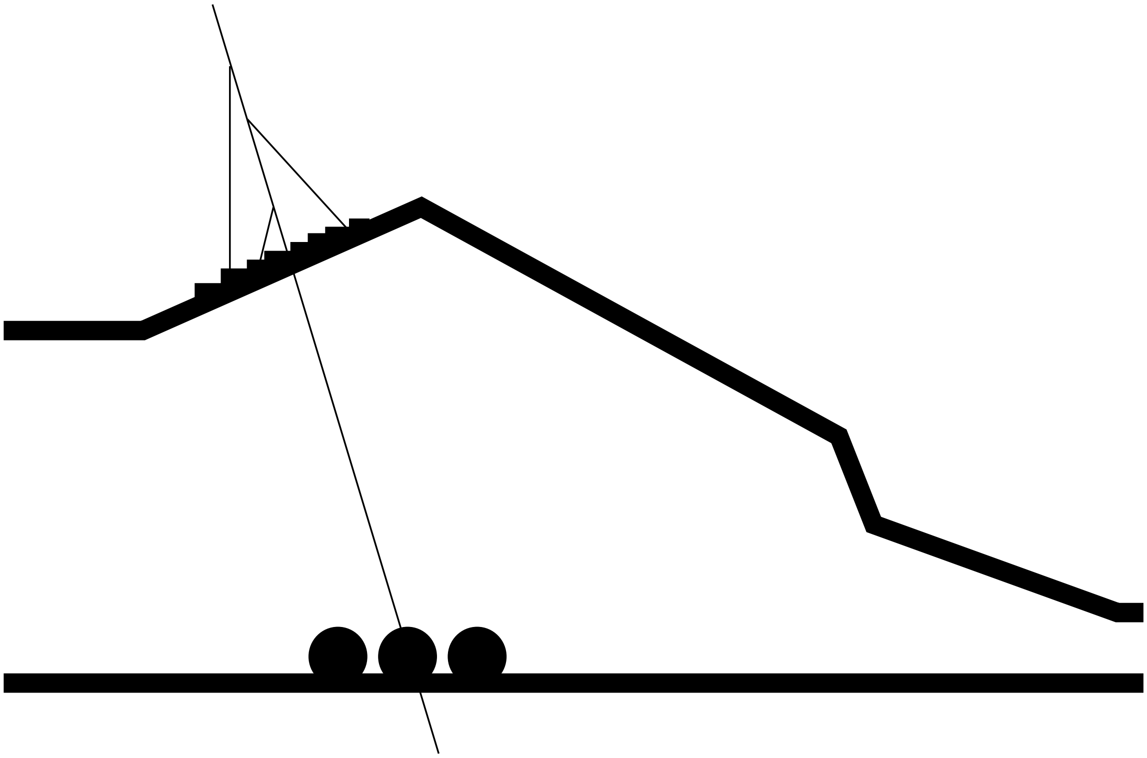
LNGS/TC-01/11
September 2011



U-LITE Unified LNGS IT Environment: a proposal for scientific computing at LNGS

S. Parlati, P. Spinnato, S. Stalio

INFN - Laboratori Nazionali del Gran Sasso

INFN - Istituto Nazionale di Fisica Nucleare
Laboratori Nazionali del Gran Sasso

LNGS/TC-01/11
September 2011

U-LITE Unified LNGS IT Environment: a proposal for scientific computing at LNGS

S. Parlati, P. Spinnato, S. Stalio

Laboratori Nazionali del Gran Sasso

Abstract

The Computing and Network Service at LNGS has developed a proposal for U-LITE, a centrally managed computing and storage infrastructure for the future experiments hosted by LNGS.

The system we propose includes, for each experiment requiring it, storage of experimental data, on-tape backup and/or off-site archiving, front end servers devoted to interactive login, and a computing infrastructure devoted to the execution of data analysis jobs, numerical simulations, and other computing jobs. The physical computing infrastructure will consist of a shared computing cluster, following a Grid-based approach, while job execution, drawing inspiration from the cloud computing paradigm, will be based on virtual machines.

In our proposal, the overall U-LITE management is accomplished by the LNGS computing service, relieving the burden of managing the IT hardware infrastructure from the collaborations. As far as financial aspects are concerned, each collaboration provides a share of IT infrastructure, with the guarantee that it will always have at its disposal the resources it contributed.

The aim of this project is to reduce purchase, maintenance and management costs for the computing and storage infrastructure of the experiments, thanks to the sharing of resources and their integration in a unified system.

1 Introduction

Since the beginning of the very first experiments at LNGS in the late 1980's, the Computing and Network Service (CNS) main goal has been to support scientific computing. A fast and reliable network, computing systems responding to the needs of experimental collaborations, assistance to experimental collaborations and single users for programming and utilization of scientific software have always been the LNGS CNS main goals in almost 25 years of activity.

The way these goals have been achieved has changed over the years, often following the natural evolution of technology. In the early 1990's, the era of the first generation experiments at LNGS (MACRO, EAS-TOP, LVD, GALLEX), the computing model was based on a highly centralized structure whose main components were a VMS cluster and the DECNET network, both managed by the CNS. When VMS and DECNET started their decline and new operating systems took over (Digital UNIX first, then, starting from the early 2000's, Linux) the LNGS computing environment became highly distributed and lost its homogeneity.

Each experimental collaboration developed its own applications on different hardware and software platforms and the CNS could not manage the experiments computing hosts any more, nor could offer a unique computing platform satisfying the requirements of every collaboration.

Nevertheless the CNS kept on offering computational resources to research groups in the forms of batch system (Condor, NQS, LSF) and disk storage and tape backup management for experimental data. Moreover the CNS kept on administering, for the benefit of all the LNGS user community, fundamental systems and services as the network infrastructure, basic network services, e-mail and web services, user support and so on.

In the last five years the LNGS CNS has made great efforts to strengthen the existing network and computing infrastructure, aiming at achieving high availability of IT services via system redundancy and other techniques.

The distributed computing model based on computer farms owned by experiments and managed by the experimental collaboration staff itself has been a good choice for large collaborations, that could afford to have skilled technical personnel on-site. This model failed for experiments that assigned the management of their computing resources at LNGS to people normally working at other INFN sites or people that did not have the necessary know-how.

More recently, new computing paradigms have arisen in the scientific community: both the Grid computing[1, 2] and the Cloud computing([1]) models rely on the idea of a common hardware infrastructure which is used by different user groups. In the Cloud computing model resources are assigned at the very moment the clients request them and, in contrast to Grid computing, applica-

[1] http://en.wikipedia.org/wiki/Cloud_computing

tions do not need to be specifically designed.
Following the Cloud paradigm the CNS developed a computing model where:

- a shared computing infrastructure is used by multiple user groups;
- experimental collaborations are free to choose the operating system and the software environment to be used for their computing needs;
- experimental collaborations are relieved the burden of managing the physical infrastructure. This task goes to the LNGS CNS.

In this document we give an overview of the new computing infrastructure, called U-LITE (Unified-LNGS IT Environment), the LNGS CNS is working on. U-LITE is mainly meant for future experiments at LNGS but can also be used by experiments that are already running.

2 Main features

2.1 System functions

In agreement with the model we just presented, computing resources will be shared among experimental collaborations. Experimental collaborations will access their storage areas exclusively, while storage areas will share a common storage infrastructure.
As far as data management are concerned, the U-LITE main functions will be:

- data transfer from the DAQ system
- middle and long term data storage
- data backup
- off-line archiving
- data distribution over the network

Data transfer from the DAQ system to the storage servers and data distribution towards other systems will have to be setup by the collaboration staff. Storage system management, backup and archiving will be taken care of by the CNS.
Regarding data processing, U-LITE will provide computing resources and custom software platforms for:

- detector simulation
- data analysis

From informal discussions with experimental collaborations, their interest for such a system has emerged, mainly because they would be relieved the burden of a task for which they seldom have internal experience.
Advantages of a centralized system managed by the CNS also include the stable and continuous presence of skilled personnel on-site, the savings guaranteed by resource optimization and the overcoming of problems related, as often happened in the past, to the integration of the experiment computing systems inside the LNGS IT infrastructure.

2.2 Target

The U-LITE project is mainly meant for future experiments at LNGS. Experimental collaborations will be encouraged to use the U-LITE infrastructure for their storage and computing needs. We wish that also experiments which are already running, or are in an advanced construction stage, and whose computing model is compatible with that of U-LITE may start using it.

2.3 Financial and organisational rationale

In our model, the collaborations provide the funding for acquiring the hardware which will constitute the shared computing and storage infrastructure, and the LNGS CNS will be in charge of managing the whole environment. Having the CNS provide for the management of U-LITE is a double plus for the collaborations. They are relieved the burden of managing their computing infrastructure with the guarantee,meanwhile, of having skilled personnel always next to the infrastructure taking care of its management.
Concerning the economical aspects, in contrast to a model in which each group has at its own disposal all (and only) the machines it bought, a shared approach is undoubtedly profitable. Namely, in the exclusive access model a collaboration must dimension its cluster in view of the peak utilisation, in order to avoid risks of resource saturation. This implies that, on average, only a fraction of the cluster is actually used. Conversely, in the shared model only the machines necessary to cope with the average use (plus a reasonable tolerance margin) need to be bought. In fact, in ULITE each collaboration has the guarantee of always having at its disposal the resources it financed, nevertheless it will be able to access unused resources provided by other groups (as long as the latter do not require them). This results in usage optimisation and cost reduction. In order to guarantee availability of resources even when a large number of collaborations reach simultaneously the utilisation peak, the CNS will make extra resources available, so that these situations will be coped with, avoiding overload.

2.4 Existing skills

The CNS staff at LNGS has a valuable experience in managing all the subsystems that are needed in order to implement the computing infrastructure we propose. Namely, the CNS:

- manages the LNGS network infrastructure starting from the physical layer up to the application layer, including the connection towards the GARR geographic network;
- manages multiple storage systems (RAID 5 or RAID 6 systems connected to data servers via Fiber Channel or iSCSI) containing mainly experimental data and simulation results for a grand total, today, of approximately 150TB;
- manages two tape libraries for data backup. Each tape library can contain up to 120 LTO4 tapes (approximate capacity: 1TB of data per tape);
- manages the data servers (NFS and AFS) that export data towards interactive login or batch job processing nodes;
- manages some interactive login servers;
- manages a small ”traditional” batch job submission farm;
- manages all the basic network services (DNS, DHCP, SMTP, IMAP, LDAP, Kerberos etc.) that are needed for the proper operation of the LNGS LAN.

All the mentioned IT services operate with high availability standards and use redundant hardware and software components.
The CNS staff has also gained a long experience working with host virtualization, with the primary purpose of offering better network services to the LNGS user community.
In the years 2009 and 2010 we developed CRM[3], a prototype for a computing farm to be used by workgroups with different computing needs. The idea, partially inspired by the INFN CNAF *WNoDeS* project[4], was to realize a computing cluster where the computing nodes are virtual hosts to be switched on at the very moment a request for a certain type of resource arrive, and to be turned off automatically when idle. This experience has been crucial for the U-LITE project which is based, for the computing part, on a cluster of virtual nodes that is dynamically reconfigured according to resource requests.

2.5 Data center characteristics

We propose the LNGS main computing room as the site for U-LITE storage and computing systems.

The LNGS computing center is distributed between two different rooms belonging to two different buildings in the external part of the laboratories.
The two buildings are approximately 200 meters apart and are interconnected by one 10Gb/s and two 1Gb/s optical fiber connections that follow two different paths.
The smaller room (called saletta router) has an area of 28 square meters. It hosts the LNGS border router, some servers containing several virtual machines dedicated to network services and one of the two tape libraries that keep experimental data, server and workstation backups.
The main computing room, where the computing and storage systems for the U-LITE project will be placed, is much larger (about 250 square meters), and is situated in the main building, the same building where most of the researchers working at LNGS have their offices.
The LNGS main computing room hosts:

- devices granting the network interconnection with other LNGS areas and with the GARR geographic network;
- data servers (NFS and AFS) managed by the Network and Computing Service staff;
- most of the servers that provide network services:
- the computing farm managed by the Network and Computing Service staff;
- various computing farms belonging to experimental collaborations and research groups: Borexino, Cuore, Icarus, Opera, Warp, Xenon, the theoretical group, the Università degli Studi dell'Aquila;
- one of the tape libraries that keep experimental data, server and workstation backups.

Today these systems use 15 sparsely filled racks. Resources needed for U-LITE will partly add to existing systems and in part replace existing systems. Part of the U-LITE resources will come from recycling available hardware.
A tape room is also available for long term storage of data tapes.

3 Architecture

The goal of U-LITE is to provide an environment able to accomplish the main computational tasks that collaborations active at LNGS need to perform. Such environment includes a computing cluster and a Storage Area Network dimensioned to collaboration needs and expandable as needs change.
The physical computing infrastructure will consist of a shared computing

cluster while job execution, drawing inspiration from the cloud computing paradigm, will be based on virtual machines([2]) (VM hereafter). This will let users access resources transparently and system managers operate the computing environment optimally.

Exploiting virtualisation provides several advantages. Namely, it makes using the cluster simpler from the user point of view. In fact the user, after developing a fully-equipped VM template, only has to replicate it and execute on such copies the data analysis, simulations and whatever else is needed for scientific computing. This allows the user to exploit the cluster resources with no need for adapting the user programs and applications to the cluster hardware. On the contrary, a VM-based approach makes very easy to adapt the computing environment to the user needs. Moreover, the use of VMs fosters the optimal use of a state-of-the-art cluster, i.e., a cluster made up of multicore multiprocessor machines (motherboards with up to 48 cores are within reach at reasonable costs nowadays). This is done transparently to the user by the VM management software, which distributes the VMs over the hardware resources, therefore optimizing the use of available resources.

Concerning management of experimental data, the storage area of each collaboration is accessed exclusively by collaboration members, while the physical disks will be hosted inside a common infrastructure managed by the LNGS Computing Service. This guarantees confidentiality for the data of each collaboration and, at the same time, a simpler and more effective management of the storage infrastructure.

Figure 1 sketches the architecture of U-LITE and shows how it works. The front-end servers provide interactive access to the cluster. Each collaboration has its servers, accessed exclusively by collaboration members. Such servers may be VMs, generated from templates developed by the collaboration itself. Front end servers, which can be remotely accessed, are devoted to software development and job submission.

Data analysis or simulations jobs submitted from the front end will be executed by VMs running on the computing farm. The management system of the computing farm will activate the VM, pass it the input parameters provided by the user, start the job, save the output data, and inform the user when the job ends.

The storage servers (here again each collaboration owns one or more servers exclusively accessed by its members) are used for all operations involving management of experimental data:

- receiving raw data coming via LAN from the data acquisition detectors located in the underground experimental halls
- preprocessing raw data

[2] http://en.wikipedia.org/wiki/Virtual_machine

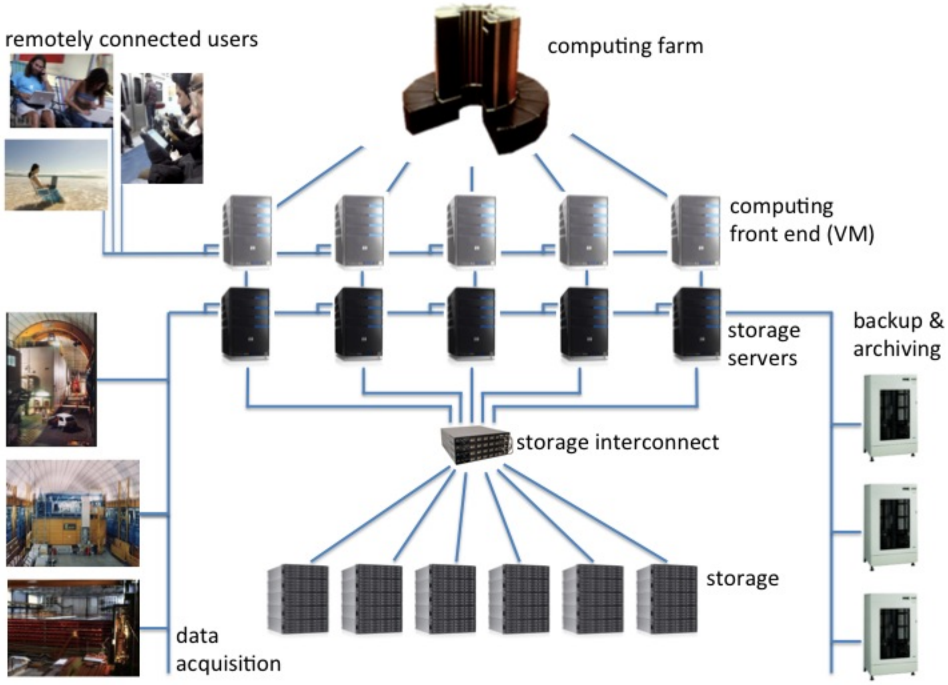


Figure 1: U-LITE functional diagram.

- storing data in the U-LITE storage area network
- data backup ad archiving

3.1 Computing environment

In our model, each collaboration develops its own worker node template according to the guidelines provided by the CNS, or alternatively asks for a standard template. The collaboration template is then taken over by the CNS that takes care of creating VM clones, which the collaboration will be able to use in the computing cluster. Job execution is performed by submission to a batch queue system managed by the open source software Torque/Maui([3]). The CNS staff developed an algorithm that interacts with Torque/Maui in order to schedule the execution of queued jobs according to the following guidelines:

- to guarantee for each collaboration immediate access to resources which it has paid for
- to distribute homogeneously free resources to all collaborations needing them

[3] http://www.clusterresources.com/products.php

- to maximise exploitation of available resources

Finally, the tasks related to VM management is accomplished by the CRM software (see section 2.4), developed by CNS. CRM takes care of switching on and providing the correct VM for the scheduled job, and putting it in suspend mode or switching it off altogether at the end of the VM job cycle.

3.2 Data storage

Researchers access their experiment data through one or more storage servers that are meant for the exclusive use of the collaboration. Storage servers are connected to storage systems via Fibre Channel or iSCSI. The storage servers main functions are:

- data copy from the DAQ system;
- short and middle term data storage;
- data preprocessing;
- raw and/or preprocessed data distribution towards the computing nodes, the backup system, the internet and towards long term storage media;
- DBMS hosting.

The storage servers have a very important function, being the connection between the DAQ system and the computing environment; for this reason their management is responsability of the experimental collaboration personnel. Experimental collaborations are free to choose the storage server hardware, its operating system and the software tools to be used for data management, as long as good security standards are satisfied. The CNS staff will give advice and guidelines for storage servers configuration and will help choosing the best strategies and tools for their operation.
The CNS staff must be able to access storage servers with administrative privileges to operate in case of emergency.
The experimental data management is responsability of the experimental collaborations as well.

3.2.1 Data backup and archiving

As for experimental data backup, two separate periodic backup schedules will be activated for each data set, to ensure a higher degree of data protection. One of the backup schedules will write data to the tape library in the LNGS main computing room, the other one will be performed on the tape library in the saletta router (in the centro direzionale building).

In order to minimize tape usage only incremental backups will be performed, while no periodic full backups are planned.
Large extents of data are not guaranteed to be kept on-line (inside the tape library). In case of lack of tape slots inside the libraries it may be necessary to manually extract tapes that contain older data.
Experimental data may be archived for long term storage on magnetic tapes to be kept off-line in a dedicated room.
Data is written on tape using an open-source format in order to allow for long term easy read-out also outside LNGS.

3.3 Authorization and Authentication

The centralized authentication and authorization tools, based on Kerberos 5 and LDAP allow for great flexibility in granting secure and personalized access to available resources. This set of tools is used to grant access to U-LITE front-end nodes and worker nodes as well.

3.4 System monitoring and availability

Almost all the U-LITE subsystems are continuously monitored using the Nagios network monitoring software.
In the event of hardware or software failures, alerts are received and handled by the U-LITE technical staff.
Continuous operation of all U-LITE subsystems is, of course, desired, but it must not be critical for the experiments data workflow.
U-LITE must be decoupled from the experiment DAQ system, each experiment is expected to have a data buffer near the DAQ that must be capable to store one week of data in the extremely improbable case of serious problems during week-ends or other holidays. Some considerations on U-LITE subsystems availability are given below.

3.4.1 Disk storage

Detector data duplication on different storage systems for better availability is usually impractical and quite expensive while not impossible a priori.
Other precautions can be taken in order to reach this objective, starting from an accurate choice of the storage hardware.
The storage systems that U-LITE will use have the following characteristics:

- redundant power supply.

- redundant FC and/or iSCSI controller.

- disks will be configured using RAID level 6 (two different parity data sets, two disks in a set can break at the same time without service interruption).
- there will be one hot spare disks every 20 active disks (approximately).
- some spare disks will be kept in the LNGS computing room, ready to replace broken disks on the system.

A next-business-day service agreement with the vendor will always be active for storage systems.

3.4.2 Storage servers

Experiments should have two storage servers connected to the same storage areas through different controllers on the same storage system. When a server, a controller or a SAN branch fails there should be some automatic or manual procedure to activate the backup server or, in case of active redundancy, to direct client requests only towards the working branch of the system.

3.4.3 Backup

The double backup policy described in section 3.2.1 ensures that if one backup system fails or a tape becomes unreadable, backup data will be still available. The fact that the two backup systems are positioned in different buildings further increases experimental data availability.

3.4.4 Computing cluster

The computing cluster is a non critical component in a non critical system, nevertheless some remarks on its availability need to be made.
The failure of a physical computing host or some worker nodes is not an issue for the normal operation of the cluster.
Conversely, the host that runs the batch system controller processes is a single point of failure for the computing part of U-LITE, so an automatic failover mechanism on a backup node has been implemented.
Also, the storage system that holds the virtual working node images, reliable as it may be, is a single point of failure. To avoid this, it must be replicated or working node images must be distributed over two or more different storage systems.
The best ways to implement redundancy of working node images is currently under investigation and will be implemented in the near future, as soon as the required hardware (a separate disk system) will be available.

4 Location, network and technological plants

The LNGS main computing room will host the U-LITE computing and storage systems. This computing room presently hosts the computing infrastructures managed by the computing service and the current experimental farms. Space availability and accessibility are not an issue, therefore the current situation allows the displacement of new technological plants.

4.1 Network

The connection between the U-LITE dataflow and the LNGS LAN is sketched in figure 2.

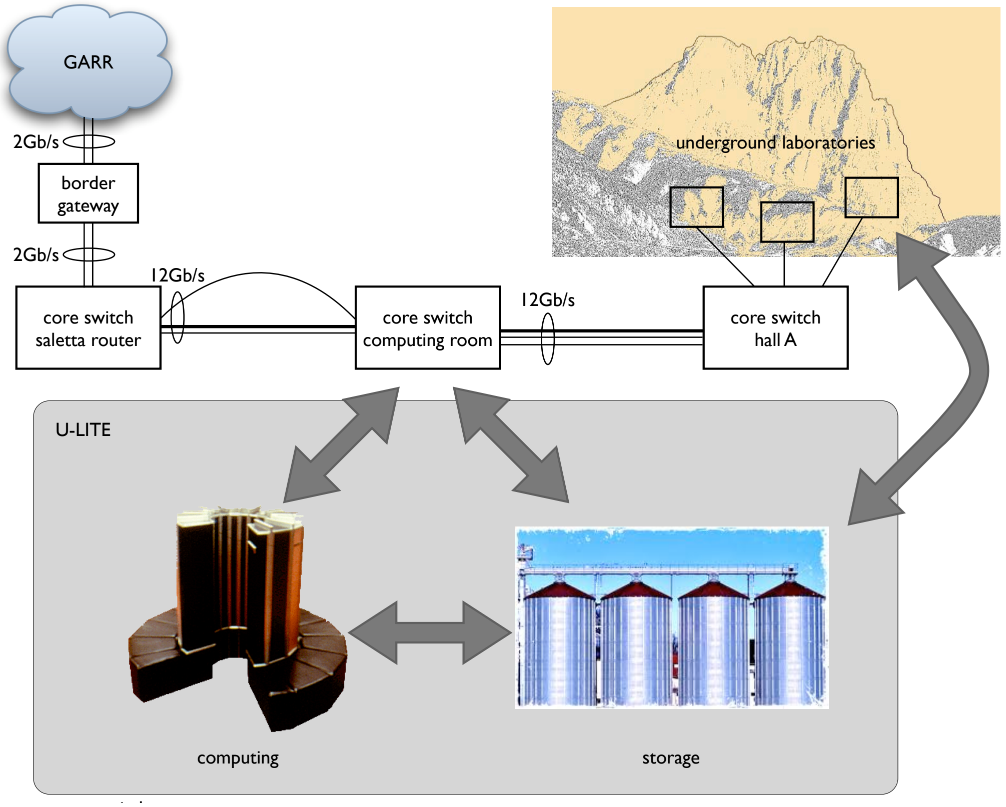


Figure 2: U-LITE dataflow and its connection with the LNGS network.

The main computing room hosts one of the two core switches of the external laboratories, where most of the external buildings and the underground laboratories LAN branches are connected.
The link with the underground laboratories is currently composed of 3 aggregated lines with a total capacity of 12 Gb/s.

The main computing room is linked to the second core switch (hosted in saletta router) by means of 3 aggregated lines, that follow two different paths, with a total capacity of 12 Gb/s.

The connection to GARR, the Italian network service provider for research institutions, is provided by two 1Gb/s Ethernet links.

Concerning the U-LITE network infrastructure, the front-end servers and the computing nodes are connected at Gbit speed by means of one or more switches, connected in turn to the core switch of the main computing room. Similarly, the storage servers are connected at Gbit speed to their own switches, which are also directly connected to the core switch.

Storage server connection to the acquisition environment in the underground laboratories is possible either through the public LAN or through dedicated point-to-point connections.

The network services and the underlying network infrastructure are included in the CNS monitoring system based on Nagios([4]). In the event of failures, alerts are received and handled by the CNS technical staff.

4.2 Technological plants

4.2.1 Electrical plant

Presently, the power consumption in the computing room is about 30KW (UPS-supplied) due to computers and active devices, while the cooling system adds up 40 more KW not supplied by UPS. The UPS infrastructure is composed by 3 devices, each providing 60KVA, which supply the whole building, including the computing room and the offices. The total load on the UPS system is now equivalent to about one third of peak power. The possibility of having an UPS system devoted exclusively to the computing room is under evaluation.

In the near future a new electricity distribution system will be installed in the computing room; it will include newly designed circuit breakers for better protection from electrical surges. At present it is not possible to accurately predict the power absorbed by the U-LITE infrastructure but, even with twice the current power consumption, the electrical distribution system of the computing room will be adequate.

4.2.2 Cooling systems

The computing room is cooled by three independent cooling systems: the centralized system that serves the entire building and two dedicated plants. Inside the computing room there are:

[4] http://www.nagios.org

- 7 machines with cooling capacity of 22 KW each, fed by cold water produced by the chillers of the general plant

- an old dedicated air conditioner with cooling capacity of about 25 KW installed in the mid-nineties

- a new dedicated air conditioner with cooling capacity of about 40 KW installed in 2010

A third conditioner will be installed in the near future, in order to have the room completely independent from the general cooling plant.

4.2.3 Fire extinguishers

The computing room is equipped with a fire suppression system based on Freon R-23 that covers the entire area.

4.2.4 Alarm systems

All technological systems are monitored by the LNGS supervision and overall control system. Alarms are received and handled on a 24/7 basis by the on-call staff within the various services.

5 Human resources

Experience with the farms belonging to experimental collaborations showed that one of the factors that has most penalized system reliability is the lack of local qualified personnel able to effectively perform computer administration or manage failures and contingencies.
One of the main strengths of U-LITE is to offer a service staffed by experienced, qualified personnel based at LNGS.
The staff involved in the setup and long-term management of the project belongs to the CNS. Moreover, an LNGS researcher who collaborates with the CNS in the field of high performance scientific computing will give advice and external feedback.
It is desirable that, during the project start-up, collaborators with training contracts will be involved in order to give a significant contribution to the development of U-LITE in terms of innovative ideas, enthusiasm and productivity.

5.1 Responsibilities and duties

5.1.1 Management and coordination

- local coordination
- contacts with the experimental Collaborations, INFN Commissions, etc.
- management of procedures for purchasing
- staff recruitment

Sandra Parlati (permanent staff)

5.1.2 Technical and operational responsibilities

- responsibility for hardware and software architecture
- responsibility for design and development
- personnel coordination

Piero Spinnato (temporary staff)
Stefano Stalio (permanent staff)

5.1.3 Supervision and control

- advisor for compliance with the objectives of the project
- advisor for the quality of service

Giuseppe Di Carlo (permanent staff)

5.1.4 Technological systems responsibilities

- responsibility for electrical, cooling, fire protection, supervision and control systems

The Technical Division is responsible for managing the general systems of the Laboratory. On-call shifts among the staff of the Division of Technical Services ensure the timely intervention of qualified personnel in the event of major emergencies related to such installations.

5.2 Technical staff

5.2.1 Basic services

- Staff of the CNS to ensure continuous operation of basic services such as network, security management, e-mail, etc..

Roberto Giuliani (permanent staff)
Sandra Parlati (permanent staff)
Piero Spinnato (temporary staff)
Stefano Stalio (permanent staff)
Nazzareno Taborgna (permanent staff)

5.2.2 Dedicated staff

- hardware and software installation, configuration, work on all components of U-LITE
- coordination of external personnel for machine installation, repair and maintenance
- managing hardware, monitoring systems and alarms

Stefano Stalio (permanent staff)
Piero Spinnato (temporary staff)

5.2.3 Human resources requirements

The human resources needed for the deployment of U-LITE are:

- 3 FTE, distributed among the existing staff and trainees, for the setup of the U-LITE project and during periods of increased activity
- 1.5 FTE distributed between two or more members of the CNS staff when U-LITE will be fully operational. Staff presence will be ensured during working hours, including holiday periods.

Based on the analysis on system availability in section 3.4, there is no need for personnel attendance outside normal working hours.

Acknowledgements

We acknowledge INFN for funding the grant assigned to Enrico Salvo, who carried out the initial development of the CRM software, under the supervision of Stefano Stalio. We are also very grateful to Giuseppe di Carlo for the inspiring discussions we had about the general infrastructure of U-LITE.

References

[1] Foster, I. and Kesselman, C. (eds), *The Grid: Blueprint for a New Computing infrastructure*, 2nd ed., Morgan Kaufmann (2004)

[2] Stockinger, H., *Defining the grid: a snapshot on the current view*, J. Supercomput., **42**, 3 (2007)

[3] Salvo, E., *VHPC, a proposal for batch computing on virtual hosts*, `http://www.lngs.infn.it/lngs_infn/contents/lngs_en/research/experiments_scientific_info/conferences_seminars/seminars/LNGSseminar2009Salvo.pdf` , LNGS, December 2009

[4] Italiano, A. and Salomoni, D., *WNoDeS, a tool for integrated Grid/Cloud access and computing farm virtualization*, CHEP 2010, Taipei, Taiwan, October 2010