Simplifying Administration and Management Processes in the Polish National Cluster

Miroslaw Kupczyk, Norbert Meyer, Pawel Wolniewicz
e-mail: {miron, meyer, pawelw}@man.poznan.pl

Poznan Supercomputing and Networking Center (PSNC)
ul. Noskowskiego 10, PL-61-704 Poznan, Poland

1. Introduction

Currently there are distributed computing and storage structures used for computing and handling large data sets by the scientific community [6]. The needs and requirements are different, starting from high performance computing up to high throughput computing and it concerns various scientific disciplines, i.e. physics, chemistry, engineering as well as large broadband network services. There are continuing advanced projects concerning the delivery of required network performance and appropriate quality of service of the data level transfer, the development of middleware layers (i.e. management of distributed resources, security of newly developed structures, unified access to storage, friendly, easy and common access to heterogeneous hardware resources) and end users applications able to use the above mentioned tools in an optimum way [7].

There has been created a structure between HPC centres in Poland connecting a few supercomputing centres, based on the national network POL-34/155. Based on this structure we developing a national HPC cluster and a geographically distributed large data set storage management system.

While connecting distributed, geographically aware systems, belonging to different institutions (centres) with their own users management policy, we are faced with the problem of managing user accounts of the whole structure, nation or world wide. The problems are due to policies which are different in each centre regarding the handling of user account databases coherency.

This article presents a concept of a tool belonging to middleware, which enables user database coherency and simplifies the process of handling users in a computing GRID. There are plans to involve the tool in the national HPC cluster.

2. Polish National HPC Cluster

Since 1999 some supercomputing centres in Poland have been connected using the LSF (Load Sharing Facility) HPC Cluster [4]. It is taken into consideration that such solutions help distribute jobs across the multicluster, what enables better machine utilisation. The current state of the situation is transitional between separate user accounts on every computer involved and distributed user account management. All work is done to provide users with transparent account access, uniform disk space, etc. The present queuing system is based on POL-34/155 backbone ATM network [2]. There is a 10 Mb/s channel dedicated to the multicluster computation (Fig. 1). This PVC channel is used both for LSF data and user file transmissions. The file transmission is based temporally on the standard NFS. Such uncomfortable solution will exist until submission to the DCE configuration is employed.

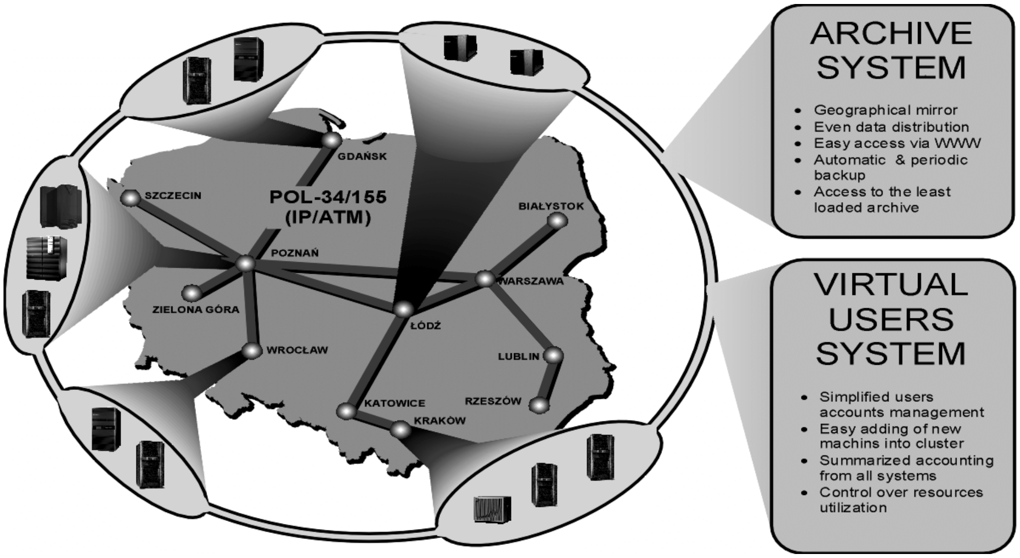
GDAŃSK
SZCZECIN
POL-34/155
(IP/ATM)
POZNAŃ
BIAŁYSTOK
WARSZAWA
ZIELONA GÓRA
ŁÓDŹ
WROCŁAW
LUBLIN
KATOWICE
RZESZÓW
KRAKÓW
ARCHIVE SYSTEM
Geographical mirror
Even data distribution
Easy access via WWW
Automatic & periodic backup
Access to the least loaded archive
VIRTUAL USERS SYSTEM
Simplified users accounts management
Easy adding of new machins into cluster
Summarized accounting from all systems
Control over resources utilization

Fig. 1 National Computing Grid

All centres are fitted with the LSF heterogeneous queuing system and define shared queues. There are some queues defined, i.e. `klaster`, `mgaussian`, and derivative with graduated memory, time and process number constraints. The `klaster` queue handles the maintenance of submitted jobs written by the users who produced their own applications. Detailed description of `klaster` queues is as follows:

- Sending jobs to other `klaster` queues.
- Receiving jobs from other centres.
- Queue jobs limit equal to 5 jobs at the same time.
- User job limit equals 2 user jobs at the same time.
- Number of processes equals 24.
- Bounded group of users with proper privileges to this queue are defined.

The Polish research society that uses supercomputing resources is guided towards the professional biochemical application, e.g. Gaussian version 94 or 98. This application has excessive resource requirements, so there is an extra need for proper resource allocation. Nowadays, in most supercomputing centres, the Gaussian98 version is installed and delivered to users. The `mgaussian` queue is defined in a computational environment, and its configuration looks as follows:

- Sending jobs to other `mgaussian` queues.
- Receiving jobs from other centres.

- Queue jobs limit equals 3 gaussian jobs at the same time.
- User job limit equals 2 user gaussian jobs in the same time.
- Bounded groups of users with proper privileges to this queue we defined.
- Memory limits are set to 400 MB.
- Temporary files are bounded to 4 GB.
- This queue doesn't allowed users to run jobs other than gaussian jobs.

The modern system should be equipped with the possibility to job restart. LSF has got such a feature and it is used in all configured queues.

Computers involved in the Polish national multiclaster have extra network interface aliases defined, especially those with an ATM interface. Everywhere possible, the PVC had been established. The group of IP addresses forms a subnetwork. These addresses create the mcl.pol34.pl domain, and the computers can be reached using both of these and 'old' primary addresses.

3. Dedicated Application Servers

Computational priorities are changing all the time. Using several supercomputers, there is a need to make computational specialisation [1]. This relies on a server dedicated to the several popular applications, e.g. Gaussian 98, Gamess or interactive ones: Matlab, Abaqus, etc. The present configuration makes it possible to submit a Gaussian98 job to the queue that starts a job on the server PowerCHALLEGE XL in PSNC.

In general, the machine should be fit to their running jobs. That is why the Cray SV1 in PSNC has got LSF configured as well, and according to its main memory it is dedicated to Gaussian 98 only. The idle time of this machine decreased to 0 %.

During a computational situation analysis that yielded us a view of the group of application usage, the most popular applications used by Polish research worker were discovered. It goes as follows: Gaussian 98, Gamess, Abaqus.

4. Simplifying user management

Joining systems into a cluster causes several problems connected with user accounts administrating. Because machines a geographically distributed it is not easy to maintain the same configuration on all machines. The main problems are:

1. **User account problems.** If the queuing system is used to balance the load of systems installed in a cluster, all of the users using this system have to have an account on all computing machines. If a new user is permitted to use a cluster, accounts should be created on all machines. Such a situation can lead to incoherence between machines in a cluster.
2. **User accounting problems.** When users are running programs on different machines in geographically distributed sites, accounting can be a problem. Users accounts can have different names and identifiers, so it is impossible to calculate global accounting for a specific user. There is a need for mechanisms that can find mapping information and collect and consolidate data from different machines.

3. **Queuing system incompatibility**. Machines in a cluster can have a different queuing system. For example Polish supercomputers have installed LSF, NQE or Load Leveler. Some of queuing systems (e.g. LSF) can send jobs to others but in a limited range.
4. **File transfer problems.** The most common situation in a local cluster is to share users home directories by NFS or other distributed file systems. Most often it is not acceptable to export disks to remote machines. Therefore before starting a job on a remote system all users' files has to be transferred from a local system to a remote one, and then the files, with the results, have to be transferred back. Some queuing systems give simple support for transferring files

We developed a specialised overlay for any queuing system which solves all of aforementioned problems.

The Virtual User Account System (Fig. 2) consists of few elements:

- any queuing system
- overlapped user interface commands to handle this queuing system
- pool of Virtual User Accounts on each machine
- Virtual User Manager Account with a program that can find and assign one of the Virtual User Accounts
- Virtual User Server responsible for global translation of user identifiers on all machines as well as user accounting and authorising a user's access to the machines.
- modules allowing jobs to other queuing systems to be sent [3]

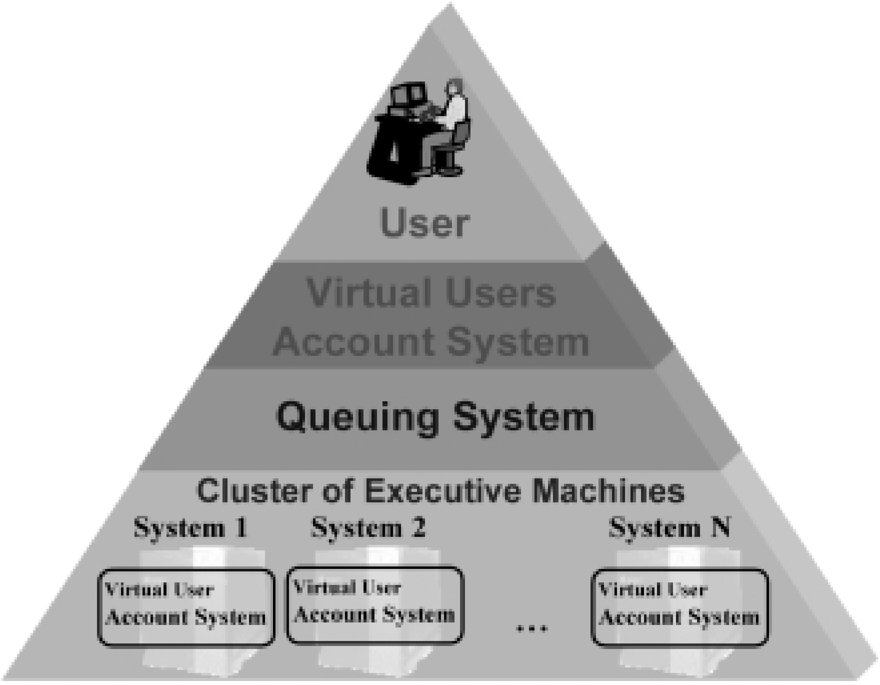
User
Virtual Users Account System
Queuing System
Cluster of Executive Machines
System 1
System 2
System N
Virtual User Account System
Virtual User Account System
...
Virtual User Account System

Fig. 2 Architecture of Virtual Users System

The Virtual User Account System is thought of as a module co-operating with normal batch processing systems, in which the interaction between a user and the program is minimal: reduced to submitting input data and acquiring results. In this situation a job can be run on any system, transparently to the user, in a way allowing load balancing of computing systems. A user can submit jobs and monitor jobs status through a user interface which is a set of specific commands similar to those in queuing system interface. Because forcing users to change the interface they know will be inconvenient for them, this is the only part of the Virtual User Account System which has to be tailored to a particular queuing system's needs. Those commands are communicating with the Virtual User Account Server (Fig. 3) to save information about users during job submissions or to get data required to properly present queues and jobs status to the user.

The main element of a virtual user system is the account of Virtual User Account Manager, which should exists on every machine in cluster. This user owns every task submitted by real users to a queuing system and this account is used to start a task on a remote system. When the queuing system starts a task on a remote machine, a special program is invoked and it assigns and manages the Virtual User Accounts on this executive machine and sends accounting information to the Virtual User Account Server. After the execution is complete, accounting information is collected and sent to a Virtual User Account Server, which keeps information about the total CPU time used by a real user. The number of tasks executed in the same time on a particular machine is limited by the number of created Virtual User Accounts and by the settings of the queuing system. An administrator can add or delete accounts and in this way regulate the maximum load introduced by the queuing system. Due to security reasons, the Virtual User Account should be used only through Virtual User Manager, which will set the UID for processes, no login should be possible for these accounts.

An essential element of our system is the Virtual User Account Server daemon. It keeps information about mapping from real users to Virtual User Account identifiers on all systems, provides an access control list to individual systems and queues and collects and processes information about CPU time used by users on all systems. To allow new users to use distributed queues, only a modification to the access control list on the Virtual User Account Server is needed. There is no need to add accounts on all physical systems, which a user is allowed to use.

In order to send jobs to other queuing systems there exists a special gateway module that works as an interface to remote queuing system. Jobs are sent to a gateway module that knows how to submit them to the system installed in a remote machine. There must be installed one gateway module for every system or cluster of systems with a different queuing system.

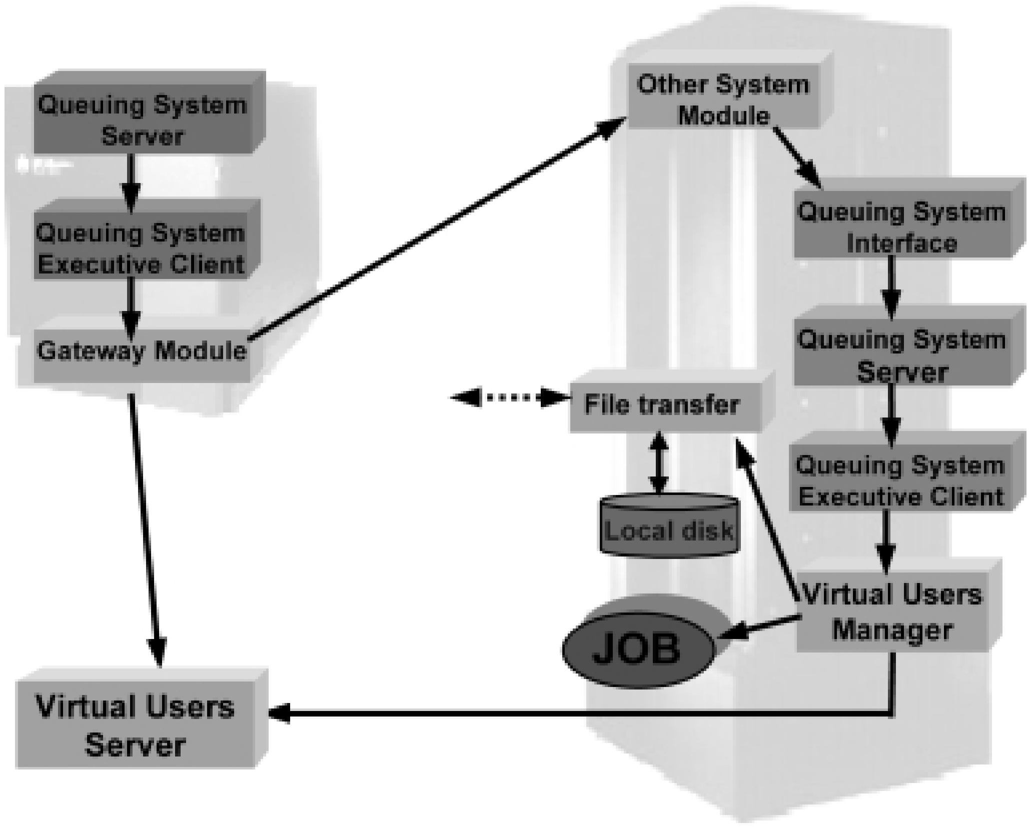

Fig. 3 Job flow in the Virtual Users Account System

In Pozna_ Supercomputer and Networking Center we have chosen LSF as the basic queuing system. We constructed a cluster of Silicon Graphics computers joined through LSF queues. On each machine there are a certain number of Virtual User Accounts. Tasks submitted to LSF are executed on remote machines on virtual user's accounts independently of the existence or not of the account for the user who submitted this task. After the task was completed, the Virtual User Account Server received completed information about the CPU time used by a user on a particular system.

Next we added to the cluster the Cray J90 with NQE queuing system and IBM SP2 with Load Leveler. Jobs where successfully transferred to remote machines and run under the control of local queuing systems. Data files and results were sent to and from remote machines automatically.

5. Future plans

There is a need of greater specialisation of some supercomputers, according to the resource critical application. Cray SV1 is not enough to satisfying all users. That's why other

supercomputers should be pointed out and configured as shared dedicated application servers. Another problem is caused by lack of an ATM interface card in some machines. The easier network configuration could be achieved using unified address space in PVN. This situation could be solved by upgrading proper hardware.
In the future we would like to install the Virtual Users Account System in the national cluster. Then creating accounts for users from remote centres will be unnecessary. We need to carefully check the system for possible security holes. We want also implement gateway modules from NQE and LL to LSF and implement versions of user interfaces for Unicos and AIX that will be similar to those from NQE and LL.

6. Conclusions

The introduction of the Virtual Users Account System will simplify user account administration. This avoids the trouble of maintaining accounts on all machines joined in a cluster. Users will need to have accounts on one of the systems and they should be added to the list of users who are allowed to use distributed queues which can send tasks to remote machines. Particularly it will be possible to easily add new machines to a cluster and it will be sufficient to add only a certain number of Virtual User Accounts to allow this machine to take part in distributed computing.

Because of the joining of HPC centres it is possible to dedicate some machines to run only one kind of the most commonly used application. Then the administrator can deny interactive logins to this machine. The disk space on such a dedicated machine can be reduced because there are no user accounts and only specific applications are installed.

7. References

[1] J.A. Kaplan, M.L. Nelson *A Comparison of Queuing, Cluster and Distributed Computing Systems*, NASA technical Memorandum 109025, NASA LaRC, October, 1993

[2] M. Nakonieczny, S. Starzak, M. Stroinski, J. Weglarz, *Polish Scientific Broadband Network POL-34*, Computer Networks & ISDN Systems 30, 1998, s 1669-1676

[3] W. Dymaczewski, N. Meyer, M. Stroinski and P. Wolniewicz, *Virtual User Account System for distributed batch processing*, HPC 99, Amsterdam, April 1999

[4] W. Dymaczewski, N. Meyer, M. Stroinski, R.Tylman, P. Wolniewicz, *Connecting Supercomputing Centres by using Virtual Users Account System* (in Polish), Miejskie Sieci Komputerowe w Nauce, Gospodarce i Admnistracji POLMAN ’99, Poznan 13-16 April 1999

[5] Gregory F. Pfister, *In Search of Clusters*, Prentice Hall, 1998

[6] Ian Foster, Carl Kesselman, *The Grid: Blueprint for a New Computing Infrastructure*, Morgan Kaufmann Publishers, Inc., 1999

[7] Ian Foster, *The Beta Grid: A National Infrastructure for Computer Systems Research*, Network Storage Symposium NetStore ’99, Seattle, October 14th-15th, 1999