent and fully fault tolerant.

3. *The "gridification" of a few genomic applications*. The purpose is to validate our approach with real genomic time-consuming problems. Three applications have been selected as a first experiment: The first one deals with intensive sequence comparison such as data bank to data bank comparison implying sensitive search. The second one concerns the implementation of a protein threading algorithm. The third one is related to the detection of repeat sequences inside full genomes. These applications share the extremely interesting property that they can easily be cut into independent tasks, leading to a very efficient degree of parallelism across the grid.

[1] www.renater.fr
[2] www.megalis.org