NSC KIPT LINUX CLUSTER FOR COMPUTING WITHIN THE CMS PHYSICS PROGRAM

L.G. Levchuk, P.V. Sorokin, D.V. Soroka and V.S. Trubnikov

National Science Center "Kharkov Institute of Physics and Technology", Kharkov, Ukraine
e-mails: levchuk@kipt.kharkov.ua; psorokin@kipt.kharkov.ua

The architecture of the NSC KIPT specialized Linux cluster constructed for carrying out work on CMS physics simulations and data processing is described. The configuration of the portable batch system (PBS) on the cluster is outlined. Capabilities of the cluster in its current configuration to perform CMS physics simulations are pointed out.

PACS: 29.85.+c, 29.40.-n

1. INTRODUCTION

Experimental searches for such manifestations of "new physics" as the Standard Model (SM) Higgs boson or supersymmetry (SUSY) partners to already known elementary particles are among the highest-priority problems of the contemporary high-energy physics (HEP). If detected, the corresponding signals, apart from bearing new evidence for the SM, could pave the ways for further development of its extensions. So, great efforts are being undertaken by the world-wide HEP community in order to provide tools for such a discovery.

Much hope is associated with the Large Hadron Collider (LHC), which is under construction at CERN and will be put into operation by the year 2006. Having two colliding proton beams with c.m. energy of 14 TeV, it will be capable of searching the Higgs boson in the whole range of its plausible masses from 100 up to 1000 GeV/c^2 and will possess a significant potential for the SUSY discovery.

For the LHC project luminosity of 10^{34} cm^{-2}s^{-1}, an average of 20 inelastic events occur every 25 ms, while the fraction of this data array that can be selected as candidates for signals of the "new physics" is very small. In case of the CMS detector [1], a two-level trigger system is developed to reduce the input rate of 10^9 events per second to the filtered rate of 10^2, with 1 Mbyte of information per event on average being stored for the further analysis. For the ten-year operational term of the LHC, the total amount of data stored by the CMS collaboration will exceed 10^{16} bytes. Of course, such huge arrays of experimental information are challenging against the data acquisition, processing and storage systems.

To meet the requirements set by the LHC era physics tasks, the concept of the data Grid [2] has been put forward. Its goal is to provide an infrastructure that would allow one a coordinated use and sharing of computational and storage resources. According to this concept, a multi-tier structure of regional centers is being created combining the computing and scientific facilities of many institutes and research centers from more then 30 countries.

The regional centers are expected to be UNIX clusters of workstations or personal computers (PC). A cluster can be defined (see Ref. [3]) as a type of a parallel or distributed processing system, which consists of a collection of interconnected stand-alone computers cooperatively working together as a single, integrated computing resource. The last several years have witnessed a considerable world-wide quantitative and qualitative growth of PC-based Linux clusters often also referred to as PC farms. They are used extensively already in the HEP laboratories, and their application area gradually broadens.

Main advantages of Linux clusters compared to other computational systems are due to their ability to have high computing performances at relatively low price. The price-per-performance ratio of a PC cluster-type machine is often estimated (see, e.g., Ref. [4]) as being three to ten times better than that for traditional supercomputers. Certainly, the efficiency of the Linux clusters compared to the supercomputers strongly depends on the character of a computational task. One may expect them being preferable in cases, when the task straightforwardly splits into a certain amount of independent or quasi-independent smaller jobs that can be distributed for execution over different processors of the system. This situation is typical for HEP computing: a needed statistics in a Monte-Carlo simulation can be gained by running several jobs with different initial random seed numbers simultaneously, and large data arrays can be analyzed through their subdivision into relatively small groups of events to be processed independently of each other.

Among other advantages of the cluster architecture, one could mention its flexibility and scalability: cluster capacities can be readily increased by adding new nodes step by step. At last, the fact that a lot of free software is available for the Linux platform results in further improvement of the performance-per-price ratio for the Linux clusters.

There exists a variety of cluster configurations from single-image systems with clients booted from the network to heterogeneous parallel systems. One can choose an appropriate hardware and a cluster scheme in accordance with the scientific problem to be solved.

What follows is a brief description of the Linux cluster created in the NSC KIPT for computing within the CMS physics program.

2. NSC KIPT CMS LINUX CLUSTER

Construction of the NSC KIPT specialized cluster to conduct computing activities on CMS physics including simulations and preparatory work for data processing and analysis has been completed in its current configuration by the end of 2001.

The cluster (see Fig. 1) consists of 5 (one “master” and four “slave”) nodes connected by the fast (100 Mbit/s) Ethernet. At present, the nodes are running Linux-2.2-16 (Red Hat 7.0).

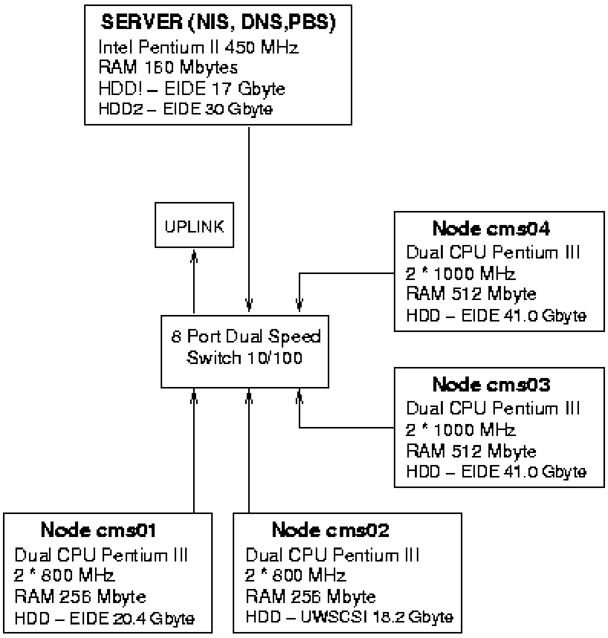

***Fig. 1**. Structure of the NSC KIPT CMS Linux cluster*

The “master” computer (450 MHz Pentium II with random access memory (RAM) of 160 Mbyte) runs (optionally) Linux Red Hat 7.0 or Fermilab Linux 6.1.2. It operates as a file-server that exports the CERNLIB and LHC++ [5] program libraries, CMS-specific software and users’ home directories to the four other computers via the network file system (NFS). Then, it provides the domain of the network information service (NIS) enabling a joint usage of passwords in the network and acts as a domain name service (DNS) server that establishes the reciprocal compliance between the network computer names and addresses. Also, it works as the portable batch system (PBS) [6] server (see below).

Two dual Pentium III (2×800 MHz, 256 MB RAM) and two dual Pentium III (2×1000 MHz, 512 MB RAM) computers have been configured as the “slave” nodes. They are used for calculations in both interactive and batch modes having the hard disk drives (HDD) formatted in a way making it possible to process single files with size of ~10 Gbyte. The local working disk space (the so-called “scratch” directory) is available on each computational node in order to reduce the network traffic. Such a configuration has been chosen in accordance with the character of the computations to be carried on with the amount of the node RAM being determined by the software demands.

At present, the total cluster HDD storage is 170 GB, and the computational capacity is about 311 SPECfpt95 or 361 SPECint95 (see Ref. [7]).

The software currently installed on the cluster includes CERNLIB (containing, in particular, the GEANT simulation package, the PAW/PAW++ physics analysis tool and such event generators as PYTHIA and ISAJET) and the LHC++ program library. In addition, we have installed the software developed by the CMS collaboration including CMSIM (a GEANT-based package for simulation of the CMS detector response), ORCA (an object oriented [based on LHC++] tool for CMS event reconstruction and analysis), and IGUANA (a package for CMS interactive data visualization and analysis). Versions of the programs are permanently refreshed according to CMS collaboration current demands.

The PBS is used as the cluster batch job and system resource management package. It accepts (see details in Ref. [6]) a batch job (a shell script with some control attributes) preserves and protects the job until running, runs the job and delivers output to the submitter. The PBS allows one to administer flexibly the system resources while carrying on the computing and may be configured to support jobs run on a single system, or many systems grouped together. It can load processors of the cluster nodes in an optimal way (in accordance with an administrator policy) and select, e.g., the highest-priority execution jobs.

The configuration of the PBS at the NSC KIPT CMS Linux cluster is presented in Fig. 2. The batch system consists (see Ref. [6]) of a command shell and three daemons: the job server, the job scheduler and the job executor, with the latter being activated on every host allocated for execution. The commands are used to submit, monitor, modify and delete jobs and are available at each of the 5 nodes of the cluster. They communicate through the network with the job server. The server main function is to provide proper processing of the “events”, i.e., such services as receiving/creating a batch job, modifying the job, protecting the job against system crashes and placing the job into execution. The job scheduler is a daemon which contains a “policy” controlling which job has to be chosen for execution, and where and when it has to be submitted. The scheduler communicates with the server to get an information about the availability of jobs to execute. To learn about the state of system resources, it addresses the job executors. (The daemon-to-daemon interface occurs via the network.) The job executor is the daemon which actually places the job into execution. It also takes the responsibility for returning the job output to the user. Once a new job to be executed is found by the scheduler, and free resources are available in the system, the job is submitted to an execution host least loaded at the moment as estimated by the batch system. At present, the maximum number of non-parallel jobs executed on the cluster simultaneously is 8, what is equal to the number of high-performance (≥800 MHz) system processors (cf. Figs. 1 and 2). The slowest processor, 450 MHz Pentium II, performs the server tasks and does not participate in batch executions by default, though can be allocated to a batch job by a special request. If there are no free nodes (i.e., all 8 execution processors are busy), new submitted jobs are put (depending on computing resources requested) into one of 5 queues. When a free processor becomes available, it is immediately allocated to a job from the queue corresponding to the least amount of requested resources.

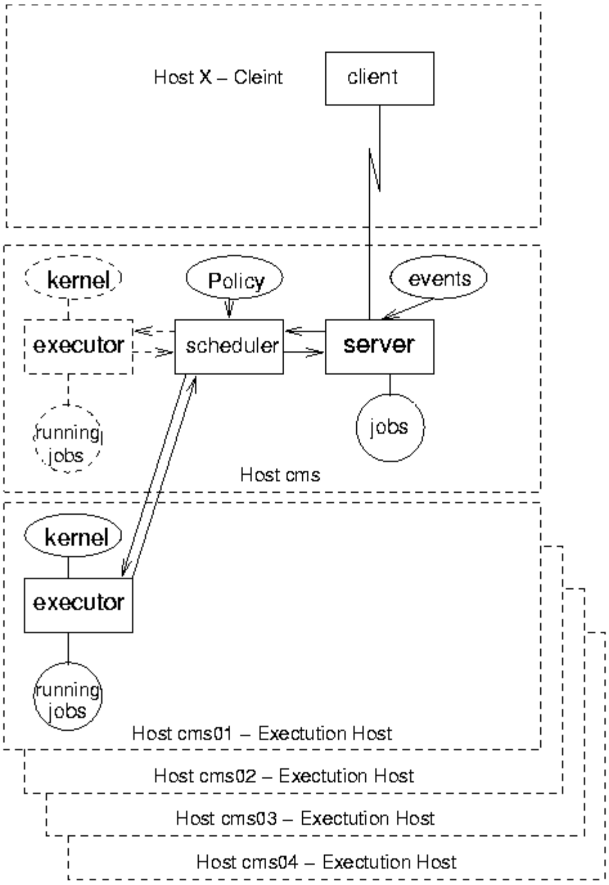
Host X – Cleint
client
kernel
Policy
events
executor
scheduler
server
jobs
running jobs
Host cms
kernel
executor
running jobs
Host cms01 – Exectution Host
Host cms02 – Exectution Host
Host cms03 – Exectution Host
Host cms04 – Exectution Host

Fig. 2. *Use of PBS at the NSC KIPT CMS cluster*

The cluster in its current configuration possesses performances that already allow us to carry on simulations within the CMS physics program. It was exploited in our Monte-Carlo studies [8] of the possibilities to observe the heavy Higgs boson in decays $H^0 \rightarrow Z^0 Z^0 \rightarrow ll\nu\nu$ at CMS. Also, we took part (tentatively) in the CMS SPRING_02 event production run (generation of high-p_T jets in pp collisions at the LHC energy) in order to determine directions of the NSC KIPT CMS Linux cluster development for the nearest future and, on the other hand, to gain an experience of working with new software packages developed by the CMS collaboration and systems supposed to be used in the Grid computations. This work is still under way.

3. CONCLUSION

Construction of the first stage of the NSC KIPT specialized Linux cluster to accomplish computing tasks within CMS physics program has been completed. The cluster consists of 5 nodes (9 processors) connected by the fast Ethernet. The PBS is used as the cluster batch job and resource management system. The cluster performances allow one to perform simulations within the CMS physics program. Further development of the cluster is planned.

REFERENCES

1. *The Compact Muon Solenoid Technical Proposal*. CERN/LHCC 94-38, LHCC/Pl, 1994.
2. http://www.EU-DataGrid.org
3. S. Cozzini. *Introductory Talk*. ICTP/INFM School in High Performance Computing on Linux Clusters, ICTP, Trieste, Italy, 2002.
4. V.V. Korenkov and E.A. Tikhonenko. The conception of Grid and computer technologies in the LHC era // *Physics of Elementary Particles and Atomic Nuclei (PEPAN)* 2001, v. 32, № 6, p. 1458-1493 (in Russian).
5. http://wwwinfo.cern.ch/asd/index.html;
http://wwwinfo.cern.ch/asd/lhc++
6. http://pbs.mrj.com
7. http://www.pcmarkt.ch/spec.shtml
8. L.G. Levchuk, *Possibilities to observe a heavy Higgs signal in $H^0 \rightarrow Z^0 Z^0 \rightarrow ll\nu\nu$ decays at CMS*, CMS/RDMS Meeting, MSU, Moscow, December 2001.