# IHEP cluster for Grid and distributed computing

**V. Kotliar**

National Research Center “Kurchatov Institute” State Research Center of Russian Federation Institute for High Energy Physics, Protvino, Russia

E-mail: Viktor.Kotliar@ihep.ru

To build a computer cluster for the Grid and distributed computing is a highly complex task. Such cluster has to seamlessly combine grid middleware and different types of the other software in one system with shared cpu, storage, network and engineering infrastructure. To be able to run effectively and be flexible for the still unknown future usage patterns many software systems must be gathered together to build a complete system with high level of complexity. This work present a general possible architecture for such systems and a cluster software stack which could be used to build and operate it using IHEP computer cluster as an example.

Keywords: Grid computing, distributed computing, computer cluster

## 1. Introduction

To build a computer cluster for the Grid and distributed computing is a highly complex task. Such cluster has to seamlessly combine grid middleware and different types of the other software in one system with shared cpu, storage, network and engineering infrastructure. To be able to run effectively and to be flexible for the still unknown future usage patterns many software systems must be gathered together to build a complete system with high level of complexity. As soon as each computer center in the world has it's own specific in the implementation depends on physical location, or engineering infrastructure, or usage patterns it is impossible to create one architecture which will serve for all. This work presents a possible general architecture for such systems and a cluster software stack which could be used to build and operate such system using IHEP computer cluster [Kotlyar V, GRID 2012] as an example .

## 2. Computing cluster for distributed computing

Distributed computing is a filed of the computer science that studies distributed systems where they consist of components located on networked computers which communicate and coordinate their action by passing messages. That components interact with each other in order to achieve a common goal. Any problem in distributed computing is divided into many tasks each of which is solved by one or more computers. Often such computers could be presented by individual processor plus memory block and communication between them is going by message passing. One of the form of distributed computing is grid computing. In Grids a “super virtual computer” presents many loosely coupled computers by network acting together to perform large tasks. As soon as connections made by conventional network interface like Ehternet it contrasts to the traditional notion of a supercomputer computing where many processors connected by a local high-speed computer bus. It also needs to be mentioned that Grid computers tends to be more heterogeneous and geographically dispersed than cluster computers and Grids are often constructed with general-purpose grid middleware software libraries.

So pay attention to all features of grid and distributed computing it needs to be mentioned that a computer cluster is presented by a set of the computer hardware, storage hardware, network hardware, engineering infrastructure, physical placement, security and usage policies. It is used for computation in more or less one field of science or more or less using similar computation technologies. The core of such cluster is software (usually open source) but to operate such distributed system it has to be used a very large list of programs. For example on IHEP cluster are used following software systems from simplest to very complex, like distributed storage systems. They are: base operating system - Debian or RHEL; data center infrastructure manager - DCIM; cluster management tool like puppet; version control tool like git; Andrew file system; Lustre file system; installation system like FAI; authentication and authorization systems like kerberos plus openldap; batch system scheduler - Maui; cluster batch system TorquePBS; storage systems based on xrootd; dCache storage systems; monitoring systems like nagios, splunk, ElasticSearch with Kibana, munin, pmacct, collectl; and some technology for virtualization like XEN, KVM together with high availability techniques based on DRBD and pacemaker.

## 3. Building a computer cluster

To start building a computer cluster infrastructure management system must be installed (figure 1). Such system describes in a human readable form compute hardware, storage hardware, network hardware, management hardware, engineering infrastructure, physical placement of the hardware and communication between each components. It mandatory that this system has a program interface to communicate with from programs.

When all hardware is described it needs to be made a decision on power and cooling systems to be used taking into account cluster capacity, reliability, connectivity  and how cluster resources are going to be distributed physically in place. After that it has to be decided what is going to be used for network infrastructure. Networking is the core of distributed computing and it must be:

- scalable;
- very reliable;
- high throughput for data transfers;
- independent from general-purpose network as much as possible (dns, gateways, dchp, proxy)
- could be several networks ( computing, storage, infrastructure, power).

Next step in building cluster is to choose authorization and authentication systems. Where authentication is the process of ascertaining that somebody really is who he claims to be (login + password) and authorization refers to rules that determine who is allowed to do what (permissions).

The main resources for the computing is compute hardware and on the next step it needs to be installed a system for their management. For distributed computing the more convenient way of running tasks is through a batch system which can orchestrate computer nodes with different types of hardware (CPU, GPU, Xeon phi). Usually  it comes with a scheduler which determines how to use resources effective and fairly.

Next two systems which are bind to the compute nodes are automatic installation and automatic configuration systems. First, make sure that it is always possible to reproduce any compute node in case of failures and second, make sure that all nodes are configured in the same way what make the cluster behavior predictable and  understandable.

A storage system is needed to store computation results on the cluster or store cluster software. Depending on usage it might be used several types of data and storage technologies:

- home directories with auto-backup;
- big data for fast analysis;
- software area for small files;
- archive storage for long term storage and backup.

At last, as soon as the cluster is ready to operate, it needs to be installed monitoring systems which will allow to know how exactly cluster works and accounting and billing systems are need to be installed for counting the cluster usage.

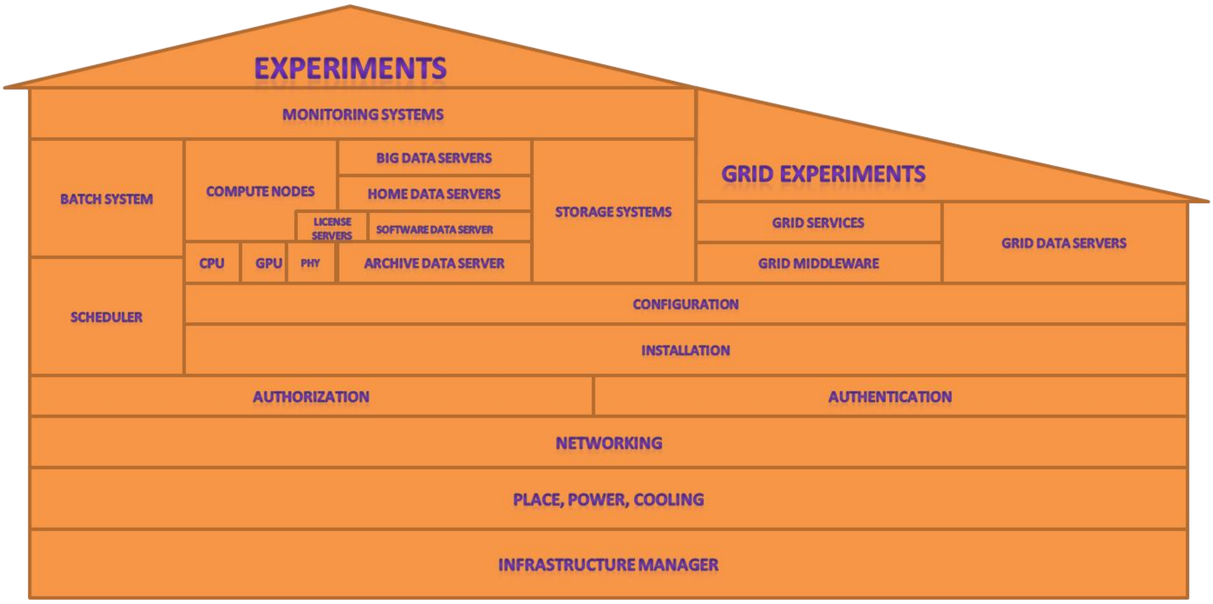

Figure 1. Building blocks for the computer cluster

On top of the cluster are real experiments which actually will use the computing resources. They are for whom the cluster has been created: all experiments have their specific usage patterns, in any time they can ask about something new. It needs to be created a communication mechanism between system administrators and users on how to use the cluster in the best way to achieve users goals.

To add to the cluster for distributed computing a Grid part it just need to be added Grid middleware software on the compute nodes, Grid data servers to store data and Grid services near the cluster ones which will link a cluster with a Grid infrastructure.

## 4. Cluster software

Lets follow the described blocks and fill it with real open source software. As an example of the possible software stack IHEP computer center will be used.

For operating cluster a data center infrastructure manager based on openDCIM is used. This software has features for site management, daily operations, auditing, reporting and it stores in a relational DB information about the whole data center infrastructure.

The next component is a cluster network. Here is used a full set of programs:

- DNS (bind) service to create local domain name space for the cluster;
- DHCP (isc-dhcp-server) for management IP addresses and network options on the cluster nodes;
- NAT (linux masquerading) for translation internal private IP space to the external with the access to the Internet;
- http proxy (squid) for caching http requests of the cluster nodes;
- NTP (ntpd) for date and time synchronization on all nodes at the data center.

To create network links also used distributed trunking and bonding (with LACP) technologies.

Authentication and authorization services are presented by openldap master&slave cluster and kerberos5 master&slave cluster. This software allows to implement SSO (Single Sign On) feature over the cluster infrastructure.

To run a task on the compute nodes at IHEP TorquePBS batch system is used. Is is a patched version for supporting kerberos5 authentication and authorization. As a general-purpose scheduler on the cluster Maui scheduler is used. There are several patches applied to it: patch to increase number of jobs in a queue system on the cluster, patch for GPU cards usage, patch for supporting software features (licenses).

A procedure of the installation on the cluster is performed by using FAI (fully automatic installation) open source software. This software allows in a non-interactive way to install, customize and manage Linux systems and software configurations on computers as well as virtual machines and chroot environments. The main benefit for the IHEP cluster is that it supports both RHEL like and Debian like operating systems. After installation to setup, configure and operate servers puppet, pdsh are used. Again they used for both RHEL and Debian systems. All puppet configurations are stored in a git repository on the different server and each commit is validated by puppet system.

For compute nodes on the cluster Scientific Linux is used as a base operating system, MPI (Message Passing Interface) and OpenMP software installed and on some nodes CUDA, Ansys Mechanical, Wolfram Mathematica are installed. To organize access to the software with licenses it is used a model with dedicated license servers reachable over the network and number of the licenses are counted in the Maui scheduler.

To store data the following storage systems are used [Kotlyar V. , NEC 2013]:

- Lustre parallel file system to store big data and data results for jobs on the cluster;
- Andrew file system to store user home directories and programs are built by users;
- CVMFS (CERN virtual file system) to provide an access to the prebuild analysis software [CVMFS homepage];

- CASTOR (CERN advance storage architecture) to provide archive storage for users [CASTOR homepage].

And, at last, to understand how everything works various monitoring tools are in place. They are common ones like Nagios for monitoring services, ElasticSearch plus Kibana for engineer infrastructure, munin for rrd graphs, Cacti for networking and also some rare tools like collectl for a real-time one glance view of the whole cluster, pmacct for network accounting, splunk for long analyzing, selft-build accounting system based on accounting  information from batch system.

To add Grid functionality to the cluster several grid services are installed in parallel with all other cluster services: grid middleware software based on UMD (unified middleware), site BDII for information system, CREAM-CE for running jobs,WLCG VO-BOXes for specific Grid VO software, Apel for Grid accounting, Perfsonar for network monitoring in Grid, Frontier/CVMFS caching squid proxies, dCache and xrootd storage systems.

## 5. Conclusion

This work presents building  blocks for a computing cluster for Grid and distributed computing. As example IHEP computer cluster is used where on one hardware may work different internal experiments like BEC, OKA, FODS, COMPASS, TNF, LDS, ORI, phenix, Panda and external LHC experiments like ATLAS, CMS, LHCb, Alice. As a matter of fact such system is very complex and need to have a highly qualified system administrators for it's creation but later it might be operated very easialy.

Step by step were described all necessary parts for creating and operating such cluster from the software point of view. Examples for possible software components are made. This work shows that it is possible to combine different computing technologies in one cluster if it is required.

Present description might be used as a starting point for building a new cluster for distributed computing or as an advice to the existing clusters if they are missing some components.

## References

*Kotlyar V., Gusev V., Kukhtenkov V., Popova E., Savin N., Soldatov A.* WLCG Tier-2 computing infrustructure at IHEP. // Distributed computing and grid-technologies in science and education (Grid`2012) —  2012 — ISBN 978-5-9530-0345-2 — стр 150-157

*Kotlyar V., Latyshev G., Popova E., Yutalova A.* IHEP Data Storage Systems for Experiments. // Proceedings of XXIV International Symposium on Nuclear Electronics & Computing (NEC`2013) — 2013. — стр. 166-172

CASTOR homepage —  [Online]. Available:http://castor.web.cern.ch/

CVMFS homepage —  [Online]. Available: ttp://cernvm.cern.ch/portal/cvmfs/release-2.0