Hagrid, a grid computing solution chosen for the Bioinformatics and Evolutionary Genomics group at the University of Ghent.

Eric Bonnet & **Yves Van de Peer**

Bioinformatics & Evolutionary genomics, Vlaams Instituut voor Biotechnologie (VIB), Department of Plant Systems Biology, Ghent University, K. L. Ledeganckstraat 35, B-9000 Gent, Belgium.

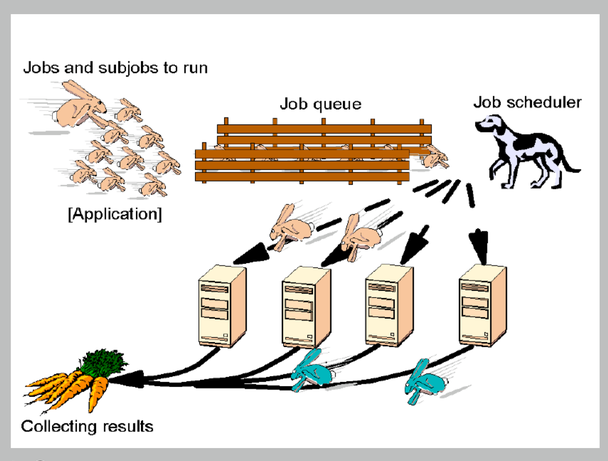
http://www.psb.rug.ac.be/bioinformatics/

Position of the problem

The two crucial material needs every bioinformatics group is faced with are:

- CPU horsepower
- Storage capacity

Since some years, the increasing power of general public processors (INTEL or AMD) contribute to solve the CPU problem by setting up “cheap” solutions based on the concept known as grid computing.



How does it work ?

One or more jobs are scheduled to run on different machines across the grid. The results are collected and assembled to produce the answer.

The solution

Our system, named Hagrid, is designed as follows:

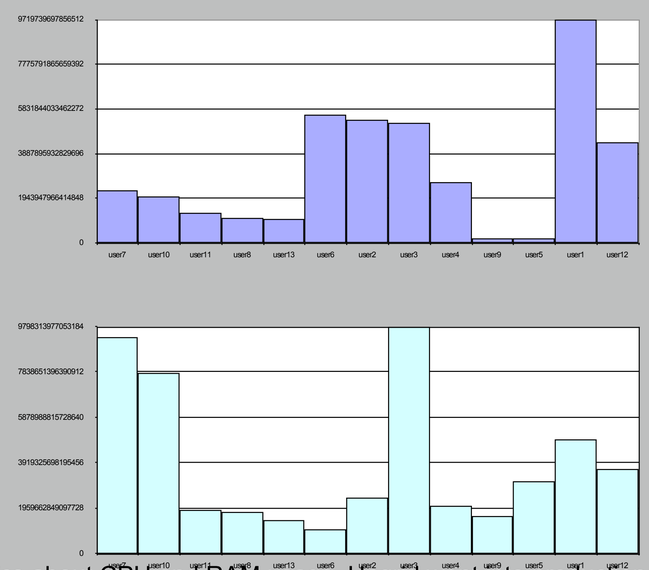
- 5 running nodes and 1 master node (total 12 CPUs and 7 Gb of RAM).

Hardware configuration for each machine:

- Dual-CPU AMD 1900+ processors.
- 1 Gigabyte of RAM (2 for the master node).
- Single hard disk 60 Gigabytes.
- 100 Megabits network backbone.
- NFS share to storage SUN server.
- 2U compact rack “pizza” boxes.

Software:

- Linux RedHat version 7.3 as operating system.
- SUN grid engine for the job scheduler.
- MPICH 1.2.4 librairies for parallelized software.



Some statistics about CPU and RAM usage. Hagrid went into production in October 2002. Approximately 3200 jobs ran on the system.

Aknowledgements

We would like to thank Luc Van Wiemeersch, Philippe Alard and Martin Kuiper from the department of Plant Systems Biology who helped us to set up this solution.

Grid computing

Grid computing can be roughly defined as applying resources from many computers in a network at the same time to a single problem.

More practically it means taking off-the-shelf computers and components and turn them *en masse* into a high-performance computing engine. The result is generally unbeatable in terms of performance/price ratio.

Even a huge organization such as CERN is now setting up grid solutions to manage computing needs.

Our needs

Although the general principle for the structure remains the same, the design of a grid solution (also known as a cluster) is a kind of combinatorial problem between the needs of the end users, maintenance considerations and available money.

In fact several technical options are available which leads to questions such as which processor to use, how many processors per node, how many nodes, how much memory, what kind of network, diskless configurations, etc.

In our case, having a bioinformatics team of about 20 people, mainly doing comparative genomics and using number crunching machine learning techniques, the crucial points were:

- High performing CPUs for number crunching.
- Comfortable memory size allowing manipulation of large data files.
- Current data storage space needs to be available everywhere.
- Flexible and reliable system available 24 / 7.
- MPI enabled system for parallelized software.

A picture of the 5 nodes of Hagrid. These are the 5 “pizza” boxes (2U). Each one has 2 AMD processors and 1 Gb of RAM.

Future Plans

In the very next future we would like to expand this system in order to have more nodes per user.

We are looking actually to purchase an IBM Blade Center, a very compact system where one server is in fact a simple “blade” enclosed in a 7U chassis. One can have up to 14 blades per chassis. Each blade can have 2 CPUs and 2 Hard disks on-board.

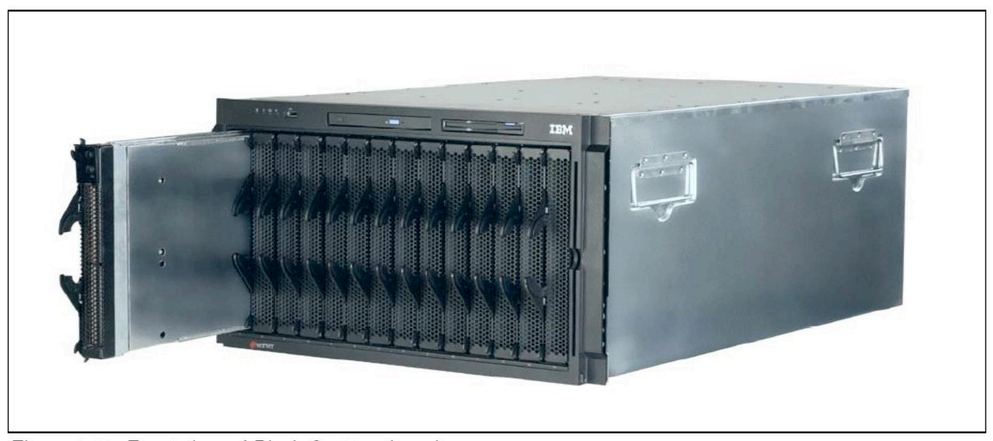


Figure 1-1 Front view of BladeCenter chassis