The Hungarian ClusterGRID Project

PETER STEFAN
research associate

NIIF/HUNGARNET
H-1132 Budapest, Kende u. 18-22
Hungary
Phone: +36 1 4503076
E-mail: stefan@iif.hu

Abstract

In the paper the goals and the main achievements of the Hungarian ClusterGRID project is summarized. The project aims to integrate more than 2000 PCs into a single country-wide virtual supercomputer called ClusterGRID. The enormous computational facility is/will be used by researchers to compute parallel jobs which require large computational performance and which would consume large amount of time on ordinary PCs. The project also goals to remove part of the computational burden from our heavily overloaded SUN E10000 supercomputers.

The beginning of the ClusterGRID project dates back to the spring of 2002, when the Hungarian Ministry of Education initiated a procurement project. Within this investigation more 2000 PCs were purchased and deployed in "PC lab units" in most of the Hungarian universities and public libraries. A single computer lab includes 20 PCs (or workstations), a single server, and a firewall, thus, more than 100 labs are involved. There are different types of PCs in the project including for example HP, Dell or IBM products; however, the common point is that all machines are based on Intel processors.

In the summer of 2002 a technical committee was set up under the coordination of ONIID in order to work out an implementation plan how the deployed PCs can be integrated into a single virtual system. The task is not easy, since the labs are placed to autonomous institutes which have their own Internet policies and defense strategy. Another difficulty is that the computers can only be partially used for computational purposes, i.e. the institutes have also got their own tasks e.g. education to accomplish. So it is easy to see that these PCs can only be used for computation when they are not used by the institutes for their individual purposes: during the nights or during the week-ends.

The applied grid architecture must fit to the "boundary conditions" mentioned above and must satisfy security requirements as well. The key idea is the spatial and temporal separation of the two different goals: institutional goals, and computational (or grid) goals. Spatial separation means to use a different operating system (OS) for the grid than for the institutional purposes, and separate the two OSs on the disk and in network access. This is necessary, since the institutes may use different OSs for their purposes, or may even use multiple systems depending on their needs. It would be an infeasible task to port grid software to all possibly used systems, so it is easier to work out the grid structure for a single OS and then to separate it. On the other

hand, temporal separation means that at the beginning of the work hours all machines are booted with any ordinary OS used by the actual institute; at the end of the work hours all machines are rebooted to use the grid OS. The chosen grid operating system is RedHat Linux 7.3 which is fairly easy to manage.

The heart of the grid is the network structure:
1. All clients are connected to the network via 802.1q encapsulation. Each client is placed into two virtual LAN segments: the "institute VLAN", which is the native one, and the "grid VLAN", which is visible only in grid mode.
2. Whenever a client is booted in grid mode, it connects to a country-wide Multi Protocol Label Switching Virtual Private Network, which is configured within the Hungarian Academic Network and connects the grid labs (in grid mode) together.
3. The clients can not be directly used in grid mode, login from the console on all clients is forbidden. There is a single entry point in the system which accepts job submission, spools and schedules jobs, as well as provides command-line/web interface to users.

From the network point of view, the architecture is not a strict grid architecture, but a three-level hierarchical system: with the entry point on the top, the institutional servers in the middle, and computation nodes (workers) at the bottom of the system. The local servers are used to provide dynamic Internet Protocol address management as well as Network File System (NFS) to the workers. The latter is used for the sake of manageability: all workers receive its root file system via NFS. (It is important to note that only the root file system is mounted as NFS file system, disks on which computation is made are locals.) The clients are also capable to boot from network. The boot image is received from the local server.

Currently a pilot system is implemented involving three institutes: Eötvös University of Sciences (100 nodes, in Budapest), Technical University of Budapest (140 nodes) and Saint Stephen University (24 nodes, in Gödöllõ). Resources are controlled by Condor resource management system. The main advantages of Condor are as follows: instant results, connection of different clusters can be accomplished by Condor's "friendly pools" and "flocking" mechanisms. Sequential jobs, jobs parallelized by Parallel Virtual Machines (PVM) communication library and array jobs are currently supported. The central entry point, as well as the local masters operate during the whole day, workers are switched to grid mode at 8:00 PM every day, and switched off (back to any ordinary mode) at 7:00 AM. User authentication is carried out by central LDAP server.

The future plans can be summarized as follows:
1. to improve the pilot system in quantity involving more institutes to the ClusterGRID,
2. to improve the level of service to the users, installing access portal via web,
3. to improve checkpointing capability of the system.