Technology of Constructing the "Dubna-Grid" Meta-Cluster[1]

A.N.Antonov[2], D.V.Belyakov[1], D.Chkhaberidze[1], E.N.Cheremisina[2],D.C.Golub[1], S.N.Dobromyslov[3], A.G.Dolbilov[1], V.V.Ivanov[1],Val.V.Ivanov[1], L.A.Kalmikova[1], V.V.Korenkov[1], Yu.A.Kryukov[2],V.V. Korenkov[1], V.V.Mitsyn[1], L.A.Popov[1], A.A.Rats[3], E.B.Ryabov[3], Yu.S.Smirnov[1,4], O.G.Smirnova[5], T.A.Strizh[1], P.V.Zrelov[1]

[1] *Laboratory of Information Technologies, Joint Institute for Nuclear Research, 141980, Dubna, Russia*

[2] *University "DUBNA", 141980, Dubna, Russia*

[3] *Administration of Dubna, 141980, Dubna, Russia*

[4] *Chicago University, USA*

[5] *Lund University, Sweden*

The project "Dubna-Grid" [1] is aimed at the creation of a distributed meta-computing environment [2] based on vacant computing resources of office computers. In early 2004, the project participants started to create a unified informational and computational environment of the city, the"Dubna-Grid" meta-cluster on the basis of resources of secondary schools, Dubna University, and the Laboratory of Information Technologies (LIT), JINR.

Various approaches to the installation of the computational infrastructure of such a scale were discussed at LIT and the available technologies were studied. Since the Microsoft Windows OS that is used everywhere for office computers does not support solving complicated and resource-consuming computing tasks in the distributed environment, it has been decided to apply a Linux-based technology of visualization of computing and network resources for construction of the meta-cluster [2]. In order to reach the goals, several technologies and all the potential resources have to be integrated into the computing infrastructure of "Dubna-Grid" meta-cluster, controlled by a unified center (LIT JINR).

The logical structure of the meta-cluster is shown on Fig.1.

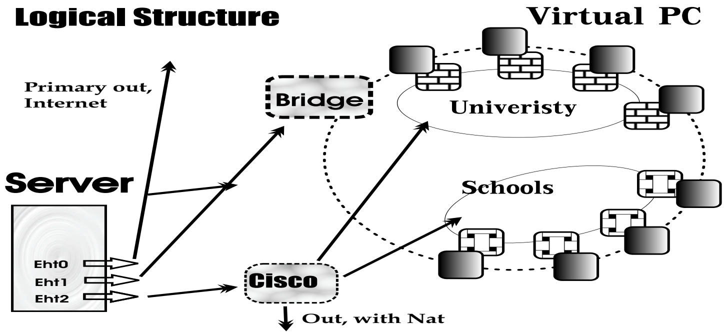

Fig.1 Logical scheme of the "Dubna-Grid" meta-cluster

In order to create the computing meta-cluster, the following equipment and approaches were used:

[1]Partially supported by the grant of the city of Dubna for high-skilled specialists and by the RFBR grant 04-01-97227.

- allocated equipment with own network infrastructure;
- time-shared equipment with common access to the resources;
- virtual clustering of the equipment.

The following software tools and technologies were used:

- software support for virtual machines maintenance (SSVM),
- virtual network (VLAN),
- virtual access to the software and data (AFS) [3],
- integration of the installation and load of the whole meta-cluster (Warewulf package) [4].

The meta-cluster comprises a central server, a software bridge and computing nodes, the so-called clients, that are virtual computers.

The following software is installed on the central server: Scientific Linux CERN OS; a package for support of cluster architecture Warewulf; Ganglia monitoring system; OpenAFS [3], and the batch system Torque with Maui scheduler [5]. At present all the server functions are performed on one computer. However, in the nearest future due to the increasing number of computing nodes, some functions of the central server will be transferred to additional servers.

Scientific Linux OS is installed also on a specialized computer that serves as a software bridge. The software bridge allows one to separate the meta-cluster network from the city network and the University and JINR LANs. The clients are computers of the computer classes of the University, city schools and JINR Laboratories. The SSVM is installed on the client nodes that emulates a particular computer with all its equipment and parameters. Such a virtual computer is started-up as a background process in the Windows environment with a priority per unit less than a standard process. At boot, the client node receives an IP address and name from the server by DHCP protocol, and a kernel and an OS image – by TFTP protocol. In the process of loading a new client, several divisions (swapping, protocol of operating the system, temporal directories for users' tasks etc.) are created in a specialized reserved partition. The main part of the system required for work of the client is loaded into RAM. It is maximally optimized and takes only 50 MB of the Virtual Machine RAM. Some directories are installed from the central server by NFS and AFS (Fig.2).

The developed logical schemes of the meta-cluster and the technology of its construction provide:

1. homogeneity of the environment and compactness of the OS,
2. simplicity of administration and possibility of a dynamic extension of the meta-cluster.

The latter is related to the fact that, as every time when the system is loaded anew on the nodes, the only place where Administrator should change something, is the server. When changing some settings, parameters and other modifications, the system administrator puts all the required changes in the image of the client's file system (it is the same for all the clients). All of them are loaded again with a fresh variant of OS. Thus, the computing complex comprising several hundreds machines, can operate as

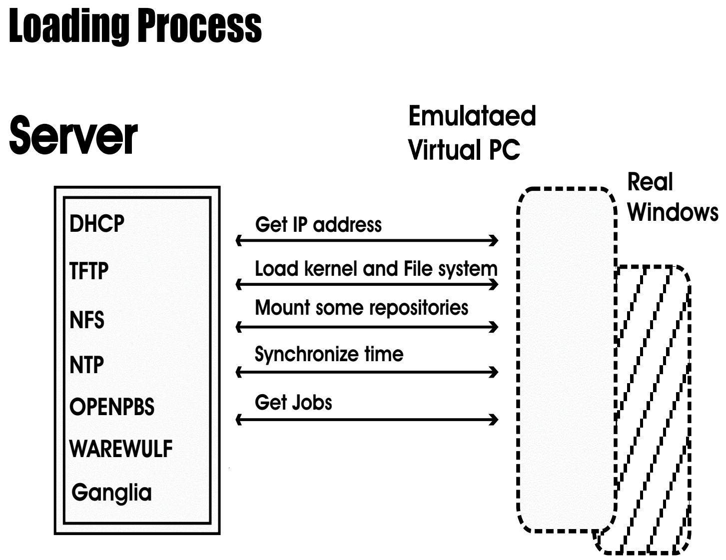

Fig.2 A schematic view of loading the computing node

one computer. Such an approach provides a way for a dynamic extension of the meta-cluster computing environment. To add a new node, one only should make a corresponding record (to register the computer name and IP address, to make alterations in Torque and IP tables [5]). Some components of the client's software (they are mainly settings of the environment and separate packages) can be synchronized via an instruction of the meta-cluster administrator. The administrator can also restart some components of the software on the client nodes by means of one instruction.

In order to support the educational process on the computers integrated into the meta-cluster, one should start up a special service of OS Windows that allows one to increase per unit the priority of the virtual machine (SSVM), giving the resources of the client computer to the processes started by the users of a particular computing node who are not aware that their computers are elements of the computing infrastructure. In the case when a computer is not loaded by educational programs, the virtual machine uses all the resources of the real computer.

In order to provide the effective operation of the meta-cluster, a monitoring of both separate elements and the whole complex is used. With enormous nodes distributed over the whole city, such information is of particular importance. The monitoring of the cluster is done with the help of the Ganglia Monitoring System [6]. Ganglia operates as a client – server setup. The server daemon (GMETAD) retrieves from clients (GMOND) information in the XTM format and archives it in a compact form with the help of RRDTOOL. The web-interface, specially written in PHP, allows one to monitor the state of the meta-cluster via the web-page http://dgrsrv.jinr.ru/ganglia/.

The monitoring system distinguishes three states of a client node:

1. computer is switched on, virtual machine is loaded and works;
2. computer is switched on, virtual machine is not loaded;
3. computer is switched on.

In order to implement the monitoring system, a specialized script has been written which, on the basis of correspondence between the addresses of Windows and virtual machines, starts a parallel query every five minutes. The look-up results are recorded in a file and are available by reference GET ERRORS at http://dgrsrv.jinr.ru/ganglia/ (Fig.3).

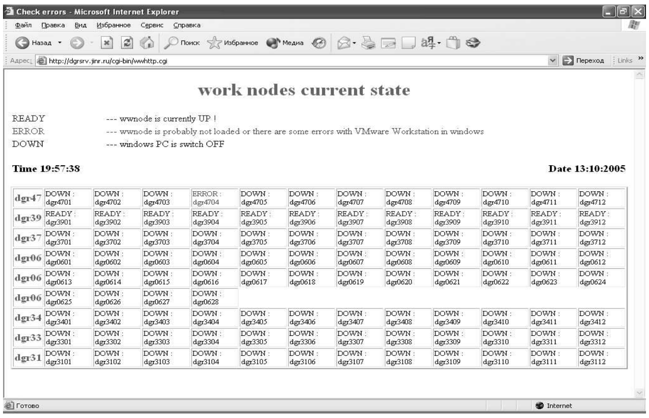

Fig.3. GET ERROR web-page

At present, work is carried out to modify the project architecture. Moreover, it is planned to modernize substantially the following components:

- to install a fresh version of Warewulf cluster 2.4 that will allow one to implement a more flexible architecture of the meta-cluster configuration;
- to change over from loading OS components in the client's RAM to assembling (installing) those components from the AFS environment thus extending the memory for the user processes.

At the same time, increase in the number of client nodes in the meta-cluster is planned. Currently, the following components of the cluster architecture are tested and their integration is planned:

- software to access large disk arrays of LIT JINR (Disk Pool Manager and dCache);
- software of international Grid projects LCG/EGEE and OSG [7]-[9].

References

[1] П.В.Зрелов, В.В.Иванов, Валерий В.Иванов, В.В.Коренков, Ю.А.Крюков, А.А. Рац, Е.Б.Рябов, Ю.С.Смирнов, О.Г.Смирнова, Т.А.Стриж, Е.Н.Черемисина: Проект "Дубна-Грид", Распределенные вычисления и Грид-технологии в науке и образовании, Труды международной конференции, стр. 48-53, Дубна, 29 июня - 2 июля 2004 г.

[2] А. Лацис. Как построить и использовать суперкомпьютер. Москва, Бестселлер, 2003.

[3] http://www.openafs.org/

[4] http://www.warewulf-cluster.org

[5] http://www.clusterresources.com/

[6] http://ganglia.sourceforge.net

[7] http://lcg.web.cern.ch/LCG/

[8] http://public.eu-egee.org/

[9] http://www.opensciencegrid.org