ISTITUTO NAZIONALE DI FISICA NUCLEARE

Laboratori Nazionali del Gran Sasso

INFN-17-06/LNGS
28th March 2017

U-LITE, 6 years of scientific computing at LNGS

Barbara Demin[1], Sandra Parlati[1], Piero Spinnato[2], Stefano Stalio[1]

[1)]INFN-LNGS, Via Acitelli 22, I-67010 l'Aquila, Italy
[2)]INFN-TIFPA, Via Sommarive 14, I-38123 Trento, Italy

Abstract

The computing infrastructure of Laboratori Nazionali del Gran Sasso (LNGS) is the primary platform for data storage, analysis, computing and simulation of the LNGS-based experiments, which are part of the research activities of the Istituto Nazionale di Fisica Nucleare (INFN). Groups running such experiments have diverse needs, and adopt different approaches in developing the computing frameworks that support their activities. Since the emergence of the Cloud paradigm, the Computing and Network Service has built on its experience in operating and managing the LNGS computing infrastructure to develop U-LITE, a versatile environment apt at hosting such varied ecosystem and providing LNGS scientific users a familiar computing interface which hides all the complexities of a modern data center management. Over the last 6 years U-LITE has proved as a valuable tool for the LNGS experiments, and provides an example of effective use of the Cloud computing approach in a real scientific context.

PACS.: 07.05.Kf, 29.85.-c

To be submitted to International Journal of Cloud Applications and Computing (IJCAC)

*Published by **SIS–Pubblicazioni***
Laboratori Nazionali di Frascati

1 INTRODUCTION

LNGS is a world-renowned research site[1], where cutting-edge research in several branches of particle physics spanning from, e.g., search for dark matter[2] to neutrino physical properties[3], or studies on nuclear reactions in stars[4] are carried on within the World-largest underground laboratory for particle physics. Data acquired from the experimental apparatuses need to be stored and analyzed. Detection systems and processes need to be modeled and simulated. Such diverse tasks require a suitable computing environment. Ever since research activities begun at LNGS in the late 1980's, the Labs Computing and Network Service (CNS) has been committed to provide the research groups active at LNGS with the IT infrastructure and the professional support required for their scientific computing needs.

The way such goal has been achieved has changed over the years, often following the natural evolution of technology. The computing model of the first experiments run at LNGS was based on a highly centralized structure, whose main components were a VMS cluster and the DECNET network, both managed by the CNS. Later on, technological evolution brought to the rise of a more inhomogeneous computing model, based on different UNIX/Linux clusters devised by staff of the experiments built at LNGS in the late 1990's and early 2000's. The resulting variety of environments was too large to be managed by the CNS. Each experiment managed its computing cluster, while the CNS was in charge of providing resources for general use, including batch systems (Condor, NQS, LSF) for those who had not a cluster of their own, disk storage and tape backup management for experimental data, as well as the basic IT services.

This arrangement, though convenient and effective in some cases, was obviously inefficient in terms of resource and staff utilization. The rise of virtualization and the consequent advent of cloud computing as a paradigm to abstract resources and provide multiple user-defined environments on a common infrastructure, was a turning point in IT resource provisioning. CNS recognized the high potential of this approach to overcome the fragmented situation that LNGS was experiencing. Re-unifying the IT substrate management under the sole responsibility of CNS, while providing each research group the most suitable and personalized computing environment for its scientific needs was the design goal for a new LNGS computing platform that virtualization tools would allow to achieve. The outcome of such effort was the Unified Lngs IT Environment (U-LITE). Characterizing features of U-LITE are a user-transparent coupling of a well-known batch system (Torque/Maui[5]) with a hypervisor (Proxmox[6]) for a seamless provisioning of virtualized resources dynamically adapting to user requests, and an overall management of the whole data stream, from the data transmitted by the data acquisition hardware in the underground Labs up to the data storage and backup systems in the Data Center located in the external Labs.

The choice of the abovementioned virtualization tools, which only run on Unix-like Operating Systems, apparently limits U-LITE versatility. In fact, in our environment the vast majority of users develops its computational tools on Linux platforms, and being limited to Linux has not been experienced as a problem by our users.

The technical features of U-LITE will be described in the next sections, together with an analysis of the usage data collected over the six years the system has been in use, demonstrating how the adoption of a cloud paradigm in a research environment can be profitable in terms of resource optimization.

2 MAIN FEATURES

U-LITE has been developed with a bottom-up approach, having in mind the needs and requirements of its final users, the scientific community of LNGS. In this perspective, its features are naturally adhering to those of a private cloud[7)], as the infrastructure is provisioned for being primarily used by the LNGS-based researchers. In fact, access to U-LITE has also been granted to a number of external users. Yet, access procedures for external users are based on scientific collaboration agreements, and do not involve financial transactions, therefore they are quite dissimilar to those commonly adopted for public clouds access.

A characterizing feature of U-LITE is its data-flow model. A key element in the *raison-d'être* of LNGS is the production of experimental data in the underground labs, which may contain evidence for yet undiscovered physical phenomena. The process leading to a scientific discovery in this context passes through a careful transport of such data from source (the experimental apparatuses underground) to destination (the U-LITE SAN within the data center in the outside labs), ensuring no data loss, and its scrupulous analysis with the software resources developed by LNGS scientists and running on the U-LITE computing platform. It is therefore clear how precious a resource are scientific data in our context, and the extreme care the CNS devotes to protect experimental data from loss.

Concerning management of such experimental data, the storage area of each collaboration is accessed exclusively by collaboration members, while the physical disks are hosted inside a common infrastructure managed by the CNS. This guarantees confidentiality for the data of each collaboration and, at the same time, a simple and effective management of the storage infrastructure.

In order to work on its data, 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. 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 the cluster, as the decoupling of the hardware and software layers allows at the same time for a straightforward add-on of computing nodes, and replacement of obsolete resources, ensuring scalability and availability of state-of-the-art computing technology.

The financial model for computing nodes is yet another peculiar feature of U-LITE. Research groups, in coordination with the CNS estimate their average needs, and acquire computing nodes using their own funds. Nodes are integrated in a common pool, and resources are assigned on request at run-time *regardless of the purchasing group*. This collaborative model, where groups contribute based on their possibilities, and benefit of the resource based on their needs, proved successful in our context. It would indeed be interesting to explore other contexts where this model could be applied. It also brings the obvious economic advantage deriving from resource sharing, as each group only needs to dimension its purchases in view of average use, utilization peaks being smoothed out thanks to the availability of the global infrastructure. Concerning storage, as storage needs of U-LITE users are essentially static, groups only contribute to the storage hardware they need. As mentioned above in this section, storage is integrated in a global SAN, and each group has its private storage area, corresponding to the storage it paid for.

3 HARDWARE ARCHITECTURE

Physical nodes (simply “nodes” hereafter), storage area and network interconnect which constitute the physical layer of U-LITE are ordinary data center state-of-the-art hardware components. What is peculiar to U-LITE is its being directly connected with the data acquisition systems of the experiments located in the underground laboratories of LNGS. This intimately links U-LITE with the experimental setups, characterizing it as a private cloud for the LNGS scientific community.

Major collaborations have their own fiber optics links where data flow from data acquisition system to the collaboration storage server in the data center. Link speed for each collaboration is (as of December 2016) 1 or 10 Gb/s, according to the experiment needs. Storage servers are connected to the U-LITE SAN via Fiber channel or iSCSI links.

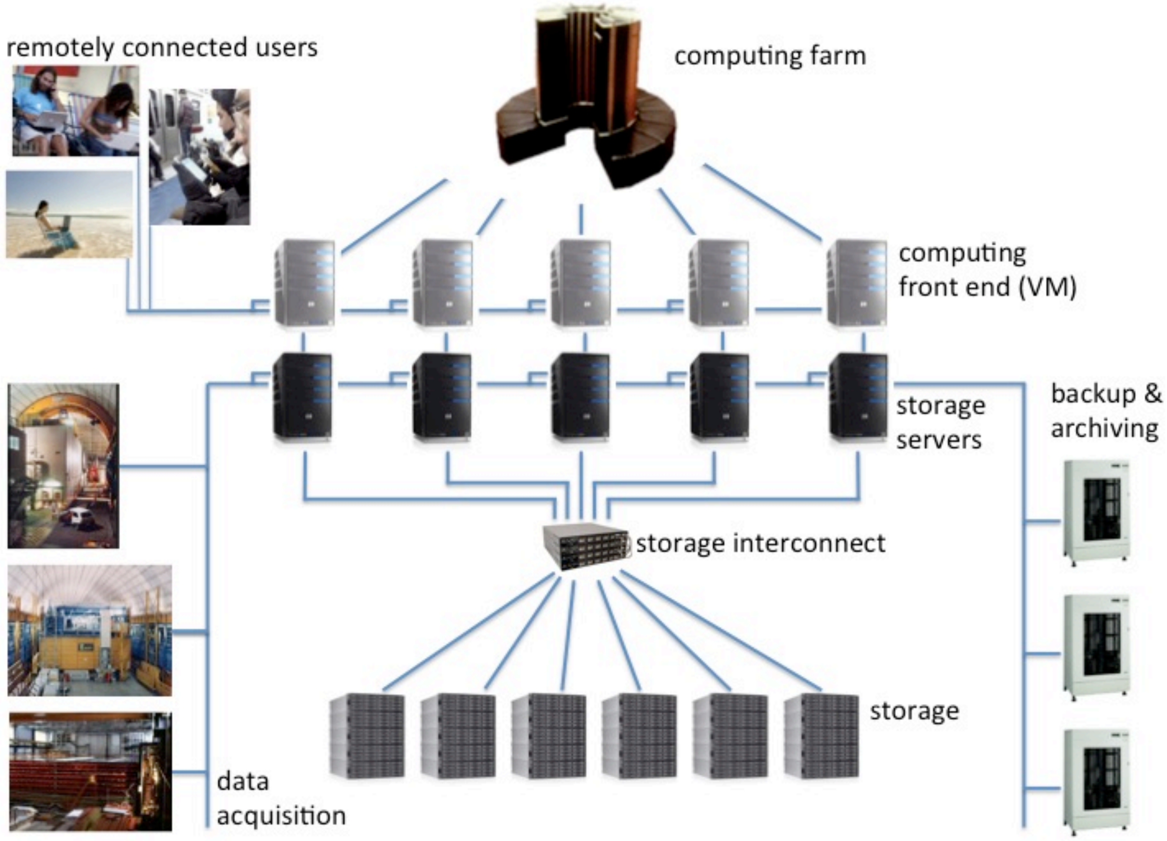

Figure 1: U-LITE conceptual architecture

Fiber channel speed ranges between 2 and 16 Gb/s, iSCSI speed is 1 Gb/s. Storage server functions, apart from data copy from the data acquisition system and management of data stored in the SAN, include data processing, raw and preprocessed data distribution towards the computing nodes, the backup system, the internet and long term storage media, DBMS hosting. Currently, the U-LITE SAN amounts to approximately 1.6 PB. Each collaboration exclusively accesses its own share, and share sizes are extremely varied, ranging from 1 TB to 1000 TB. Storage reliability is attained through redundant power supply and redundant Fiber channel and/or iSCSI controller in the storage boxes, and RAID6 disk organization with hot spare.

Long-term storage and backup of scientific data are performed on two tape libraries in different locations within LNGS. Each data set is duplicated on each tape library, with separate periodic backup schedules for each tape library, in order to ensure a high degree of data protection. In order to minimize tape usage, only incremental backups are performed. Experimental data to be archived for long-term storage are kept off-line on magnetic tapes.

The U-LITE SAN also serves as a storage platform for VM images. Three independent storage areas containing VM images are shared among all the computing hosts. Using RAID storage systems with redundant controllers and power supplies reduces the risk of failure of the VM repositories, and the only single point of failure is represented by the single switch that makes VM images available to the computing hosts using the iSCSI protocol. Different storage and network architectures deploying higher availability standards might be investigated in the future. Logically, images are stored as LVM volumes available to all hosts, therefore each VM can always be started on any computing host, depending on the overall system usage. Live migration is possible and has been implemented, but is not used today. This storage setup is very flexible and allows for very fast VM provisioning, at the cost of sub-optimal disk I/O performances, since the system disk is not local to the computing host. Yet, VM images are only used to load the OS and sometimes the applications to be run. A storage area local to the hosting server, acting as a high performance, temporary data buffer can be created and mounted on a VM at boot time and destroyed when the VM is powered off.

The U-LITE computing platform is not monolithic, being constituted of a number of nodes produced by a variety of different vendors. All nodes feature multicore processors, Intel or AMD, currently amounting to a total of 808 (partially hyperthreaded) distributed over 20 nodes, with clock frequency ranging around 2.5 GHz. Average RAM per core (or thread) is 2 GB, and total node RAM ranges between 24 and 128 GB, providing ample room for VM memory allocation. Node hard disk storage, which as already mentioned is mainly intended for fast, volatile VM disk space, ranges between 100 GB and 1000 GB. Nodes run the open source Debian-based Proxmox VE (currently 4.x) virtualization platform, which uses the KVM hypervisor for full virtualization and LXC as container technology.

The front-end servers and the computing nodes are connected at either 1 Gb/s or 10 Gb/s speed to the U-LITE network switch, connected in turn at 10 Gb/s to the core switch of the LNGS data center. Similarly, the storage servers are connected at 1 Gb/s or 10 Gb/s speed to their own storage switches, which are also directly connected to the core switch.

All U-LITE components are monitored by the CNS monitoring system based on Nagios[8].

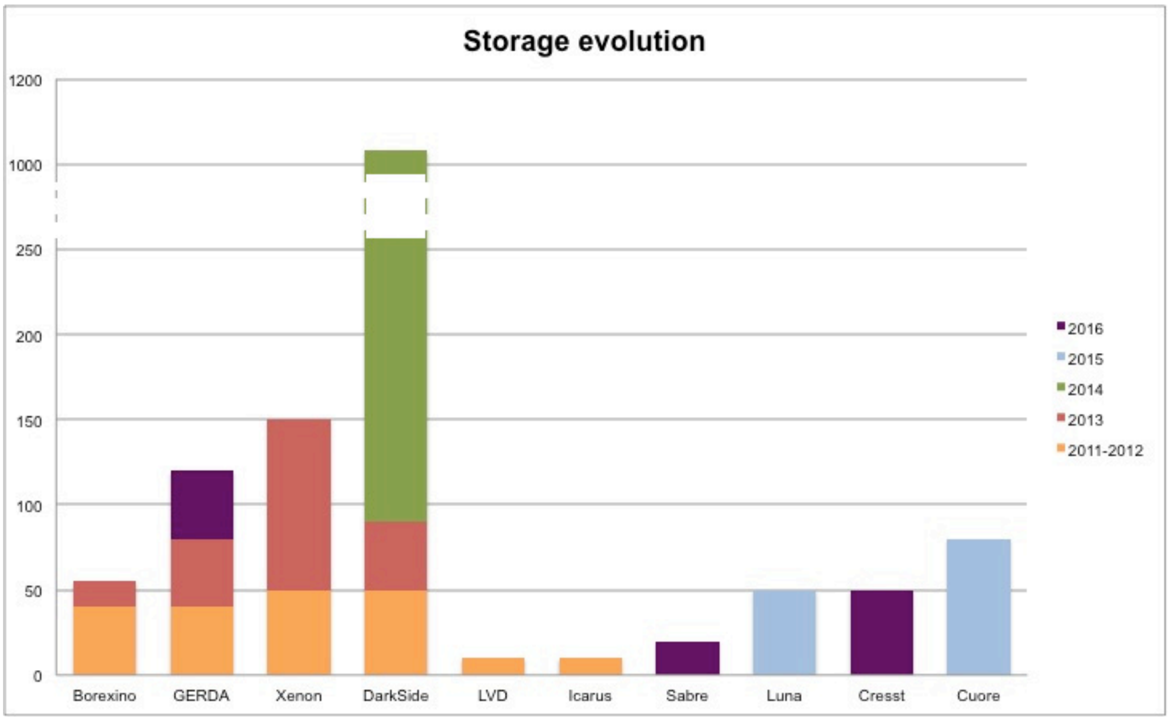

Figure 2: Chronological evolution of the storage in U-LITE. Contributions are shown per group.

4 HARDWARE RESOURCE EVOLUTION

We give here a sketch of how U-LITE hardware resources have evolved across time. In fig. 2 we show the evolution of the storage resource. As mentioned in the previous section, storage is funded directly by groups, therefore storage shares are statically assigned to them. Accordingly, we show storage evolution data for each group. Typically, when new groups enter the U-LITE community, they provide funding for the storage they will need for the following years. In a few cases (Gerda, Xenon, Darkside), storage is incremented after the first provision. U-LITE storage needs are different among groups; some use storage as a buffer for fast access, while the data bulk is stored off-line or elsewhere, whereas other groups keep all their datasets on-line in U-LITE.

A small amount of storage (about 10 TB) has been provided directly by the CNS, mainly for U-LITE housekeeping (VM image repository, VM scratch areas).

In fig. 3 we show the overall storage evolution. We can see that storage has incremented steadily across time, except in 2014, when a very large amount was provided by Darkside.

Concerning computing resources, we show in fig. 4 the overall evolution, splitting the contributions into CNS-funded and group-funded. In this case we do not separate contributions from each single group, since access to computing resources, compared to storage, is much more dynamic and not strictly related to financial contributions. In this case also, we see a steady increment of resources across time, except in 2014 when groups started to contribute substantially to U-LITE.

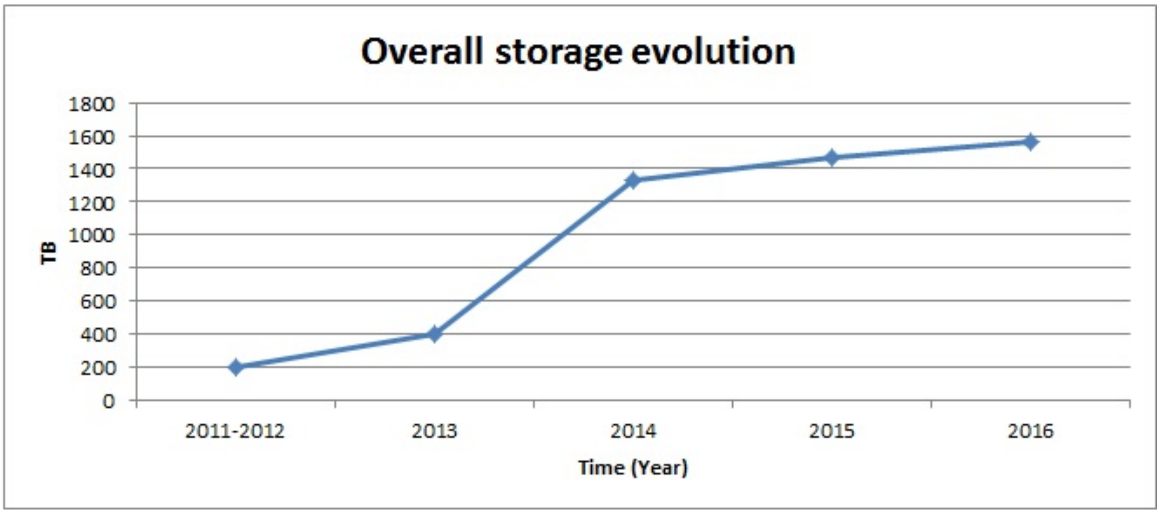

Figure 3: Overall evolution of storage in U-LITE

5 SOFTWARE ARCHITECTURE

As already mentioned, the U-LITE computing cluster is made of a number of heterogeneous, off-the-shelf multi-core computers whose function is to provide hardware resources for the VMs they house. VMs are in charge of executing user programs, typically data analysis jobs and Monte Carlo simulations. All VMs are clones of a limited number of templates, each template being set up according to the requirements of a project or workgroup in terms of, e.g., operating system, installed software, registered users, access to remote storage areas.

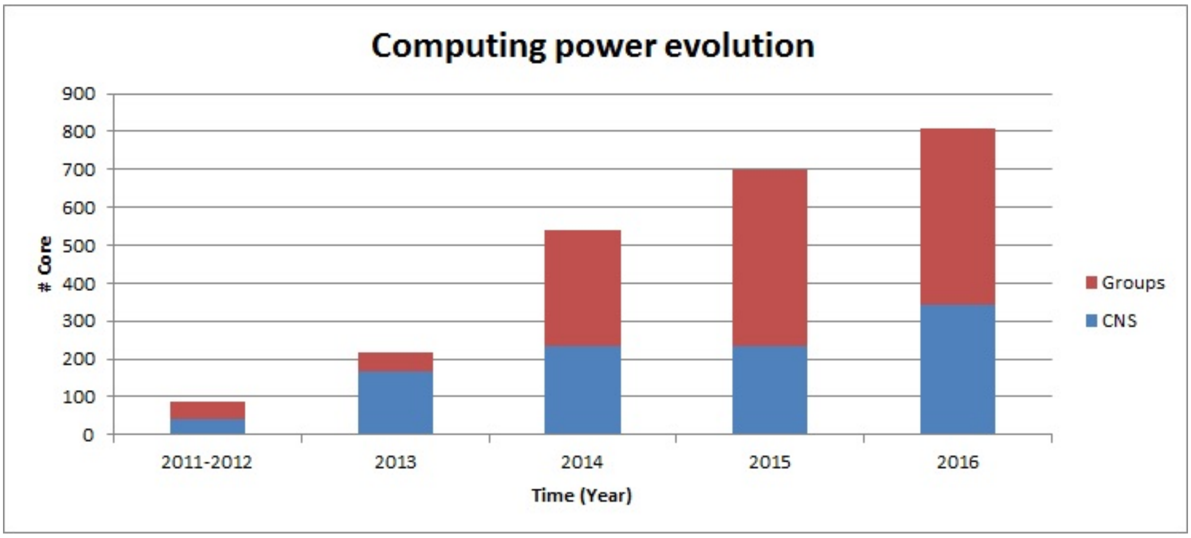

Figure 4: Evolution in time of computing power in U-LITE.

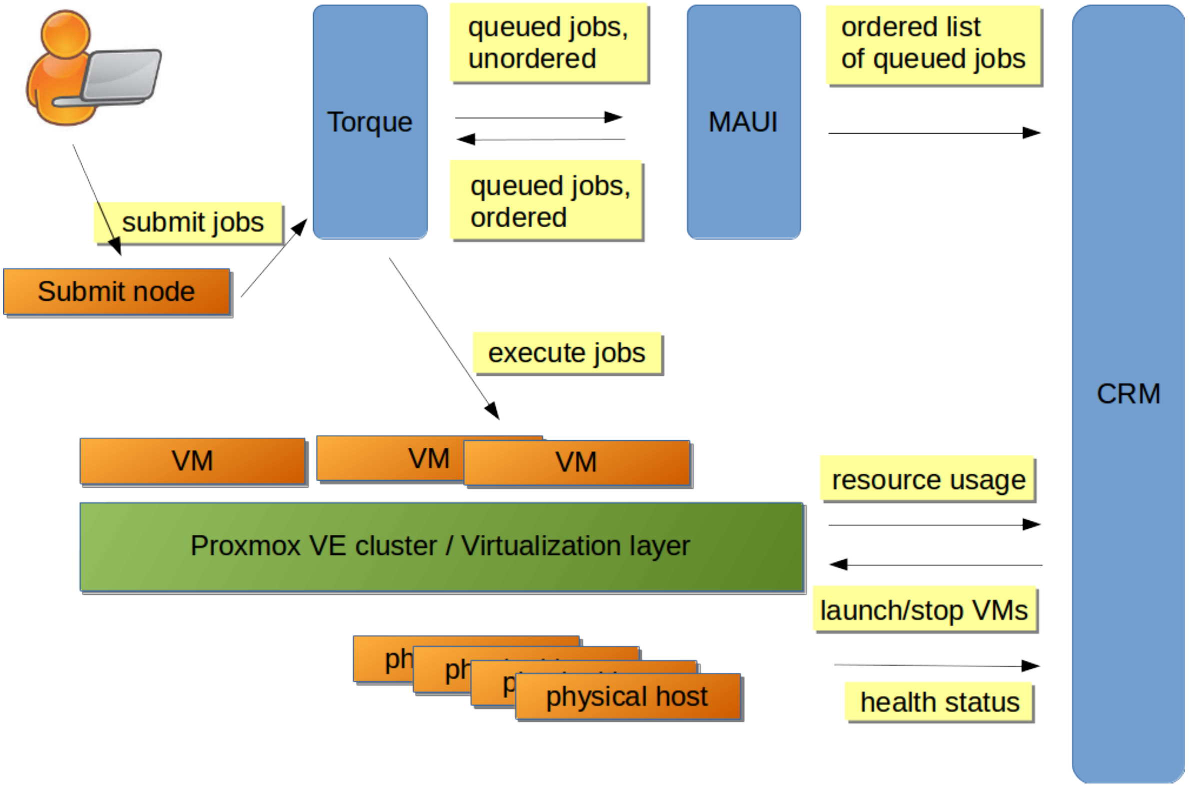

Figure 5: Functional diagram of U-LITE software architecture

5.1 The Batch System

Resource requests, in terms of VMs, are triggered by users submitting jobs to a traditional batch queue system based on the Torque resource manager for job delivery and on the Maui job scheduler for a fair and balanced sharing of resources. Users do not necessarily need to know their jobs will be run on virtual hosts, they may as well think to be working on a traditional system where all computing nodes are real computers.
Torque and Maui[5)] are consolidated batch system software tools, developed before the rise of host virtualization technologies. In order to attain an efficient use of Torque/Maui in a virtualized environment, a specific software tool, CRM (Computing Resource Manager), was developed at LNGS.

5.2 The Computing Resource Manager

CRM represents a middleware layer that collects information from the physical servers, from the Proxmox VE cluster management system, from the Torque server and from the Maui scheduler and operates with the goal of ensuring the resources requested by users via the batch system are available. Furthermore, CRM is in charge of releasing unused resources.

At every iteration (typically once or twice a minute) CRM obtains the list of queued jobs, ordered by priority, from the job scheduler; the status of the running jobs, of the queues and of the computing VMs from the Torque resource manager; that of the physical nodes from the nodes themselves and the resource usage for each physical node from the Proxmox VE cluster.

Based on this information CRM takes actions in order to provide resources for queued jobs, by making VMs available to the batch system, and releasing idle resources, by powering off unused VMs. More in detail: in order to improve the responsiveness to user requests, if the system has enough available resources for new VMs, idle VMs are not powered off but only made unavailable to the batch system by toggling their "offline" flag on. This way, making a VM available again is often just a matter of setting a flag rather than having to wait for the whole boot process. Making resources available or unavailable to the batch system, and having them in full control of the CRM middleware, is necessary for granting each project or experiment their share and avoid unbalanced usage of the computing cluster.

The CRM middleware layer also supports multi-processor jobs and jobs requiring considerable amounts of RAM, both within the limits of U-LITE physical hosts: it reads the job resource requests from the Torque queue status and provides VMs sized as to satisfy the job requirements.

A very important task CRM is in charge of is making sure that the resources shared by each physical server are not overbooked. This means that the total amount of RAM, CPU cores and disk space allocated by the VMs on a server must never exceed the resources that the server itself can provide. In order to reach this goal CRM always starts VMs on physical nodes that have enough idle resources to host them and, should for any reason a server become "overbooked", CRM will stop idle VMs on that server.

5.3 U-LITE queues

While in many HPC/HTC computing clusters the main reason for using different queues is to prioritize short jobs or jobs that require limited resources with respect to more resource intensive ones, in U-LITE different queues are also used to allow users select the VM type, and thus the software platform, to employ for a specific job. Each batch queue is associated to a single VM type, a group of VMs that are clones of the same template. Each VM type can be associated to one or more queues. Usually each VM type and its associated queues "belong" to a single project or working group, while a single project or working group may "own" one or more queues and the associated VM types. The use of VMs that can be automatically powered on or off gives each user or collaboration the chance of using, in principle, all the cluster capabilities when needed and allows for quick, automatic reallocation of idle resources.

6 U-LITE MANAGEMENT

The U-LITE computing infrastructure heavily relies on automated tools for monitoring the hardware and middleware status as well as the computing resource usage, for resource usage accounting and for the management and configuration of its physical host as well as its VMs.

In order to on-line monitor the status of the U-LITE computing cluster a dedicated web page shows in real time important information both for U-LITE administrators and users, such as the status of the U-LITE software components (Torque, Maui and CRM), the overall use of the system, the list of users who have submitted jobs with the number of queued and running jobs, the U-LITE physical hosts list, the list of running worker nodes, the list of queues with running or queued jobs. This page also points to an application showing accounting records on the usage of the compute resources.

At a lower level the health of hardware components and of basic network and computing services is monitored using a centralized Nagios[8] instance that is in charge of monitoring the status of the LNGS computing and network infrastructure as a whole and can send alarms to administrators via e-mail or SMS.

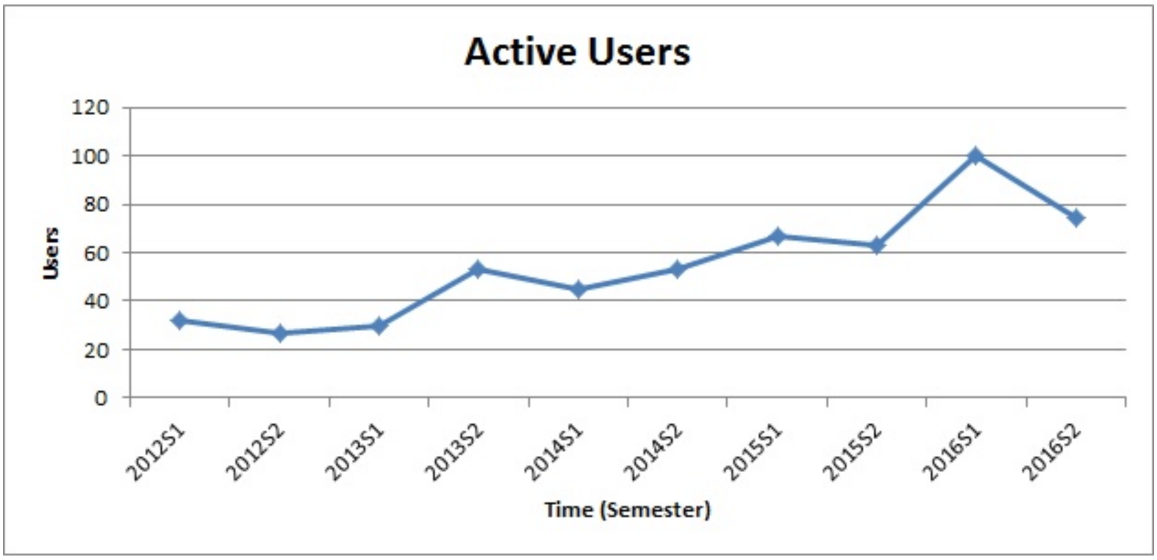

Figure 6: Evolution in time of the number of active users in U-LITE

Since year 2016 both U-LITE physical hosts and VMs are managed through Puppet[9] , a very popular open-source configuration management tool that is designed to manage Unix-like and Microsoft Windows systems declaratively. Puppet allows for an easy administration, maintenance and upgrade of the U-LITE servers and offers a very efficient way of satisfying the experiments requests in terms of software upgrades and installations, network storage configuration and other customizations to be performed on their VMs.

7 U-LITE UTILIZATION

The first U-LITE prototype started to be developed at the beginning of 2011. A pre-production platform was ready in September 2011, and was introduced to the LNGS scientific community with an official proposal[10]. Since the beginning of 2012 the LNGS user community started using U-LITE, and in this section we provide an analysis of the evolution of the platform utilization, starting from the first semester 2012.

7.1 Overall Usage Evolution

In order to have a measure of U-LITE utilization, we analyze here the accounting data from U-LITE scheduling system. Data reported in the following figures are shown with a six month granularity. Specifically, we have considered the amount of submitted jobs, number of active users, CPU time and wallclock time.

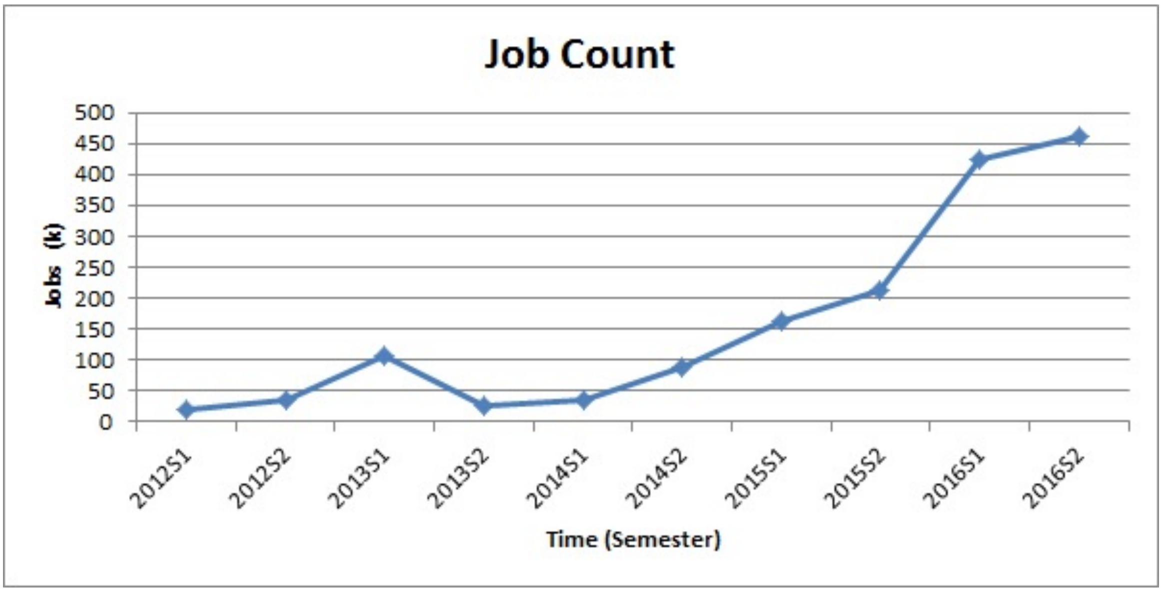

Figure 7: Time evolution of the number of jobs submitted to U-LITE

As a first metric for evaluating the utilization of U-LITE, we looked at the number of active users, i.e., those users who have requested at least a job to the system within the given time bin. This number has gradually grown over time until reaching a peak in the first semester 2016 (see fig. 6).

Another metric we used was the amount of submitted jobs. We can see in fig. 7 how it has grown over time. No decrease appears in the last half of 2016, contrarily to the behavior shown in the active user number (fig. 6), suggesting however a tendency to saturation.

This discrepancy in the behavior of the two metrics in 2016 is a hint of a change in either the composition of U-LITE users, or the way users utilize the platform.

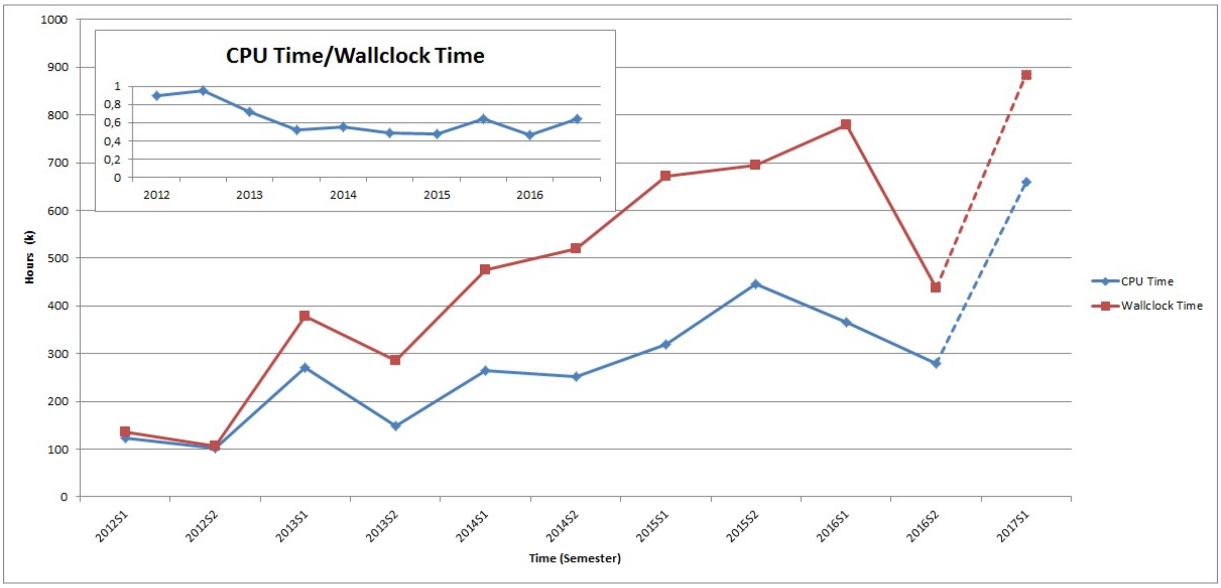

Figure 8: Comparison of CPU time and wallclock time for jobs submitted to U-LITE. The inset shows the ratio of the two values.

In order to better understand this aspect, we analysed the amount of cpu hours and wallclock hours consumed by submitted jobs (see fig. 8). From 2012, values have continually grown over time until a peak, respectively in the second semester of 2015 for CPU time and in the first semester of 2016 for wallclock time. Comparing absolute values of CPU hours and wallclock hours, we can see that initially the two values have been quite similar, and from 2013 a progressive gap between the two patterns has broadened until a peak in the first semester of 2016. This is an indication that the type of jobs submitted by users has changed over time: jobs have been less and less CPU-intensive. Then, during the last period of 2016, the utilization of CPU hours and wallclock hours has significantly decreased, while their trends seem to be converging. In fact, preliminary data from the first half of 2017 (shown as dashed lines in fig. 8) suggest that this decrease is not a global tendency, but part of a fluctuating sequence. The inset of fig. 8 shows in more detail the ratio between the two patterns, evidencing that the apparently high fluctuations between CPU time and wallclock time are in fact very limited in percentage, and close to 50%.

The decrease in wallclock time in the second half of 2016, which is to correlate with a corresponding decrease in active users (see fig. 6), can be explained with a tendency of groups that in the previous years had taken a large share of U-LITE computing resources to move towards external resources. Nevertheless, as already mentioned, early data from 2017 suggest a change in trend, as a large number of compute-intensive jobs have been submitted to U-LITE. The dotted lines in fig.3 show the extrapolation of the amount of CPU hours and wallclock hours in the first half of 2017 considering the data available at this document submission date, March 24th 2017.

7.2 Evolution of U-LITE User Composition

In order to understand the evolution of the U-LITE user composition, we carried out an analysis of the CPU hours consumed by each group over time, and how groups have evolved in user number. The research activities of experimental groups active at LNGS are described in the Labs website[11].

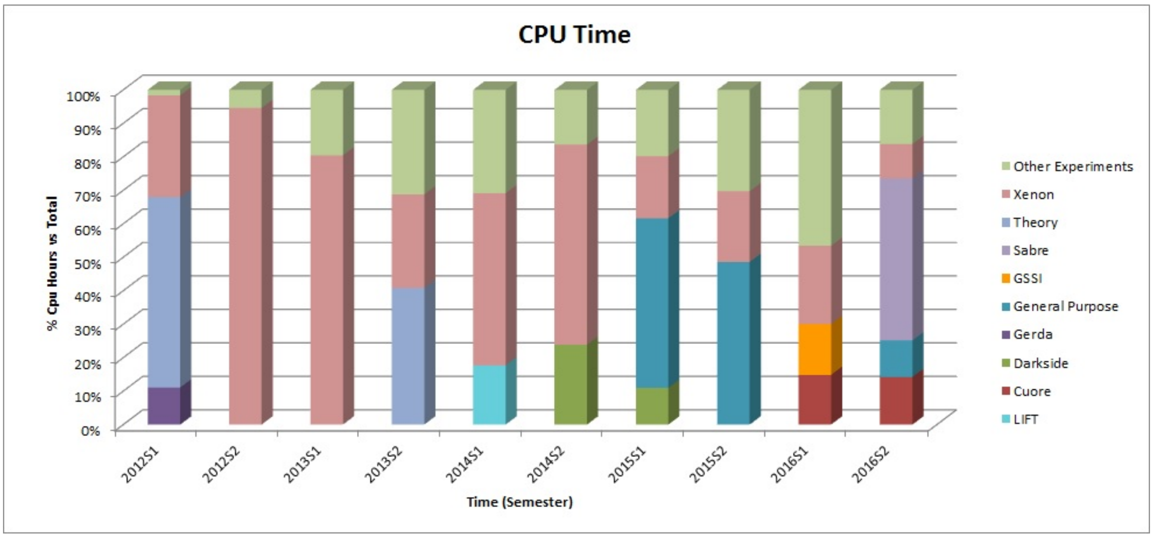

Figure 9: Utilization of U-LITE by the various groups which have been using the system over time

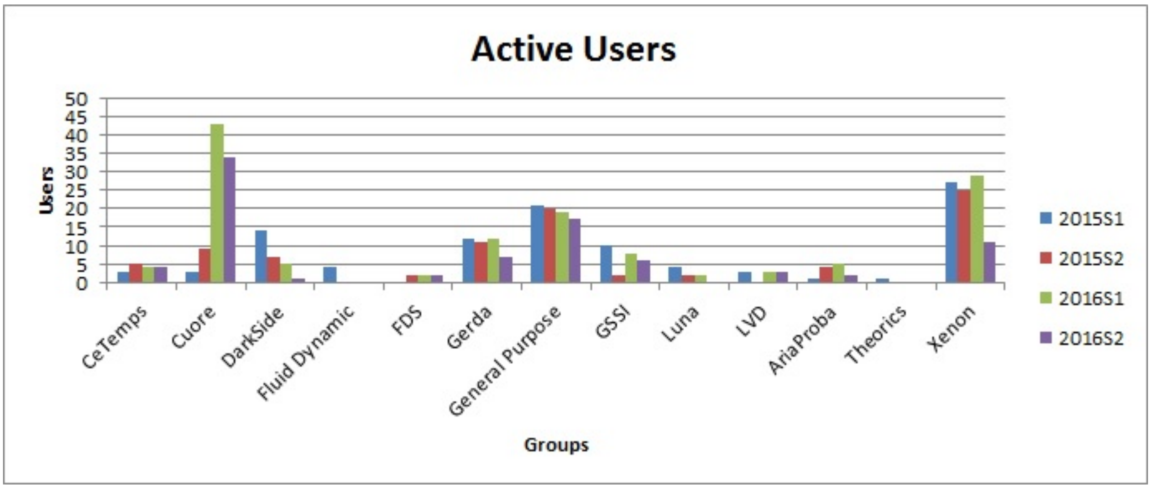

Figure 10: Time evolution of active users per group

In the plot shown in Figure 9, groups in each bin are singled out if their cpu hours utilization has been at least 10% of the total in the corresponding time window. All groups with CPU hours below such threshold are included in Other Experiments. For instance, consider the activity of the Theory group. In the early stages of U-LITE activity, they heavily used the system, accounting for more that 50% of CPU utilization in the first half of 2012, and about 40% in the second half of 2013. Conversely, they do not appear in the other time bins since their utilization of U-LITE was below 10 % in all such cases, and their contribution is included in the Other Experiments shares.

From fig. 9, it is clear how system utilization has been very variable over time. In the early years (2012-2013) approximately 80% of CPU time was used by Xenon, one of the most active experiments on U-LITE. Xenon jobs are CPU intensive, and correspondingly the CPU-time/wallclock-time ratio shown in the inset of fig. 8 is close to 1 in this time frame. Subsequently, other experiments have started using more and more the system. From 2015 a number of experiments which do not have dedicated queues, indicated in the figure as General Purpose, begun to use significantly the system.

This evolution in the user composition explains the pattern shown in fig. 8. As mentioned above, the initial CPU-time/wallclock-time ratio is close to 1 due to the high amount of Xenon jobs, which are CPU-intensive. Jobs submitted by the groups which subsequently joined U-LITE are more interactive, which results in a decrease in the ratio. In particular, interactive jobs are typically submitted by Cuore users. In fact, the peak of active users in the first half of 2016 (see fig. 6), and the concomitant minimum in the CPU-time/wallclock-time ratio in fig. 3 can be correlated with the evolution in the number of Cuore active users. Fig. 10 shows the evolution of active users for the various groups in the time frame 2015-2016. Cuore users are the most in number, and indeed their maximum is in the first half of 2016.

In general, groups show a tendency to use U-LITE heavily for limited periods of time, and decrease their need for CPU time otherwise. The activity of Xenon is an epitome of such pattern, with very large shares of CPU time taken in 2012S2, 2013S1 and both halves of 2014, and relatively limited shares in the other time bins. Other experiments show the same tendency but, as their shares during periods of reduced activity is below 10%, their contribution is not visible because is included in Other Experiments. A new experiment, Sabre, started using U-LITE in the last half of 2016, already contributing for about 50% of CPU time. This once more demonstrate the extreme flexibility of U-LITE in providing computational resources to research groups, and the great advantage in terms of usage optimization resulting from sharing resources among groups.

U-LITE flexibility can be further appreciated quantitatively by looking at the percentage of CPU hours consumed by external projects or groups, i.e., groups whose scientific activity is not related to research performed at LNGS. We report this values in fig. 11. Computational needs of such groups are all CPU-oriented. Their scientific activities are summarized in the table below:

CETEMPS	Excellence Center at University of l'Aquila, using the U-LITE resources for research purposes in the field of data assimilation for the initialization of Weather Research and Forecasting (WRF[12]) model. Several case studies have been simulated using both the three-dimensional and four-dimensional variational data assimilation methods (3D-Var and 4D-Var) of the WRF model. This allowed the Cetemps group to prepare the WRF simulations for the HyMeX and MesoVICT important international projects in which Cetemps is involved.
Ariaproba	Atmospheric science group at the Department of Physical and Chemical Sciences and CETEMPS at University of L'Aquila. Winner of "LNGS Computing Award 2015", using the resources to validate and improve current capabilities of operational air quality forecasts over Italy using a meteorological model (WRF[12]), and a chemistry-transport model (CHIMERE[13]), to simulate the atmospheric chemical composition of main pollutants relevant for human health (ozone, nitrogen dioxide and particulate matter) over Europe and Italy. Ariaproba simulated the years 2008-2012 and compared the results against available ground-based measurements from the European air-quality monitoring network. They further used this relatively long-term dataset to statistically characterize the modelling system bias and developed strategies to improve model performance in order to fit in the uncertainty limits established by the law.
GSSI	Computer Science at Gran Sasso Science Institute (GSSI), using U-LITE to test the effectiveness of innovative Genetic Algorithm (GA) techniques in tackling complex real-life modeling problems, symbolic regression, from many different applicative domains, from Yacht Hydrodynamics to Parkinson diagnosis using voice features to many others. GA is part of the Evolutionary Computation techniques and draws inspiration from the process of natural evolution. The intrinsic complex and stochastic nature of these algorithms needs a relevant number of experiment repetitions to statistically validate theoretical insight.
LIFT	The LIFT laboratory of Engineering Department "Enzo Ferrari" at University of Modena e Reggio Emilia, using U-LITE computational resources in order to perform the verification and validation of a novel, parallel numerical procedure for the numerical simulation of turbulent heat transfer in forced and natural convection regimes.

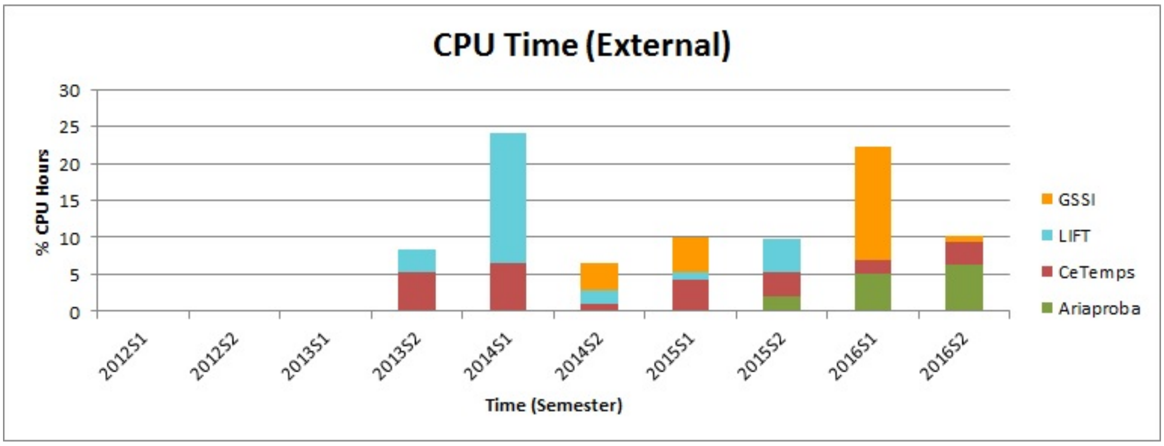

Figure 11: CPU time used on U-LITE by external users

As clearly visible in fig. 11, since the last half of 2013 at least 10% of the total CPU time has been used by external users, with peaks close to 25%. This demonstrate how U-LITE, although developed for the LNGS user community, is able to serve a larger spectrum of users thanks to the versatility arising from the adoption of a cloud paradigm.

7.2 U-LITE multicore computing

The various possibilities provided to U-LITE users also include multicore computing. Figures 12 and 13 show job count and CPU time disaggregated by number of cores. Multicore jobs appear since the early phases of U-LITE utilization (5% of job count and 30% of CPU use in the first half of 2012) and their incidence fluctuates around 20% of CPU time, with a peak close to 50% in 2014. Typically, multicore jobs do not require a large number of processors; a notable deviation from this tendency appears in the first half of 2014, when 20% of CPU time was taken by 32-cores jobs. This is due to the Fluid Dynamics simulations of the LIFT group, as clearly visible comparing fig. 13 with fig. 9 and 11.

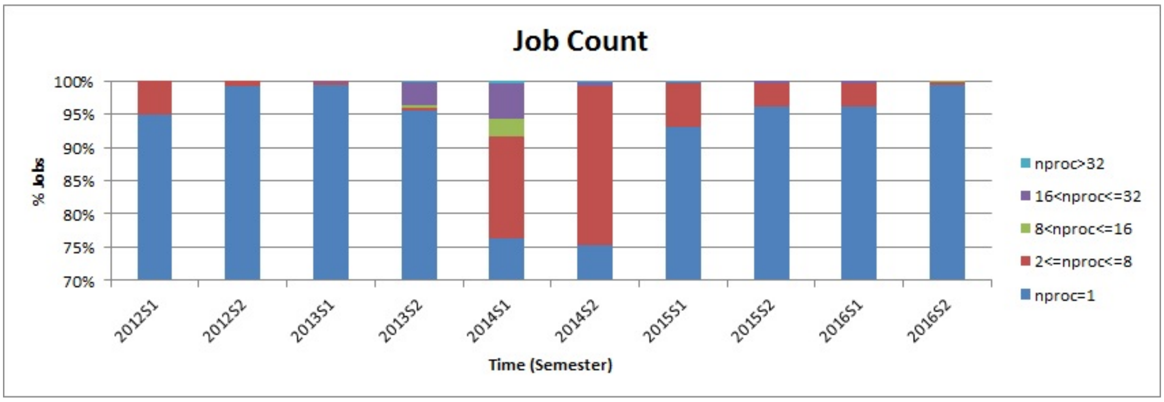

Figure 12: Time evolution of number of jobs submitted to U-LITE, disaggregated by requested number of cores. Note that Y scale starts at 70%.

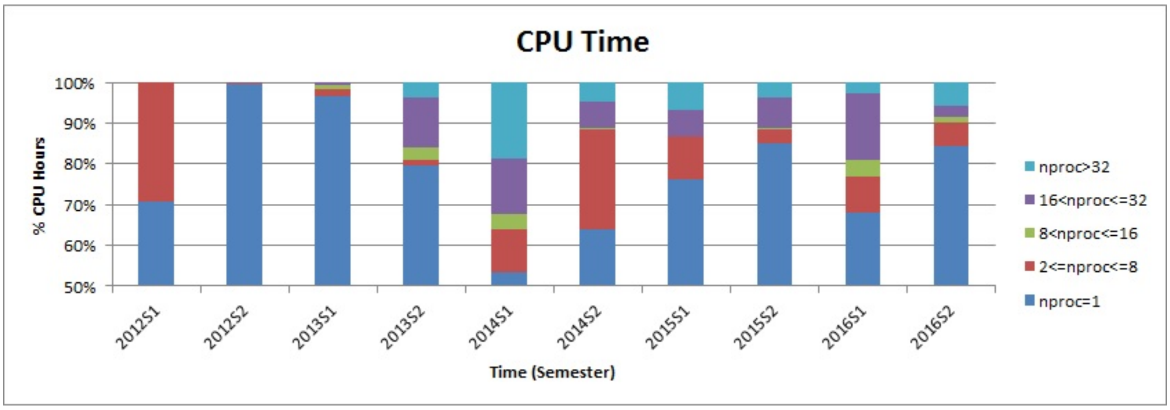

Figure 13: Time evolution of CPU time for jobs submitted to U-LITE, disaggregated by requested number of cores. Note that Y scale starts at 50%.

7.3 Exit Status

A useful metric to evaluate system reliability is the job exit status. Once a job under Torque has completed, the exit status attribute contains the result code returned by the job script. This value is useful in diagnosing problems with jobs that may have unexpectedly terminated. Possible values are:

1. 0 for successful completion.
2. Negative values in case Torque was unable to start the job.
3. Positive values for user-related errors.

In fig. 14 we show the exit code distribution divided on a six-month base. Non-zero codes are always below 20%, and the overall average for successful completion is above 90%. System-related error codes (i.e., negative values) are negligible.

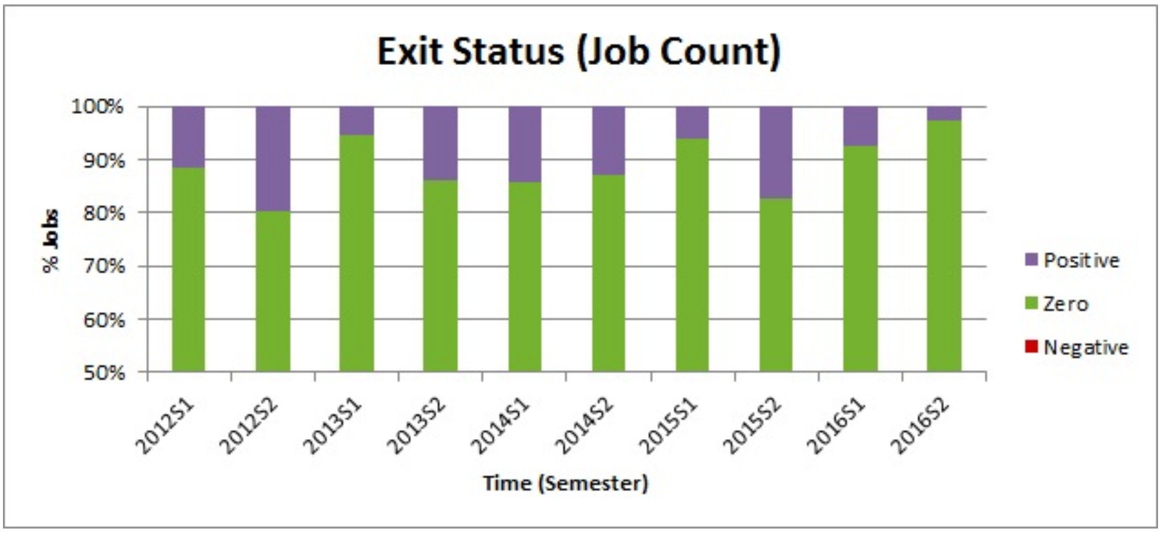

Figure 14: Evolution in time of exit codes. Note that Y scale starts at 50%.

Fig. 15 shows aggregate data over the six years of U-LITE operation. In this way, the system-related contribute appears, as a mere 0.26%. Successful completion, as noted before, is above 90%.

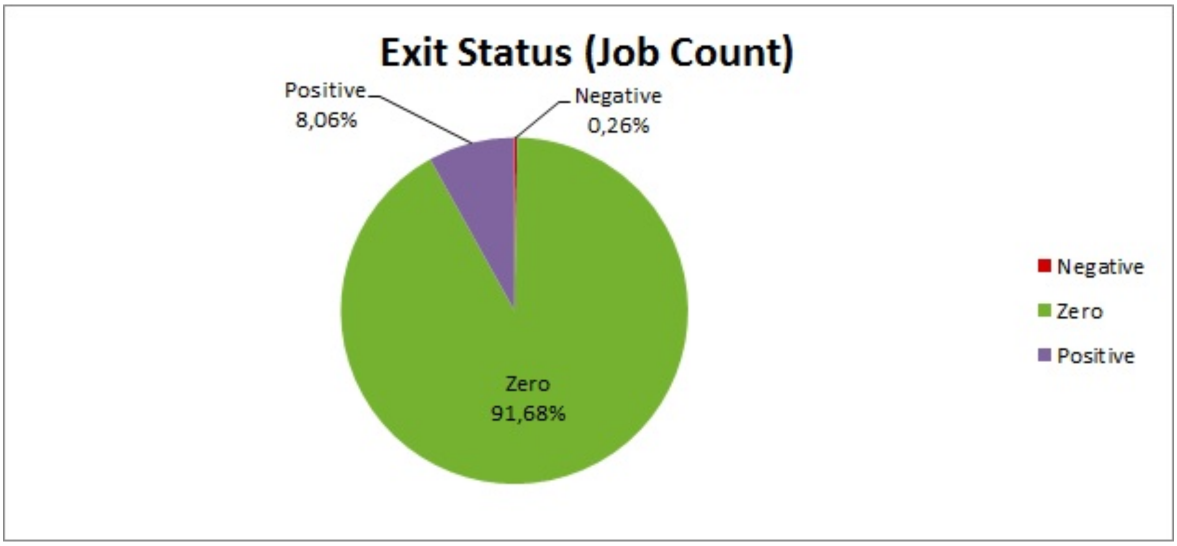

Figure 15: Aggregated data for exit status over 6 years of U-LITE use.

These data show how well U-LITE has performed in terms of system reliability, with only 0.26% of system-related failures. The user-related unsuccessful completions, apart from actual errors (such as missing input data, wrong file permissions, code bugs), also include test runs, application tuning, and other kinds of trials, i.e. jobs that are often purposely terminated before their normal completion. This further reduces the already limited incidence of the user-related terminations.

8 CONCLUSIONS AND FUTURE WORK

This article presents the main features and the usage data of U-LITE, the computing platform of LNGS (Laboratori Nazionali del Gran Sasso). Activities carried out at LNGS include fundamental research on various topics related to Particle and Astroparticle Physics. Experimental groups active at LNGS base their activity on the analysis of data produced within the experimental caves of the underground labs. U-LITE is used to store and analyze such data, and for modeling and simulation of data acquisition apparatuses and physical processes occurring therein. A Theory group is also active at LNGS, with specific needs in terms of computing that U-LITE has been able to fulfill. Moreover, a number of external groups have been granted access to U-LITE within specific research agreements to carry out their computations. Across its 6 years of operation so far, U-LITE has been able to provide computational resources for such large variety of users, demonstrating a high degree of flexibility thanks to its design, which was inspired to a cloud-like model since its inception. U-LITE hardware substrate has evolved over the years according to uses needs. Its current dimensions are those of a small sized data center (about 1.6 PB of storage and 800 cores distributed over 20 hosts), with a number of active users which has steadily grown to about 100 until the first half of 2016. A similar pattern is shown by the wallclock time data, which also peaks at about 800,000 hours in the first half of 2016. The subsequent fall in active users and wallclock time is likely to be correlated to the move of a group among the heaviest users of U-LITE in terms of CPU time towards external resources. U-LITE still remains a resource of choice for a large variety of groups active at LNGS and for external groups, which are attracted by U-LITE versatility and ease of use. In fact, given that during the early months of 2017 a large number of compute-intensive jobs have been submitted to U-LITE, the decrease in usage shown in 2016 might well be a normal fluctuation.

Our overall evaluation on U-LITE, after 6 years at LNGS community's service is by all means positive. LNGS research staff and external groups have profited substantially from using U-LITE. Concerning future development, two quickly emerging technologies are being investigated in order integrate them into the U-LITE workflow.

The container technology, an approach to virtualization in which the virtualization layer is extremely light and runs as an application within the operating system, could, at least in part, substitute full virtualization in U-LITE, allowing for optimal exploitation of hardware resources. The use of containers instead of fully virtualized hosts might also speed up the system response to newly submitted jobs, as a container is created and started up much faster than a VM, and simplify the maintenance and upgrade of the software platform of each project. Much work in this direction has already been performed, and container based environments are already fully integrated and available in U-LITE, although they have not been yet used by scientific collaborations.

The second direction for architecture upgrades in U-LITE is in the way scientific data is accessed. U-LITE compute nodes use today almost exclusively the NFS[14)] file system for shared access to scientific data. This quite traditional approach, although efficient and easy to implement and manage, requires a tight coupling between the storage systems and the compute resources and is not resilient to temporary failures in the data servers. Modern data access paradigms, like Object Storage system or XRootD[15)], may not be as efficient under certain circumstances, but are more resilient to system failures and do not need such tight coupling. This means that local access to remote data as well as remote access to local data would become possible and transparent to the final user. Breaking the barrier of the single data center for LNGS experimental data analysis and storage is important for the LNGS scientific community that is spread all over the world.

In the near future these new paradigms to data access will be proposed to scientific collaborations, mainly to new experiments that do not have an already defined data workflow.

9 ACKNOWLEDGEMENTS

We kindly acknowledge Giuseppe di Carlo (INFN-LNGS) for his insightful comments during our discussions in the initial stages of development of U-LITE.

10 REFERENCES

(1) http://www.lngs.infn.it
(2) http://www.lngs.infn.it/en/dark-matter
(3) http://www.lngs.infn.it/en/neutrino-physics
(4) http://www.lngs.infn.it/en/nuclear-astrophysics
(5) http://www.adaptivecomputing.com/products/open-source/torque/
(6) http://www.proxmox.com
(7) http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf
(8) http://www.nagios.org/
(9) http://puppet.com/
(10) https://www.lngs.infn.it/images/REIS/Annual_Report/PREPRINT/preprint113.pdf
(11) https://www.lngs.infn.it/en/current
(12) http://www.wrf-model.org/
(13) http://www.lmd.polytechnique.fr/chimere/
(14) https://en.wikipedia.org/wiki/Network_File_System
(15) http://xrootd.org