ELEMENTS OF THE GRID MIDDLEWARE AT THE KIPT CMS LINUX CLUSTER

L.G.Levchuk, P.V.Sorokin, D.V.Soroka, S.S.Zub

NSC Kharkiv Institute of Physics and Technology, 61108 Kharkiv, Ukraine
e-mail: stah@kipt.kharkov.ua

Requirements set by physics goals of the LHC experiments that put forward the concept of distributed (Grid-based) storage and processing of large-scale arrays of experimental information are outlined. Operation of the portable batch system at the KIPT CMS Linux cluster is described. Preparation of the cluster for its integration into the Grid structures is discussed.

Introduction

The nominal Large Hadron Collider (LHC) luminosity of $10^{34}cm^{-2}s^{-1}$ corresponds to 10^9 proton-proton interactions per second. In case of the Compact Muon Solenoid (CMS) detector (see, e.g., Ref. [1]), about 10^{-7} of this event flow will be selected by a multi-level trigger for the off-line event processing and analysis. Thus, the data should be archived in a high performance storage system (SS) with the rate of ~100 Hz. Since the size of one CMS event written to the SS is supposed to be ~1 Mbyte, about 100 Mbyte of the information per second (or more than 1 Pbyte annually) has to be transferred to the SS.

A discovery of signals manifesting an evidence for the "new physics" requires a huge amount of data to be processed and thoroughly analyzed. Typically, $\sim 10^5$ pb^{-1} of the integral luminosity is needed in order to separate such manifestations against a huge background (see, e.g., Ref. [1]). It means that 10^9 CMS events (or 10^{15} bytes of information) have to be processed and analyzed.

All this sets hard requirements upon the data acquisition system and the SS. The extreme requirements for data storage, terms of computing and networking that the LHC experiments will need have put forward development and certification of a new concept called Grid computing (see, e.g., Ref. [2]). The Grid is a new form of a distributed system for research and development. Major challenges to be met by the scientific community involved in the LHC experiments are supposed to be overcome via communication and collaboration at a distance, network-distributed computing and data resources and remote software development and physics analysis (see, e.g., Ref. [3]).

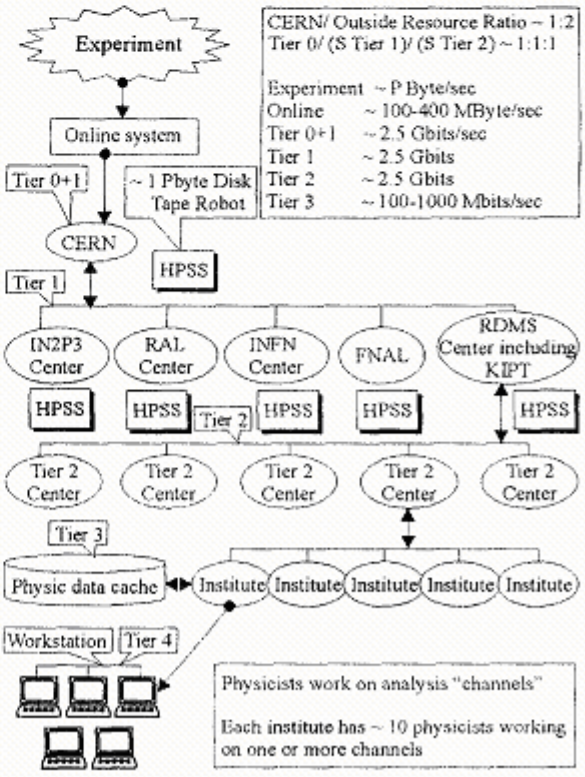

Fig.1. Hierarchy of Grid structures for LHC experiments [3].

Elements of the European Data Grid (EDG) middleware have been installed at the KIPT CMS Linux Cluster (KCLC) (see, e.g., Ref. [4]) which is a part of the Moscow distributed regional center (the hierarchy of the Grid regional centers is shown in Fig. 1). The KCLC architecture, operation of the batch system at the cluster, and installed EDG elements are briefly described below.

KIPT CMS Linux Cluster

The KCLC (see Fig. 2) is a specialized PC farm for conducting computation activities on the CMS physics. Ten dual nodes, viz., 4x800 MHz, 4x1000 MHz, and 10x1400 MHz Pentium III and 2x2000 MHz Xeon processors are allocated for either interactive or batch jobs providing the total PC farm CPU power of ~50×10^3 bogomips. The total amount of the hard disk drive (HDD) storage is more than 1 Tbyte.

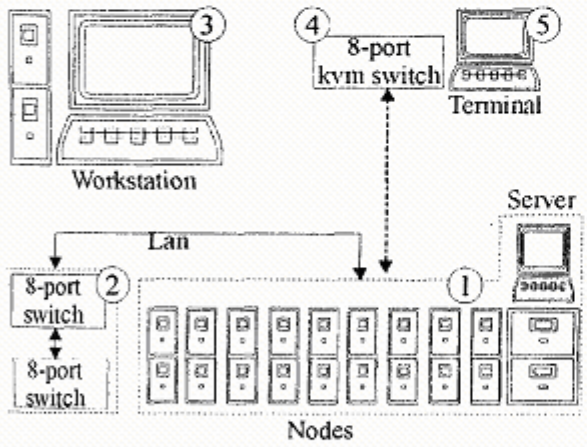

Fig.2. KCLC layout: 1 – server and nodes; 2 – two switches; 3 – workstation; 4 – kvm switch; 5 – terminal.

The nodes run Linux as operation system (OS) and PBS (see, e.g., Ref. [5]) as a batch system. Other important components of our cluster are the network file system (NFS) and network information service (NIS). They provide the joint access to the useful resources such as CERNLIB, ROOT, PYTHIA, GEANT and LHC++ and to specific CMS software such as CMSIM (a GEANT-based package for simulation of the CMS detector response), ORCA (an object oriented tool for CMS event reconstruction). 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. [5]) 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. 3. The batch system consists (see Ref. [5]) 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 10 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 20. The dual Intel Pentium III 800 MHz computer 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 20 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.

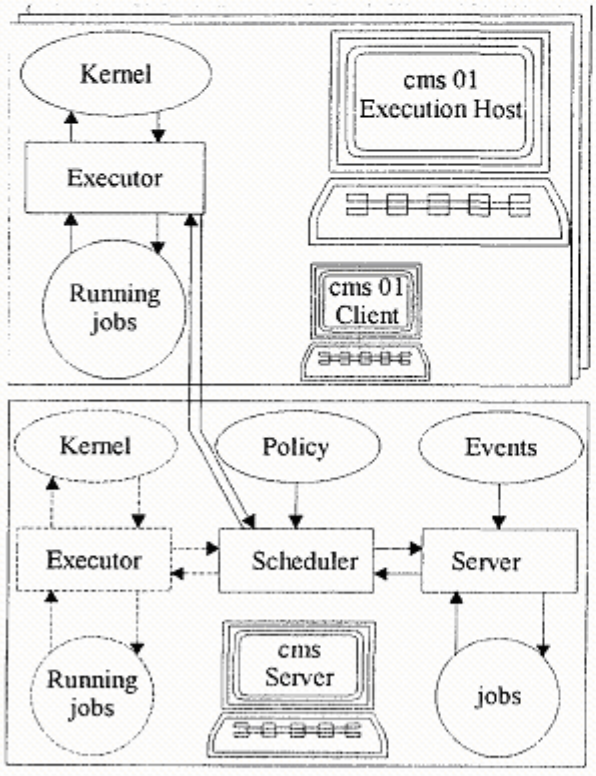

Fig.3. Use of PBS at the KCLC.

Adaptation of the KCLC to the Grid implies that the PBS scheduler has to be supplemented by a Grid scheduler (so-called Globus scheduler).

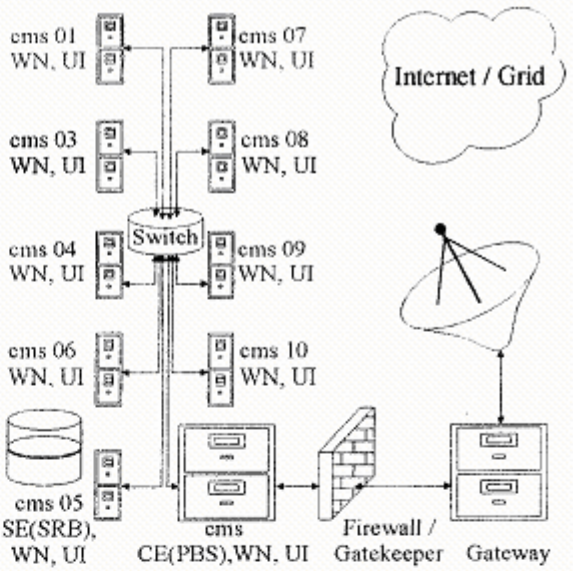

Fig.4. KCLC with Grid elements.

Grid middleware at the KCLC

The Grid is an environment that makes widely distributed computing resources transparently available to the end-user. Configuration of the site in terms of EDG software comprises four types of testbed machines: a computing element (CE), a worker node (WN), a storage element (SE) and a user interface (UI). Within the testbed, the replica catalogue (RC) and server are to be set up and maintained for each virtual organization (VO). At present, we have configured some elements of Grid infrastructure such as the storage resource broker, which is used for the generated data transfer and its place on the SE. The installation of the other Grid components is now in progress. For the KIPT CMS site, it includes setting up a Gatekeeper node, which acts as the portal for the site and as the front-end of the PBS controlled set of the WN's. Nodes 'cms01'–'cms10' run the UI software, which allows us to interact with the EDG infrastructure. Each such UI element provides an entry to log into the Grid to submit jobs and retrieve the result of job execution. Node 'cms' is our CE with the PBS server and gatekeeper on board. This machine provides users with computational resources. Gatekeeper is the front-end of CE.

This machine interacts with other sites of the Grid environment. It accepts jobs, dispatches them for the execution and returning to the output. Apart from being the UI elements, the nodes 'cms01'–'cms10' serve also as the WN's. These WN's are hidden behind the gatekeeper and are operated by the PBS locally. The gatekeeper conceals the details of this operation from the outside users, however, it is these nodes that are actually perform users' computations. They do not run any EDG daemons but have a client interface for accessing EDG services and information. At last, the SE is the node, which provides a unified access to the large storage spaces. In our case it is cms05 having the RAID controller of both levels 0 and 1 and the SRB software installed.

Conclusion

Installation and configuration of the EDG software at the KCLC is under way. Integration of the cluster into the Grid structures established for distributed CMS data storage and processing is planned for the nearest future.

References

1. The Compact Muon Solenoid Technical Proposal, CERN/LHCC 94-38, 15 December 1994.
2. I. Foster, C. Kesselman and S. Tuecke, The Anatomy of the Grid "Enabling Scalable Virtual Organization", Intl J. Supercomputer Applications, 2001, p.138.
3. D. Stickland, CMS Computing and Core-Software Report, In: Proc.VIth Annual RDMS CMS Collaboration Meeting, "Physics Program with the CMS Detector", IHEP, Moscow, Russia, December 19, 2001, p.293-308.
4. L.G.Levchuk, P.V.Sorokin, D.V.Soroka, and V.S.Trubnikov, NSC KIPT Linux Cluster for Computing within the CMS Physics Program, In: Problems of Atomic Science and Technology, Ser."Nuclear Physics Investigations". Kharkov 2002. N2(40), p.49-51.
5. http://www.pbspro.com/tech_overview.html

ЕЛЕМЕНТИ ПРОГРАМНОГО КОМПЛЕКСУ GRID НА CMS LINUX КЛАСТЕРІ ННЦ ХФТІ

Л.Г.Левчук, П.В.Сорокін, Д.В.Сорока, С.С.Зуб

ННЦ "Харківський фізико-технічний інститут",
вул.Академічна, 1, Харків, 61108
e-mail: stah@kipt.kharkov.ua

Розглянуто вимоги, що викликані фізичними цілями експериментів на LHC, які висувають на передній план концепцію розгалуженного зберігання (базується на Grid) та обробки великих масивів експериментальної інформації. Розглянуто функціювання виконуючої пакетної системи на CMS Linux кластері ННЦ ХФТІ. Викладено заходи щодо підготовки кластера до його інтеграції в структури Grid.