ANNUAL REPORT 2013

INFN CNAF

INFN-CNAF Annual Report 2013

www.cnaf.infn.it/annual-report

Editors

Luca Dell'Agnello
Francesco Giacomini
Claudio Grandi

Address

INFN CNAF
Viale Berti Pichat, 6/2
I-40127 Bologna
Tel. +39 051 209 5763, Fax +39 051 209 5478
www.cnaf.infn.it

Cover Images

The tape library hosted at the CNAF Tier-1 Computing Center.
Grid activity as shown by the WLCG dashboard (*dashboard.cern.ch*). ©2009 GeoBasis-DE/BKG, Image Landsat ©2014 Google, US Dept. of State Geographer.
View of the entrance to the Physics and Astronomy Dept. of the University of Bologna, where CNAF is hosted.

Cover design: Francesca Cuicchio

Contents

Scientific Exploitation of CNAF ICT Resources

The INFN-Tier1 Center and National ICT Services

Software Services and Distributed Systems

Additional Information

50 Years of CNAF

Introduction

This first CNAF Annual Report is dedicated to CNAF itself: to its first 50 years, to those who founded it, to those who contributed to improving it until it became one of the main data centers for experimental physics and particle physics especially, to those who succeeded in attracting copious external funding to our Institute. European funds, regional funds and prize funds have in fact allowed us to inject our center with innovation and development, to remain technologically advanced and to face with adequate tools the increasingly engaging challenges our experiments require. But this first activity report is primarily dedicated to all CNAF personnel, both staff and temporary collaborators, who day by day make it happen.

Since 2005, CNAF has hosted the Italian Tier1 computing center for LHC experiments, but many are the INFN experiments dealing with particle and astro-particle physics that it has attracted. An approximate 75% of the Tier1 computing resources is used by LHC; the rest by other experiments. The first section of this Annual Report, "Scientific Exploitation of CNAF ICT Resources", gathers and recounts the experience and results of all users of our Institute. In 2013 new experiments were added to the list CNAF leads, thus demonstrating CNAF's importance in Scientific Computing, both nationally and internationally.

The second section, "The INFN-Tier1 Center and National ICT Services", describes the "data center", with its architecture and organization. In 2013 the computing power of the Tier1 reached 200000 HS06 (18000 cores), a fast storage capacity of 13 PB and a long-term storage capacity (tape) of 15 PB. The computing center also hosts services of national utility for INFN, such as the top level domain DNS and high availability core DNS, mail relay services, mailing lists services, centralized web sites management, NTP.

"Software Services and Distributed Systems" is the third section of this report. The highlight is on software development activities, geographically distributed systems (Cloud and Grid) and related services (e.g. SRM/StoRM), as well as technological tracking activities aimed at exploring new technologies in computing systems (CPU architecture, storage, memory, etc.). Among these activities: the recent collaboration with the KM3NeT experiment for the development of their trigger and data acquisition; the completion of the COKA project on new multicore processors; important activities of technological transfer of Cloud Computing towards the Public Administration of Regione Marche.

Year 2013 also set the end of the last European projects financed by the 7th Framework Programme (e.g. EGI-Inspire) and the opening of Horizon 2020 calls for proposals. One important activity that was carried out during this year is in fact the study of the H2020 programme, focused on e-Infrastructures and ICT themes, in order to identify the most suitable calls for our work programmes and future development. The project Open City Platform, consisting of 20 partners led by INFN, with an overall budget of 12 million Euros, was one of the winners of the call for proposals promoted by MIUR on "Idee progettuali per Smart Cities and Communities and Social Innovation". CNAF, with an approved budget of half a million Euros, has been playing a leading role in this project both in managing and technical procedures. Starting year of all activities: 2014.

Our core activities — the pillars that guarantee the center's sustainability and future — are scientific computing, software development, management of external projects and technology transfer towards civil or industrial subjects throughout the Country. They are complementary activities, but the challenge we face is to make them grow harmoniously and interdependently, effectively yielding existing competences and impressive synergies in the diverse areas.

In 2013, CNAF celebrated its first 50 years (1962–2012). The most important moments of the center

have been recounted in the volume "CNAF 50 Years of Technological Excellence", which was published on the occasion of the celebrations. We open this Annual Report with the introduction to Mauro Morandin's book. Director of CNAF until September 2013, he saw in digital revolution the common thread uniting the first glorious ante-litteram "scanners" built in our center's very first quarters, and the refined powerful computing center Tier1 that CNAF hosts nowadays and with which INFN participates in the analysis of all data resulting from LHC experiments. The most significant articles in the book can be found in the appendix of this Annual Report.

Gaetano Maron
CNAF Director

CNAF 1962–2012
50 years of innovation and technological excellence

When the National Institute of Nuclear Physics (INFN) created the CNAF Center in 1962, nobody could have imagined the crucial impact that digital technologies would play many years later, and their role in all aspects of scientific activities and, to an increasing extent, of everybody's life. At that time computers were really big and intimidating machines that could be talked to only via very specialized and clumsy interfaces. Digital network did not exist, and calculations were mostly performed with the help of logarithmic tables and slide rules.

However, it was at about that time that the very early phase of the digital revolution began. Scientists in the most advanced fields of research felt the need to have at their disposal electronics devices capable of transforming analog information into digital data streams to be immediately digested by computers, and that urgency paved the way for a radical technology shift meant to revolutionize the way we live. The creation of CNAF was an example of a general transition driven by compelling scientific motivations: far before the invention of digital photography, physicists had the urgency to digitize huge amounts of pictures of elementary particle trajectories that started to be produced at a high rate.

In hindsight, the decision taken by INFN to create CNAF was remarkable for a number of reasons. First of all, the commitment to build a Center specifically focused on technological activities showed how important it was to involve INFN personnel in the conception, development and operation of the advanced tools required to preserve competitiveness in cutting edge research. In the following decades, this approach proved to be vital for INFN scientists, who were able to play a key role in the most advanced research activities all over the world.

Secondly, the choice of investing specifically on digital technologies turned out to be quite far-reaching. Although adapting to the rapidly evolving technological context, the original mission of CNAF to be at the forefront of digital technologies has basically remained the same for more than 50 years, and it is still the core of its activities today. Finally, the special location of CNAF (which is firmly embedded in the University context) was part of a general strategic plan that saw INFN pursuing to keep their activities strongly linked to that fertile and inspirational environment. The story of CNAF is the story of how Science can stimulate and exploit novel technological developments for its own benefits, and possibly reshape the way our society works. This book is a recollection of the exceptional story of this project, and a tribute to the ingenuity and commitment of all the people who contributed to its success.

Mauro Morandin
Former CNAF Director

Valerio Venturi (1975–2013)

Our colleague and friend Valerio Venturi suddenly passed away on Christmas day while having lunch with his family.

Valerio was born in Florence on April, 5th 1975. He studied Mathemathics at the University of Florence, and graduated in 2003 discussing a thesis on algebraic topology in symplectic manifolds. After the thesis, he obtained a MsC in applied mathematics at the University of Bologna, focusing on the study of invariant signatures for planar shape recognition.

In May 2004 Valerio joined INFN-CNAF as a member of the VOMS development team. His contribution in this role was of great importance for enstabilishing VOMS as the core of the Grid middleware authorization stack. He first worked on the implementation of the VOMS server, clients and APIs and then focused on VOMS integration in SAML federations.

From 2008 to 2010 Valerio led the INFN team that worked in the ETICS-2 project. Besides representing INFN in the technical board of the project, Valerio was responsible for the project work-package in charge of handling support request from users of the infrastructure, helping new customers integrating their projects with the ETICS system.

After a brief work experience in Lepida, in late 2011 he returned to INFN to lead the development group of the Italiand Grid Infrastructure project. His responsibility was to ensure that the middleware that runs the distributed storage and computing infrastructure supporting the LHC experiments and other scientific communities worked reliably and evolved to meet emerging new requirements.

Valerio was a delightful person, with a great vision and a natural ability to relate with people. He will be dearly missed for his valuable contributions at work, but above all for his friendship and fine humour.

Scientific Exploitation of CNAF ICT Resources

The CNAF User Support Service

Claudio Grandi
INFN Bologna
E-mail: Claudio.Grandi@bo.infn.it

Abstract. The CNAF User Support service provides direct day-by-day operational support to the research groups using the Tier-1.

1. Introduction

The role of the CNAF User Support service is to provide direct day-by-day operational support to the research groups, also known as Virtual Organizations (VO), using the Tier-1. The User Support Service has been configured as an independent service with respect to the Tier-1 Service, in order to guarantee its independence in representing the interests of the users, being they the stakeholders of the Tier-1.

2. History

The User Support Service was born, in its current configuration, in April 2012. Direct support of the LHC experiments exists since when the Tier-1 was created. More personnel has been assigned to the support of other VOs in the following years.

With the consolidation of the activities related to the LHC experiments and the increase of the number of active VOs, it appeared clear that a stronger coordination of the activities of the user support staff could favour the transfer of know-how from the bigger groups to the smaller ones, and a more efficient utilization of the resources.

3. Human resources

The service is provided by one senior permanent staff working at 50% of his time as coordinator and six fixed term staff (*Assegno di Ricerca*) with post-doctoral education or equivalent work experience in scientific research or computing service. Contracts are two-years, renewable for two more years if age requirements are satisfied.

Given the relatively frequent turn-around and the academic nature of the contract, training and educational activities are integral part of the staff activities.

4. Supported experiments

The LHC experiments represent more than three quarters of the total CNAF utilization. They use heavily the Grid technology on top of which they have built experiment specific services to better support data management, job management and monitoring. For all LHC experiments CNAF has the rank of a Tier-1 site. LHCb has also a Tier-2 at CNAF (in the following the LHCb accounting includes also the Tier-2 usage). Furthermore CNAF hosts a Tier-3 of the

INFN Bologna section that supports ATLAS, CMS, LHCb and other local research groups. In the following accounting for the Tier-3 is not included.

The non LHC research groups that have been or are directly supported by CNAF are: CDF, Babar, SuperB, Belle II, NA62, KLOE, LHCf, Argo, AMS2, Auger, Icarus, Fermi/GLAST, Pamela, Borexino, Xenon, Gerda, Virgo, CTA, Opera, Agata, a few theoretical physics and local astronomy groups. Contacts have been taken for the support of more VOs such as Km3NET and PANDA. Most of these VOs also use Grid technologies at some level, but in some cases the activities are done only locally or using ad-hoc tools.

5. Activities

Day-by-day operational support of the VOs includes:

- Installing, operating and maintaining the VO-specific services running at the Tier-1 on dedicated machines (VOBOXes) for LHC experiments; given the variety of ways in which the VOs computing model are built, the service is not currently providing support for VO-specific services of other VOs.
- Responding to VO-specific tickets on the VOs ticketing system, support mailing list or on direct user contact, depending on the VO rules.
- Acting as a filter from the VO and the Tier-1 staff, forwarding to the appropriate Tier-1 Department the requests that cannot be satisfied autonomously, e.g. those concerning common services such as Grid Computing and Storage Elements.
- Attending the VO Computing Operations meetings and representing the CNAF Tier-1 inside the VO.
- Bridging the communication between the VOs and the Tier-1 in both directions; e.g. anticipating major reprocessing campaigns of the VO or communicating best practices to the VO for the use of the Tier-1 resources.

Furthermore the service provides consultancy for the evolution of the VO Computing Model, promoting the use of common technologies to access the Tier-1, in particular Grid tools.

In some cases the personnel of the User Support Service is deeply integrated in the experiments they support; this is the case especially of the LHC experiments because in the past the contact person working at CNAF was coordinated directly within the experiment. This has created competences at the highest level in HEP computing that can now be transferred and used also by other VOs. In this respect a great achievement of the team has been to help some of these groups to modify their computing model in order to include the use of common Grid tools.

The service provides the role of the Tier-1 Run Coordinator. The goal of this task is to have an overview of the VO activities at the site, represent CNAF at the Daily WLCG calls and report about resource usage and problems at the monthly meeting of the Tier-1 management body (*Comitato di Gestione del Tier-1*).

Besides the operational activities, the service participates to a few software development projects intended to improve the usability of the Tier-1. It is worth mentioning the development of tools for the collection of the Tier-1 usage statistics. They automatically retrieve CPU, disk and tape assignments and usage statistics from the dedicated monitoring systems and present them in a consistent way to help the Tier-1 management body (*Comitato di Gestione del Tier-1*) control and build plans.

6. Resource usage

The following figures show the CPU, (Figure 1) disk (figure 2) and tape (figure 3) usage by the supported VOs since January 2012 compared with the assignment and the pledge to the

experiment. In all cases the pledge for the year is normally assumed to be provided in April of that year.

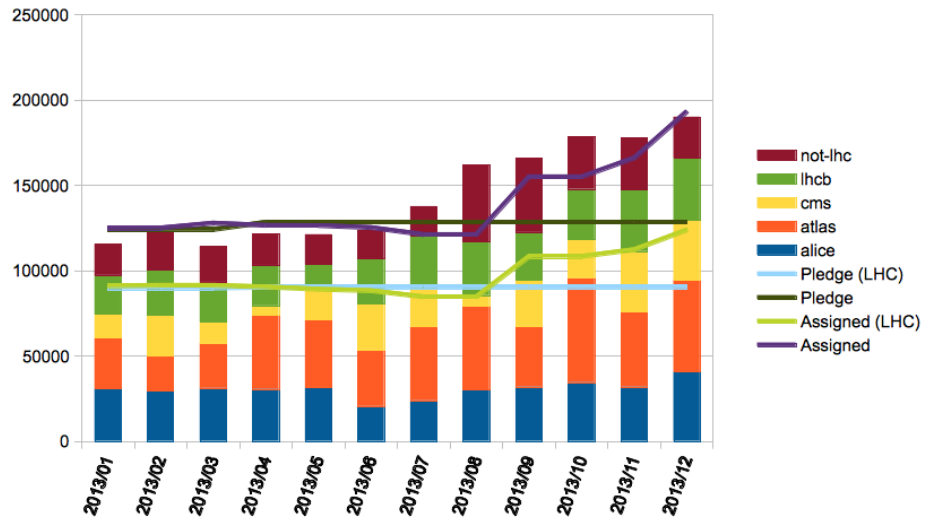


Figure 1. Average daily CPU usage ($HS06 \cdot days$). The non-LHC VOs are grouped together. The lines show the pledges and the assignments for the LHC experiments and in total. Even though individual VOs have significant fluctuations, the total resource usage only shows small fluctuations when resources have been unavailable for technical interventions. The graphs only account for VOs directly supported by CNAF. Part of the resources are assigned to other VOs (e.g. biomed) but they can be used in case they are underutilized.

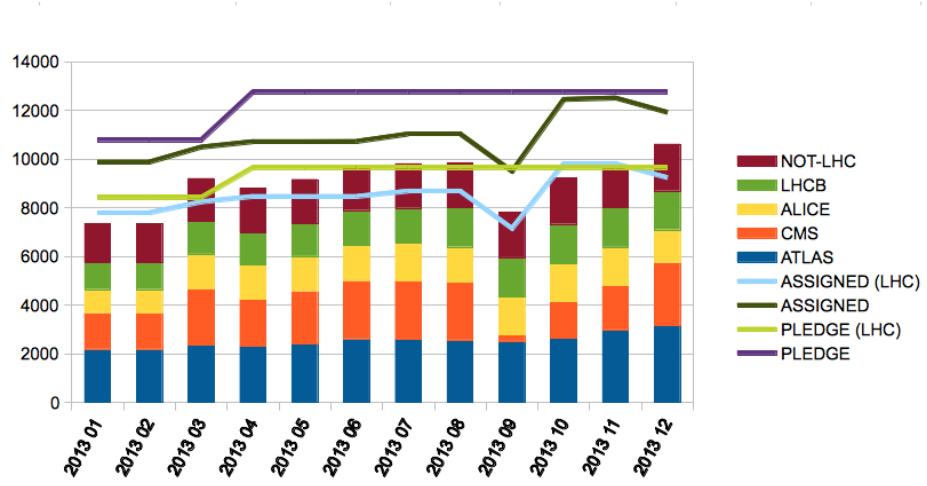


Figure 2. Disk usage (TB) for all VOs. The non LHC VOs are grouped together. The lines show the pledges and the assignement. The dip in september is due to an intervention on the CMS storage.

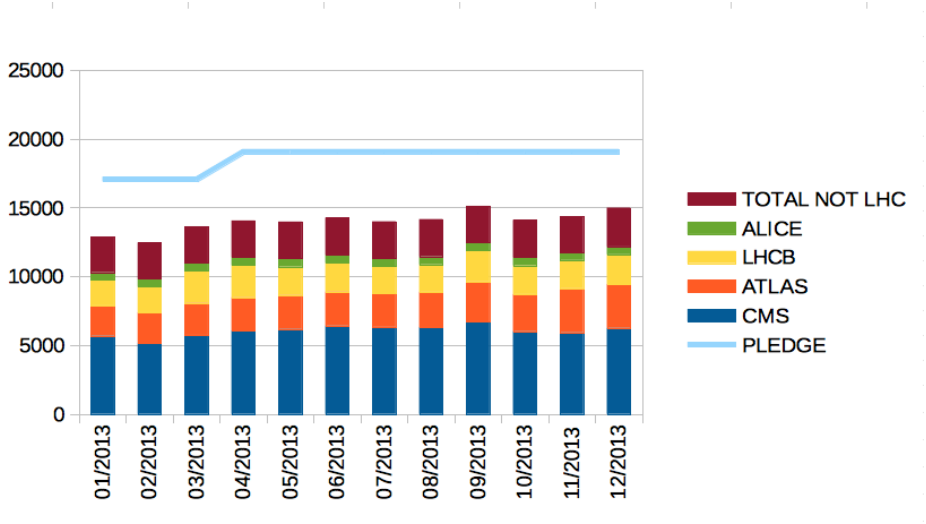


Figure 3. Tape usage (TB) for all VOs. The non LHC VOs are grouped together. The line shows the pledge. Tapes are assigned to the VOs only when needed.

The ALICE experiment activities at CNAF

S. Bagnasco[1], **D. Elia**[2]
[1]INFN Torino, [2]INFN Bari
E-mail: Stefano.Bagnasco@to.infn.it

1. Experimental apparatus and physics goal

ALICE (A Large Ion Collider Experiment) is a general-purpose heavy-ion experiment specifically designed to study the physics of strongly interacting matter and QGP (Quark-Gluon Plasma) in nucleus-nucleus collisions at the CERN LHC (Large Hadron Collider).

The experimental apparatus is composed of a central barrel detector, which measures event-by-event hadrons, electrons and photons, and a forward spectrometer to measure muons. It has been designed to cope with the highest particle multiplicities theoretically anticipated for Pb–Pb reactions and has been operational since the start-up of the LHC in 2009. The central part, embedded in a large solenoidal 0.5 T field magnet, covers polar angles from 45° to 135° over the full azimuth. It consists of a Inner Tracking System (ITS) of high-resolution silicon detectors, a cylindrical Time Projection Chamber (TPC), three particle identification arrays of Time-of-Flight (TOF), Ring Imaging Cherenkov (HMPID) and Transition Radiation (TRD) detectors, a high-resolution electromagnetic calorimeter (PHOS) and a large Pb-scintillator sampling calorimeter designed also for jet studies (EMCAL). The forward muon arm (covering polar angles from 2° to 9°) consists of a complex arrangement of absorbers, a large 0.7 T field dipole magnet, and fourteen planes of tracking and triggering chambers. Several smaller detectors (ZDC, PMD, FMD, T0, V0) for global event characterization and triggering are located at small angles. An array of scintillators (ACORDE) on top of the large solenoidal magnet is used to trigger on cosmic rays. An extension (DCAL) of the EMCAL, a second arm complementing EMCAL at the opposite azimuth and thus enhancing the jet and di-jet physics, is being installed during the Long Shutdown 1 period of LHC. A detailed description of the ALICE sub-detectors can be found in [1].

The main goal of ALICE is the study of the hot and dense matter created in ultra-relativistic nuclear collisions. At high temperature the Quantum CromoDynamics (QCD) predicts a phase transition between hadronic matter, where quarks and gluons are confined inside hadrons, and a deconfined state of matter known as Quark-Gluon Plasma [2, 3]. Such deconfined state was also created in the primordial matter, a few microseconds after the Big Bang. The ALICE experiment creates the QGP in the laboratory through head-on collisions of heavy nuclei at the unprecedented energies of the LHC. The larger the colliding nuclei and the higher the centre-of-mass energy, the greater the chance of creating the QGP: for this reason, ALICE has also chosen lead, which is one of the largest nuclei readily available. In addition to the Pb–Pb collisions, the ALICE Collaboration is currently studying pp and p–Pb systems, which are also used as reference data for the nucleus-nucleus collisions.

2. Main physics results

After the first three years of operations the ALICE has obtained important results studying in detail the hot matter produced in Pb–Pb collisions, with particular emphasis on correlations, heavy flavour and particle production. These results impose more stringent constraints for the various QCD models describing the QGP and the hadronization phase. The physics results of ALICE also include several measurements in proton-proton and proton-nucleus collisions, as well as photoproduction using ultra-peripheral collisions in Pb–Pb.

The experiment has collected data on Pb–Pb collisions at 2.76 TeV per nucleon pair in 2010 and 2011, pp collisions up to 8 TeV from 2009 to 2012 and p–Pb collisions at 5.02 TeV in 2012 and 2013. The measurements with proton beams, which are the baseline for the study of the heavy-ion collisions, have also produced genuine physics results: among them, the very first measurements of the charged-particle multiplicity and rapidity density at the LHC [4, 5] provided the basic ingredients for the tuning of the Monte Carlo generators. The first measurements with Pb–Pb collisions provided results on the charged-particle multiplicity (found to be around 1600, well below most of the extrapolations from RHIC) [6], the size and life time of the system created in the collision (increased by factors of 2 and 1.4 with respect to RHIC, respectively) [7] and the elliptic flow coefficient v_2 (increased by about 30% with respect to RHIC and in agreement with the hydrodynamic description of a strongly interacting very low viscosity fluid) [8].

Along the last years many other results have provided detailed information on the system created in Pb–Pb collisions at the LHC and how such system is described by the various hydrodynamic and thermal models. On the heavy flavour sector, results on the production of J/ψ, Υ and D mesons (nuclear modification factors R_{AA}, flow, particle ratios etc) [9, 10, 11] have cast new light on the interaction of hard probes with the medium, providing the first indications of a mass dependence of the energy loss in the QGP. In the light flavour sector, the spectra and correlations of identified particles have been published both for Pb–Pb and p–Pb collisions [12][13]. In particular, the latest data collected in 2013 on p–Pb collisions have allowed the evaluation of cold nuclear effects in the long rapidity range correlations (ridge effect) [14] and are currently being compared with different calculations including shadowing, saturation and hydrodynamical expansion, in order to elucidate the origin of such effects. Finally, results on J/ψ production in ultra-peripheral Pb–Pb collisions have been obtained for the first time at the LHC [15], allowing comparison with data from HERA and providing discrimination with respect to QCD models and new input on nuclear particle distribution functions.

3. Computing model

The ALICE computing model is heavily based on Grid distributed computing; since the very beginning, the base principle underlying it has been that every physicist should have equal access to the data and computing resources [16]. According to this principle, the ALICE peculiarity has always been to operate its Grid as a cloud of computing resources (both CPU and storage) with no specific role assigned to any given centre, the only difference between them being the Tier to which they belong. All resources are to be made available to all ALICE members, according only to experiment policy and not on resource physical location, and data is distributed according to network topology and availability of resources and not in pre-defined datasets.

Thus, Tier-1s only peculiarities are their size and the availability of tape custodial storage, which holds a collective second copy of raw data and allows the collaboration to run reconstruction passes there. In the ALICE model, though, tape recall is almost never done: all useful data reside on disk, and the custodial tape copy is used only for safekeeping. All data access is done through the xrootd protocol, either through the use of native xrootd storage or, like in many large deployments, using xrootd servers in front of a distributed parallel filesystem like GPFS.

In the original MONARC architecture, the Grid had a hierarchical layout, with large Tier-1

centres acting as a pivot for a number of smaller Tier-2 centres in the same region; with the partially unforeseen availability of reliable high bandwith network connections between centres, at a relatively cheap price, the hierarchy is becoming more and more blurred. At the same time, alongside the Grid centres providing batch-like computing power, a number of Interactive Analysis Facilities have been born based on PROOF, that complement the Grid centres providing rapid turn-around resources for analyses that require a smaller dataset or for tuning on smaller samples analyses that will then be run on full statistics on the Grid [17].

The Italian share to the ALICE distributed computing effort (currently about 20%) includes resources both form the Tier-1 at CNAF and from the Tier-2s in Torino, Bari, Catania and Padova/Legnaro, plus some extra resources in Cagliari, Bologna and Trieste.

Since March 2013, the activity to define and develop the new computing framework for the post-Long Shutdown 2 phase has been officially started within the Collaboration. The corresponding project (O^2 Project), which is mainly based on the concepts of Online-Offline integration and Cloud computing, has been organized in several dedicated Computing Working Groups which are expected to provide a Technical Design Report by October 2014.

4. Role and contribution of the INFN Tier-1 at CNAF

CNAF is a full-fledged ALICE Tier-1 centre, having been one of the first to enter the production infrastructure years ago. According to the ALICE computing model, it has no special given task among Tier-1 centres, but provides computing and storage resources to all the collaboration, along with offering a valuable support staff for the experiments computing activities.

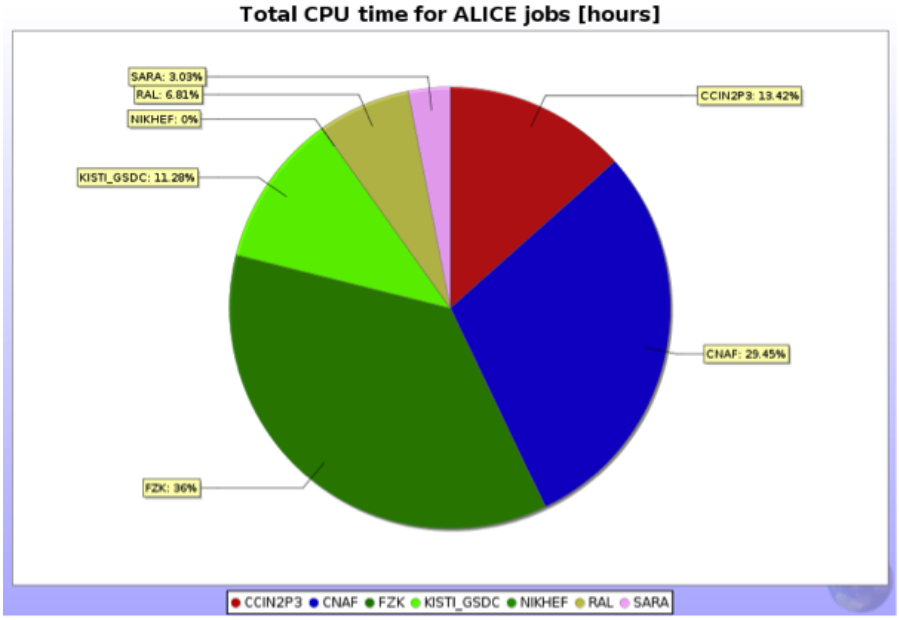


Figure 1. Ranking of CNAF among ALICE Tier-1 centres in 2013

It provides reliable xrootd access both to its disk storage and to the tape infrastructure, though a TSM plugin that was developed by CNAF staff specifically for ALICE use.

The computing resources provided for ALICE at the CNAF Tier-1 centre were fully used during last year, matching and exceeding the pledged amounts by as much as 150% when exploited resources unused by other collaborations (Figure 1). Overall, about 60% of the activity was Montecarlo simulation, 10% reconstruction (which takes place at the Tier-0 and Tier-1

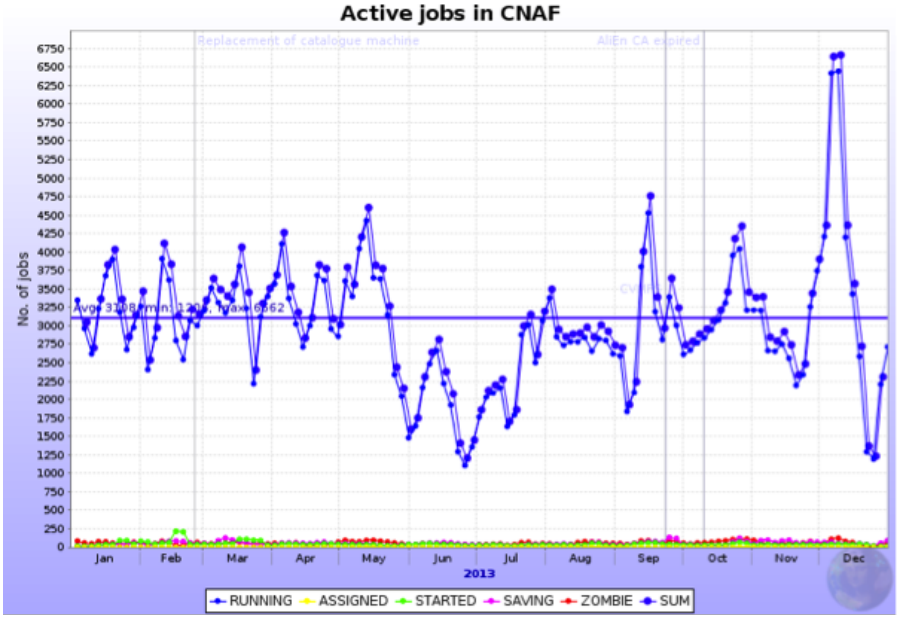


Figure 2. Running jobs profile at CNAF in 2013

centres only), 10% organized analysis (the so-called analysis trains) and 20% chaotic end-user analysis.

In 2013, CNAF provided 7.9% of the total CPU hours used by ALICE, thus ranking third of the ALICE sites (following Prague and FZK in Karlsruhe and excluding the Tier-0 at CERN), corresponding to about 60% of the total INFN contribution: it successfully completed more that 6.6 million jobs, for a total of 21.9 millions CPU hours (Figure 2).

ALICE keeps on disk at CNAF about 1.3 PB of data, plus about 700 TB of raw data on custodial tape storage; the reliability of the storage infrastructure is commendable, even taking into account the extra layer of complexity introduced by the xrootd interfaces. Also network connectivity has always been reliable; furthermore, the recent upgrade to 40Gb/s of the WAN link makes CNAF one of the better-connected sites in the ALICE Computing Grid.

5. References

[1] K. Aamodt et al. (ALICE Collaboration), JINST **3** S08002 (2008).
[2] N. Cabibbo, G. Parisi, Phys. Lett. **B 59** 67 (1975).
[3] E.V. Shuryak, Phys. Rep. **61** 71 (1980).
[4] K. Aamodt et al. (ALICE Collaboration), Eur. Phys. J. **C 65** 111-125 (2010).
[5] K. Aamodt et al. (ALICE Collaboration), Eur. Phys. J. **C 72** 2124 (2012).
[6] K. Aamodt et al. (ALICE Collaboration), Phys. Rev. Lett. **105** 252301 (2010).
[7] K. Aamodt et al. (ALICE Collaboration), Phys. Lett. **B 696** 328-337 (2011).
[8] K. Aamodt et al. (ALICE Collaboration), Phys. Rev. Lett. **105** 252302 (2010).
[9] B. Abelev et al. (ALICE Collaboration), J. High Energy Phys. **9** 112 (2012).
[10] E. Abbas et al. (ALICE Collaboration), Phys. Rev. Lett. **111** 162301 (2013).
[11] B. Abelev et al. (ALICE Collaboration), Phys. Rev. Lett. **111** 102301 (2013).
[12] B. Abelev et al. (ALICE Collaboration), Phys. Rev. Lett. **109** 252301 (2012).
[13] B. Abelev et al. (ALICE Collaboration), Physics Letters **B 728** 25-38 (2014).
[14] B. Abelev et al. (ALICE Collaboration), Physics Letters **B 726** 164-177 (2013).
[15] B. Abelev et al. (ALICE Collaboration), Physics Letters **B 718** 1273-1283 (2013).
[16] P. Cortese et al. (ALICE Collaboration), CERN-LHCC-2005-018 (2005).
[17] S. Bagnasco et al., Journal of Physics: Conf. Series **368** 012019 (2012).

ATLAS activities

A De Salvo[1]

E-mail: Alessandro.DeSalvo@roma1.infn.it

Abstract. In this paper we describe the computing activities of the ATLAS experiment at LHC, CERN, in relation to the Italian Tier-1 located at CNAF, Bologna. The major achievements in terms of computing and physics results are briefly discussed, together with the impact of the Italian community on the computation of the results.

1. Introduction

ATLAS is one of two general-purpose detectors at the Large Hadron Collider (LHC). It investigates a wide range of physics, from the search for the Higgs boson and standard model studies to extra dimensions and particles that could make up dark matter.

Beams of particles from the LHC collide at the centre of the ATLAS detector making collision debris in the form of new particles, which fly out from the collision point in all directions. Six different detecting subsystems arranged in layers around the collision point record the paths, momentum, and energy of the particles, allowing them to be individually identified. A huge magnet system bends the paths of charged particles so that their momenta can be measured.

The interactions in the ATLAS detectors create an enormous flow of data. To digest the data, ATLAS uses an advanced trigger system to tell the detector which events to record and which to ignore. Complex data-acquisition and computing systems are then used to analyse the collision events recorded. At 46 m long, 25 m high and 25 m wide, the 7000-tons ATLAS detector is the largest volume particle detector ever constructed. It sits in a cavern 100 m below ground near the main CERN site, close to the village of Meyrin in Switzerland.
More than 3000 scientists from 174 institutes in 38 countries work on the ATLAS experiment.
ATLAS has been taking data from 2010 to 2012, at center of mass energies of 7 and 8 TeV, collecting about 5 and 20 fb-1 of integrated luminosity, respectively.

The experiment has been designed to look for New Physics over a very large set of final states and signatures, and for precision measurements of known Standard Model (SM) processes.

[1] INFN, Sez. Roma1

Its most notable result up to now has been the discovery of a new resonance at a mass of about 125 GeV, followed by the measurement of its properties (mass, production cross sections in various channels and couplings). These measurements have confirmed the compatibility of the new resonance with the Higgs boson, foreseen by the SM but never observed before.
As an example of the wide range of Physics measurements and searches carried on by the ATLAS experiment, a summary of the main results obtained during the year 2013 is detailed below.

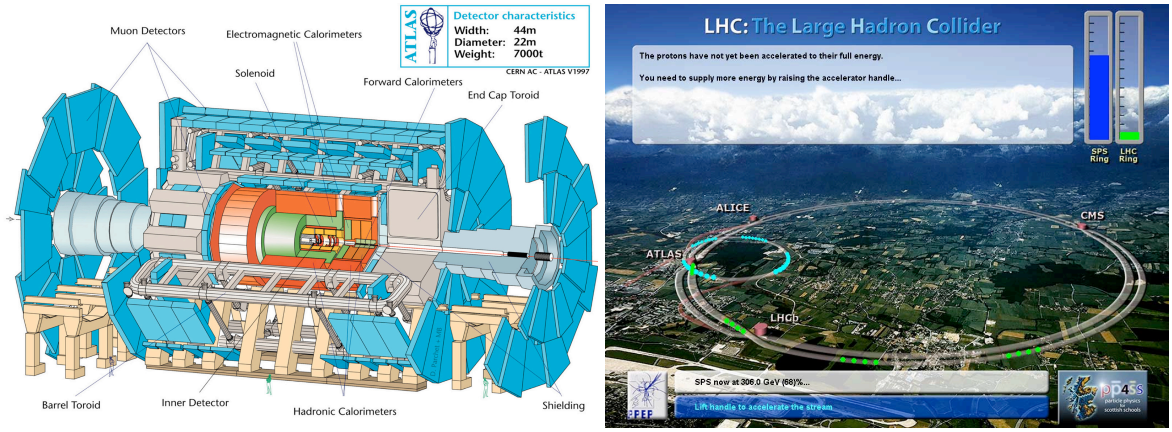


Figure 1 - The ATLAS experiment at LHC

2. Main Physics achievements in 2013

Here we briefly summarize the main physics results obtained by ATLAS in 2013.

2.1.1. SUSY

ATLAS has published 5 papers in 2013 and 21 conference notes on SUSY searches. These are designed to cover a broad spectrum of different models: standard scenarios with prompt production and R parity conservation; scenarios with R-parity violation; long-lived particles.
Concerning models with prompt production and R-parity conservation, searches have been updated with the data collected at 8 TeV. These comprise strong production of squarks and gluinos with final states from 0 to 2 leptons and several (b-)jets. Squarks and gluinos with masses below about 1.5 TeV have been excluded.
Direct stop production have been searched looking for stop decays to top neautralino, bottom chargino and charm neutralino.
Final states with 0 to three leptons and 0 to 3 (b-)jets have been considered, excluding the stops with a mass less than about 650 GeV.
Electroweak production have also been investigated, looking both for neutralinos and charginos decays to sleptons and for decays to W/Z bosons plus lightest neutralinos to take into account models where only neutralinos are below the TeV mass scale.
Limits on the chargino and next-to-lightest neutralino masses ranging from 300 to 600 GeV have been put. In models with R-parity violation, the lightest neutralino could decay to two charged leptons and a neutrino or two quarks and a charged lepton. Seaveral searches have been performed looking for prompt and displaced decays with final leptons in the final state.
Further searches have been done for long-lived particles also in R-parity conserving models. In both the cases, no deviation from the SM expactation has been observed.

2.1.2. Exotics

In 2013 ATLAS has published about 25 paper on searches of new exotic particles and more than 10 conference notes.
New limits have been put on the existence of new heavy gauge bosons (Z' and W') up to masses of about 2.8 TeV; on resonances forseen by models with extra dimensions, like gravitons up to masses of 2.5 TeV, depending on the specific model, or quantum black holes; on excited quarks or leptons, where the excited quarks have been excluded up to masses of 3.8 TeV.
Of particular importance are the searches for a new 4th generation quark family and in particular for vectorlike quarks (VLQ). These are predicted, for instance, by models of partial compositness. The following decays have been looked for: T->Wb,Zt,Ht and B->Wt,Zb,Hb, where T and B are two VLQ. This 4th family couples strongly with the third one, leading to final states with several tops and b's.
The searches therefore require final states with one or more leptons, several jets with two or more tagged as b-jets. These states are excluded up to masses of about 700 GeV.
Top resonances have been also searched, looking for two tops in the final state: new states as axigluons, KK gluons, leptophobic Z' have been excluded up to masses of about 2.5 TeV.
Dark matter searches have been performed looking for events where two weakly interacting particle escaping detection are produced in association with W or Z bosons or gluons. interesting limits have been put in particular in the low mass region where direct detection experiment have less sensitivity.
Hidden-sector models predict the existence of dark photons that could convert into pairs of ordinary leptons. The simultaneous decay of 2 or more energetic dark-photons would lead to events with 2 or more very collimated leptons, so called lepton-jets.
Searches have been performed both for muon and electron jets and for prompt and displaced vertexes. No signal have been found of such events allowing us to set limits in on branching ratios and/or lifetimes of these particles.

2.1.3. Heavy Ions

ATLAS published in 2013 6 papers and 5 conference notes on heavy-ions physics.
Both proton-lead and lead-lead collision data have been analysed.
Proton-lead data have been collected in a short run in September 2012 and a longer run in January/February 2013.
The energy of protons was 4 TeV and that of Pb208 was 82x4 TeV, leading to a nucleon-nucleon CM energy of 5.02 TeV.
An important outcome of proton-lead data was the observation of a long-range rapidity-correlation as observed in proton-proton collisions, the so-called ridge-effect. The effect is stronger in p-Pb collisions and is present both in the near and away sides.
Proton-lead collisions provide also a unique opportunity to study violation of factorization in hard scattering and for observing parton saturation effects. Such effects have been investigated by measuring the centrality dependence of jet production.
Strong suppression of high energy jets for central collisions with respect to peripheral collisions has been observed.
Central-to-peripheral ratio of jets has been measured by ATLAS in Pb-Pb collisions at nucleon-nucleon CM energy of 2.76 TeV.
Jet production is found to be suppressed a factor two in the 10% most central collisions relative to peripheral collisions.
On the contrary no suppresion has been observed in weak boson production, consistent with the hypothesis that the suppression observed for hadrons is due to final-state interactions with the hot medium.

The properties of the new state produced in Pb-Pb collisions have been investigated by measuring the azimuthal anisotropy of particle emission in the transverse plane, finding good agreement with hydrodynamic-model descriptions.

2.1.4. Higgs

In 2013 ATLAS published 5 papers and 8 conference notes on Higgs searches and its properties.
The focus of the work in Higgs analysis during 2013 has been carrying on and completing the analysis of run-1 data.
The results for all the main Standard Model (SM) channels have been updated based on the full dataset of about 5 fb-1 at 7 TeV and 20 fb-1 at 8 TeV (for some measurements the dataset at 8 TeV has been used).
Among the most important addressed points there was the measurement of the properties of the new resonance, whose discovery was announced in July 2012.
First of all the mass, measured with the two high-precision decay channels H->ZZ*->4-lepton and H->2gamma. The combined result for the Higgs mass is mH=125.5 ±0.2 (stat) +0.5-0.6 (sys) GeV.
The combined signal strength for the two channels, at the combined mass value, is mu=1.43 ± 0.16 (stat) ± 0.14 (syst).
The results for the other important discovery channel, i.e. H->WW->lnulnu were also updated using the full available luminosity of about 25 fb-1, observing an excess above the backgrounds of 3.8 standard deviations (3.7 expected for a SM Higgs Boson).
The combination of all SM channels was updated, including 25 fb-1 for H->ZZ->4l, H->2gamma and H->WW, and about 18 fb-1 (of which 4.7 fb-1 at 7 TeV) for the fermionic H->bb and H->tau+tau- channels.
The combined signal strength at a mass of 125.5 GeV was determined to be mu=1.30 ±0.13 (stat) ±0.14 (syst), and the cross section ratio.
The full set of fits testing the fermion and vector coupling sector, couplings to W and Z and loop-induced, showed no significant deviation from the SM expectation.
The spin and parity of the new resonance were studied in in the H->ZZ*->4l, H->2gamma and H->WW->lnulnu channels, leading to the evidence for the spin-0 nature of the Higgs boson.
The data strongly favour the J^P=0^+ hypothesis. The specific J^P=2^+ hypothesis was excluded with a confidence level above 99.9%. The 0-,1+,1- hypotheses were also excluded at a confidence level of 97.8% or higher.
First measurements of the differential production cross sections were performed during 2013, in the H->2gamma channel. All measurements were also in this case compatible with the SM expectation.
The search for the Higgs boson with a mass of about 125 GeV in the H->tau+tau- channel has been updated, looking at the 20.3 fb-1 at 8 TeV. The observed (expected) deviation from the background-only hypothesis corresponds to a significance of 4.1 (3.2) standard deviations, while the measured signal strength is mu=1.4 +0.5-0.4. This is the evidence for H->tau+tau- decay, and the results are consistent with the SM expectation.
Also the search in the H->bb decay channel, in particular in the associated production VH (H->bb) was updated on the full run-1 dataset.

2.1.5. Standard Model

In 2013 ATLAS published 10 papers and 7 conf-notes with results of Standard Models Physic.
Other Standard Model measurements were performed on the 4.6 fb-1. Those included W boson production in association with b-jets and of jets in association to a Z boson.

Also the production cross section of a W boson in association with a charm quark was measured using the same dataset.
High-mass Drell-Yan differential cross section and di-jet production cross section. The fraction of events with double-parton interactions was measured using W->lnu+2 jets events.

2.1.6. B-physics

In 2013 ATLAS published 2 papers and 7 conf-notes with results of B Physics.

2.1.7. Top physics

In 2013 ATLAS published 4 papers and 11 conf-notes with results of Top Physics (excluding Top/Exotic searches).

3. The ATLAS Computing System

The ATLAS Computing System[1] is responsible for the provision of the software framework and services, the data management system, user-support services, and the world-wide data access and job-submission system. The development of detector-specific algorithmic code for simulation, calibration, alignment, trigger and reconstruction is under the responsibility of the detector projects, but the Software & Computing Project plans and coordinates these activities across detector boundaries. In particular, a significant effort has been made to ensure that relevant parts of the “offline” framework and event-reconstruction code can be used in the High Level Trigger. Similarly, close cooperation with Physics Coordination and the Combined Performance groups ensures the smooth development of global event-reconstruction code and of software tools for physics analysis.

3.1.1. The ATLAS Computing Model

The ATLAS Computing Model [2] embraces the Grid paradigm and a high degree of decentralisation and sharing of computing resources. The required level of computing resources means that off-site facilities are vital to the operation of ATLAS in a way that was not the case for previous CERN-based experiments. The primary event processing occurs at CERN in a Tier-0 Facility. The RAW data is archived at CERN and copied (along with the primary processed data) to the Tier-1 facilities around the world. These facilities archive the raw data, provide the reprocessing capacity, provide access to the various processed versions, and allow scheduled analysis of the processed data by physics analysis groups. Derived datasets produced by the physics groups are copied to the Tier-2 facilities for further analysis. The Tier-2 facilities also provide the simulation capacity for the experiment, with the simulated data housed at Tier-1s. In addition, Tier-2 centres will provide analysis facilities, and some provide the capacity to produce calibrations based on processing raw data. A CERN Analysis Facility provides an additional analysis capacity, with an important role in the calibration and algorithmic development work. ATLAS has adopted an object-oriented approach to software, based primarily on the C++ programming language, but with some components implemented using FORTRAN and Java. A component-based model has been adopted, whereby applications are built up from collections of plug-compatible components based on a variety of configuration files. This capability is supported by a common framework that provides common data-processing support. This approach results in great flexibility in meeting both the basic processing needs of the experiment, but also for responding to changing requirements throughout its lifetime. The heavy use of abstract interfaces allows for different

implementations to be provided, supporting different persistency technologies, or optimized for the offline or high-level trigger environments.
The Athena framework is an enhanced version of the Gaudi framework that was originally developed by the LHCb experiment, but is now a common ATLAS-LHCb project. Major design principles are the clear separation of data and algorithms, and between transient (in-memory) and persistent (in-file) data. All levels of processing of ATLAS data, from high-level trigger to event simulation, reconstruction and analysis, take place within the Athena framework; in this way it is easier for code developers and users to test and run algorithmic code, with the assurance that all geometry and conditions data will be the same for all types of applications (simulation, reconstruction, analysis, visualization).
One of the principal challenges for ATLAS computing is to develop and operate a data storage and management infrastructure able to meet the demands of a yearly data volume of O(10PB) utilized by data processing and analysis activities spread around the world. The ATLAS Computing Model establishes the environment and operational requirements that ATLAS data-handling systems must support and provides the primary guidance for the development of the data management systems.
The ATLAS Databases and Data Management Project (DB Project) leads and coordinates ATLAS activities in these areas, with a scope encompassing technical data bases (detector production, installation and survey data), detector geometry, online/TDAQ databases, conditions databases (online and offline), event data, offline processing configuration and bookkeeping, distributed data management, and distributed database and data management services. The project is responsible for ensuring the coherent development, integration and operational capability of the distributed database and data management software and infrastructure for ATLAS across these areas.
The ATLAS Computing Model defines the distribution of raw and processed data to Tier-1 and Tier-2 centres, so as to be able to exploit fully the computing resources that are made available to the Collaboration. Additional computing resources are available for data processing and analysis at Tier-3 centres and other computing facilities to which ATLAS may have access. A complex set of tools and distributed services, enabling the automatic distribution and processing of the large amounts of data, has been developed and deployed by ATLAS in cooperation with the LHC Computing Grid (LCG) Project and with the middleware providers of the three large Grid infrastructures we use: EGEE, OSG and NorduGrid. The tools are designed in a flexible way, in order to have the possibility to extend them to use other types of Grid middleware in the future.
The main computing operations that ATLAS have to run comprise the preparation, distribution and validation of ATLAS software, and the computing and data management operations run centrally on Tier-0, Tier-1s and Tier-2s. The ATLAS Virtual Organization allows production and analysis users to run jobs and access data at remote sites using the ATLAS-developed Grid tools.
The Computing Model, together with the knowledge of the resources needed to store and process each ATLAS event, gives rise to estimates of required resources that can be used to design and set up the various facilities. It is not assumed that all Tier-1s or Tier-2s are of the same size; however, in order to ensure a smooth operation of the Computing Model, all Tier-1s usually have broadly similar proportions of disk, tape and CPU, and similarly for the Tier-2s.
The organization of the ATLAS Software & Computing Project reflects all areas of activity within the project itself. Strong high-level links are established with other parts of the ATLAS organization, such as the T-DAQ Project and Physics Coordination, through cross-representation in the respective steering boards. The Computing Management Board, and in particular the Planning Officer, acts to make sure that software and computing developments take place coherently across sub-systems and that the project as a whole meets its milestones. The International Computing Board assures the information flow between the ATLAS Software & Computing Project and the national resources and their Funding Agencies.

4. The role of the Italian Computing facilities in the global ATLAS Computing

Italy provides Tier-1, Tier-2 and Tier-3 facilities to the ATLAS collaboration. The Tier-1, located at CNAF, Bologna, is the main centre, also referred as "regional" centre. The Tier-2 centres are distributed in different areas of Italy, namely in Frascati, Napoli, Milano and Roma. 3 out of 4 Tier-2 sites are also considered as Direct Tier-2 (T2D), meaning that they have an higher importance with respect to normal Tier-2s and can have primary data too. Frascati, which is the only Tier-2 not yet Direct, will probably be upgraded to the higher category in the next months, when its size will reach the minimum required to be considered as T2D. The total of the T2 sites corresponds to more than the total ATLAS size at the T1, for what concerns disk and CPUs; tape is not available in the T2 sites.
A third category of sites is the so-called Tier-3 centres. Those are smaller centres, scattered in different places in Italy, that nevertheless contributes in a consistent way to the overall computing power, in terms of disk and CPUs. The overall size of the Tier-3 sites corresponds roughly to the size of a Tier-2 site. The Tier-1 and Tier-2 sites have pledged resources, while the Tier-3 sites do not have any pledge resource available.
In terms of pledged resources, Italy contributes to the ATLAS computing as 10% of both CPU and disk for the Tier-1. The share of the T2 facilities corresponds to 7% of disk and 9% of CPU of the whole ATLAS computing infrastructure.
The Italian Tier-1, together with the other Italian centres, provides both resources and expertise to the ATLAS computing community, and manages the so-called Italian Cloud of computing. Up to 2013 the Italian cloud does not only include Italian sites, but also T3 sites of other countries, namely South Africa and Greece.
The computing resources, in terms of disk, tape and CPU, available in the Tier-1 at CNAF have been very important for all kind of activities, including event generation, simulation, reconstruction and analysis, for both MonteCarlo and real data. Its major contribution has been the data reprocessing, since this is a very I/O and memory intense operation, normally executed only in Tier-1 centres. In this sense CNAF has played a fundamental role for the Higgs discovery in 2012 [3] and the subsequent refinement in the measurements in 2013. All the efforts lead to the Nobel Prize to Peter Higgs in 2013, for the Higgs boson prediction and the following discovery at LHC.
The Italian centres, including CNAF, have been very active not only in the operation side, but contributed a lot in various aspect of the Computing of the ATLAS experiment, in particular for what concerns the network, the storage federations and the monitoring tools.
The T1 at CNAF has been very important for the ATLAS community in 2013, for some specific activities:

1) test and fine tuning of the Xrootd federation using the StoRM storage system, completely developed by CNAF within the LCG and related projects, funded by EU;
2) initial setup and test of the WebDAV/HTTPS access for StoRM, also using it in a HTTP storage federation, alternative of the Xrootd one;
3) setting up, testing and maintenance in production mode of the first multi-core queues with the LSF resource management system;
4) early-adoption of the Perfsonar-PS network monitoring system, to be used with the LHCONE overlay network, together with the other T1 and T2 sites.

5. Performance of the Italian sites within the ATLAS Computing infrastructure

Up to the end of 2013 ATLAS collected and registered at the Tier0 about 34 PB of raw and derived data, while the cumulative data volume distributed in all the data centres in the grid was of the order of 155PB.
The data, in 2013, has been replicated with an efficiency of 100% and an average throughput of more than 50 Gbps with monthly average peaks above 90 Gbps. For just Italy, the average throughput was of the order of 4 Gbps with weekly average peaks above 8 Gbps. The data replication speed from

Tier0 to the Tier2s has been quite fast with a transfer time lower than 4 hours. The increase in performance of the data transfer has been also helped a lot by the new generation of network connections and the dedicated overlay network LHCONE, designed to connect the Tier2 sites of LHC. Already since mid 2012 almost all the Tier2 centres have been connected at >= 10 Gbps via the NREN networks, while the T1s were already connected at 10 Gbps via the dedicated infrastructure LHCOPN. The Italian T1 will soon upgrade its bandwidth to 40 Gbps, allowing for better performance and increased capacity, suitable for direct WAN access and similar techniques, very important, for example, in the storage federation domains.
The average number of simultaneous jobs running on the grid has been of about 55k for both production (simulation and reconstruction) and data analysis with peaks up to 62k, corresponding to a weekly average of about 2.9M of completed jobs with success.
The use of the grid for analysis has been stable on ~35k simultaneous jobs, with peaks around the conferences' periods to over 50k, showing the reliability and effectiveness of the use of grid tools for data analysis. In order to improve the reliability and efficiency of the whole system, in 2012 ATLAS introduced the so-called Federation of Xrootd storage systems (FAX), on top of the existing infrastructure. Using FAX, the users will have the possibility to access remote files via the XRootd protocol in a transparent way, using a global namespace and a hierarchy of redirectors, thus reducing the number of failures due to missing or not accessible local files, while also giving the possibility to relax the data management and storage requirements in the sites. The testing of the FAX federation started in mid 2012 and Italy joined it with 3 Tier2 sites in November 2012. During 2013, with the storage upgrades and experience gained during the pilot phase, FAX was extended and now is very close to pre-production mode.

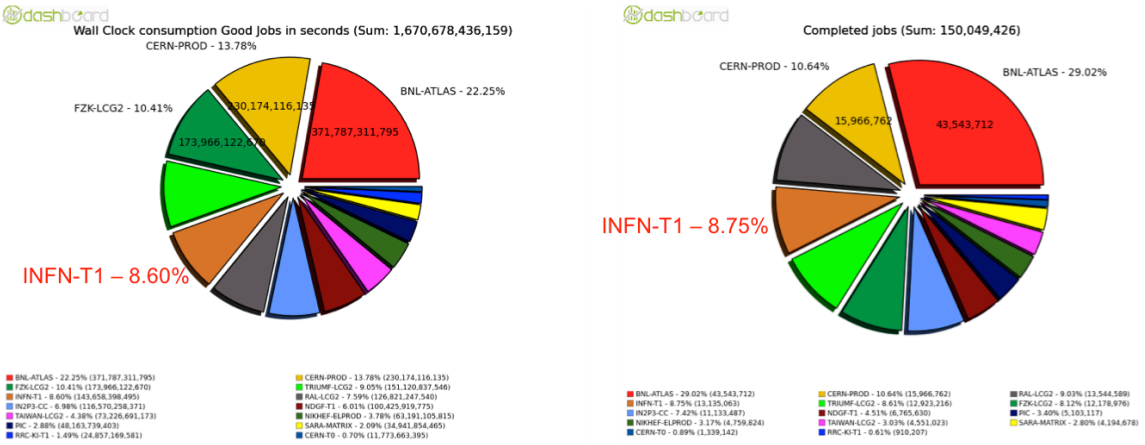


Figure 2 - The contribution of CNAF to the overall ATLAS computing, in terms of Wall Clock consumption and number of completed jobs in 2013

Concerning the INFN T1 specifically, the resources available to ATLAS has been constantly fully used, and in most of the cases even used above the dedicated share and pledges, when they were not used by other activities. In Figure 3 the Wall Clock Time from 2011 to the beginning of 2014 is shown, together with the pledges of the respective years. From the plot is clear that the amount of resources used were almost all the time > 20% over the pledges. This numbers well compare with the same values in the total group of T1s in ATLAS, show in Figure 3, right plot.
The efficiency of the jobs running at CNAF is above 90%, in line with the values of the other T1 sites of ATLAS, as show in Figure 4.

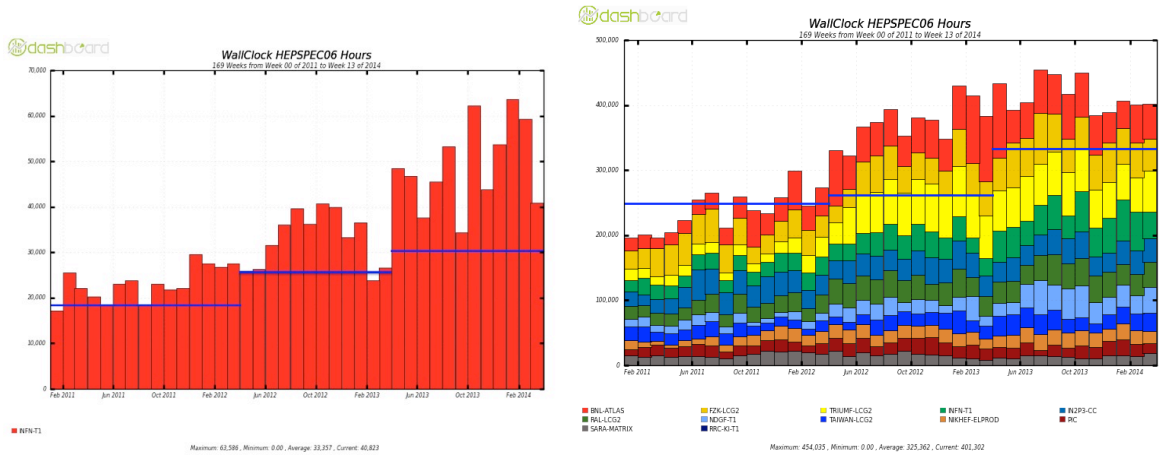


Figure 3 - The left plot shows the Wall Clock Time used at CNAF by the ATLAS jobs since 2011. The blue lines are representing the pledges of the respective years. In the right plot the Wall Clock Time in all the ATLAS T1s is also show, together with the global pledges.

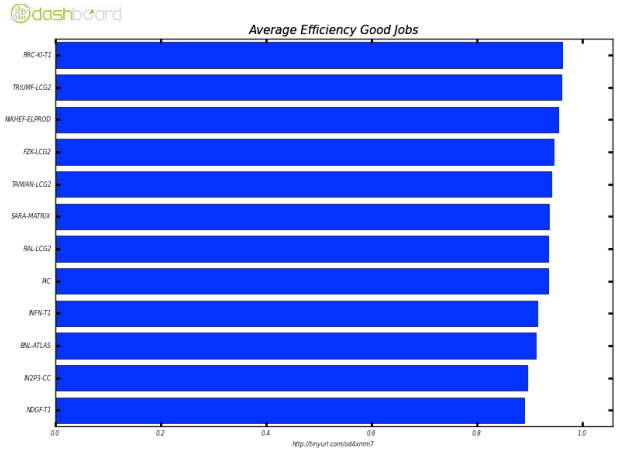


Figure 4 - The average efficiency of the ATLAS Tier-1 Centers in 2011-2013. CNAF has an average efficiency of more than 90%, comparable to the efficiency obtained in the other sites.

The disk space used at CNAF has also been fully exploited. The bigger amount of data stored on disk is represented by the .NTUP data format, corresponding to the n-tuples used for the analysis, followed by the AOD (Analysis Data Objects) and ESD (Event Summary Data), as shown in Figure 5. The tape occupancy in show in Figure 6: in this case the biggest volume is represented by the RAW data, stored as custodial data in all the T1s.
From the point of view of the data transfer rate, as show in Figure 7, CNAF has always been over 250 MB/s for the ATLAS-related data transfer, with an average efficiency of more than 90% at the first attempt and 100% including the retries.

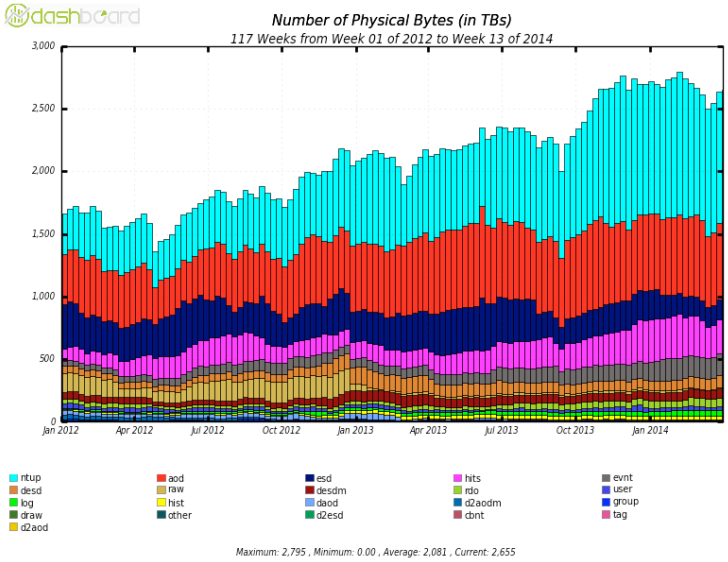


Figure 5 - Number of physical bytes, grouped by data format, stored by ATLAS on the INFN Tier-1 disks from 2012 to the beginning of 2014. The most popular data format is the NTUP, corresponding to the ntuples used for analysis.

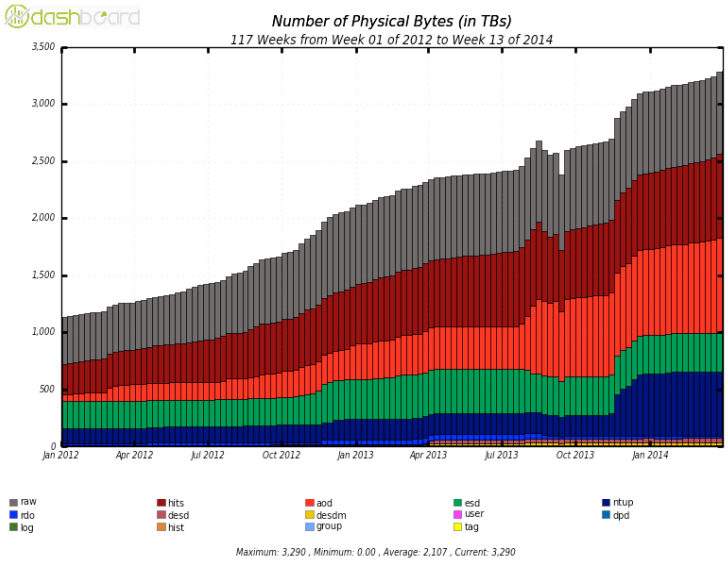


Figure 6 - The evolution of the tape space used by ATLAS in the INFN Tier-1 centre between 2012 and the beginning of 2014. The biggest volume stored is represented by the RAW data.

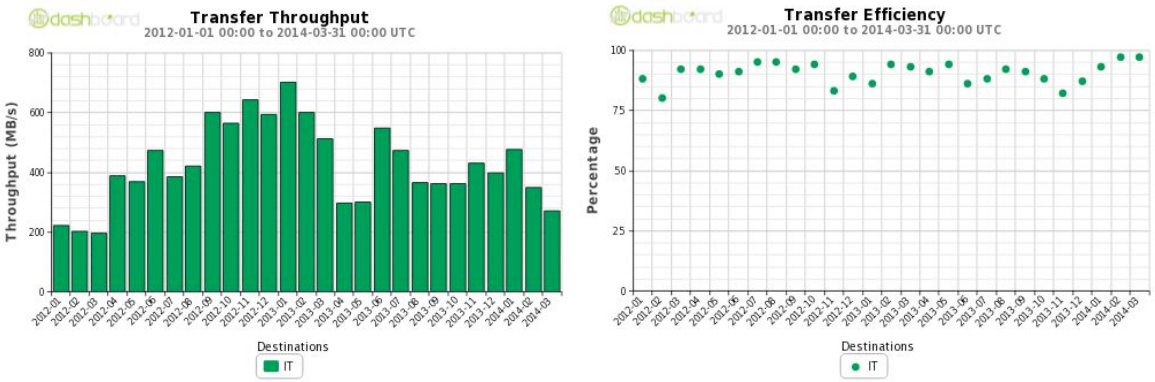


Figure 7 - Transfer rate for ATLAS data to CNAF (right plot) and transfer efficiency (left plot).

6. References

[1] The ATLAS Computing Technical Design Report ATLAS-TDR-017; CERN-LHCC-2005-022, June 2005
[2] The evolution of the ATLAS computing model; R W L Jones and D Barberis 2010 J. Phys.: Conf. Ser. 219 072037 doi:10.1088/1742-6596/219/7/072037
[3] Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC, the ATLAS Collaboration, Physics Letters B, Volume 716, Issue 1, 17 September 2012, Pages 1–29

Pierre Auger Observatory Data Simulation and Analysis at CNAF.

G Cataldi[1] and the Pierre Auger Collaboration[2]

[1] Istituto Nazionale Fisica Nucleare, sezione di Lecce, Italy.
[2] Observatorio Pierre Auger, Av. San Martìn Norte 304, 5613 Malargüe, Argentina
(Full author list : http://www.auger.org/archive/authors_2013_06.html)

E-mail: Gabriella.Cataldi@le.infn.it

Abstract. The Pierre Auger Observatory is described. The adopted computing model is summarized and the computing organization of the Italian part of the collaboration is explained. The flux of cosmic rays above 3×10^{17}eV has been measured with unprecedented precision at the Pierre Auger Observatory based on data in the period between January 1st 2004 and December 31st 2012. For realizing such a precision a reliable measurement of the exposure must be calculated.

1. Introduction

The Pierre Auger Observatory, built near the town of Malargüe in Argentina, has been gathering data since January 2004 [1]. It reached its baseline design covering 3000 km^2 with 1600 water Cherenkov surface detectors (SD) overlooked by 24 fluorescence telescopes (FD) by mid 2008 and by the end of 2009 had accumulated a total exposure of about twenty thousand km^2 sr yr, much larger than that of all previous air shower experiments combined. The surface detector has a duty cycle of almost 100% collecting then the vast majority of the data which are used for spectrum measurements and anysotropies search. The simultaneous observations with both the fluorescence and surface detectors (hybrid observations) are possible for $\sim$ 13% of the events (those observed during moonless and clear nights), for which the longitudinal development in the atmosphere as well as the lateral profile on the ground can be measured. This allows the cross calibration between the two detection techniques and to determine the depth of maximum development of the shower, which encodes precious information on the composition of the primaries and the properties of the first hadronic interactions. The studies of the cosmic rays at the highest energies with the Auger Observatory has already allowed to start addressing many of the old questions that motivated its construction by measuring the features present in the spectrum, searching for anisotropies in the cosmic ray arrival directions distribution or constraining the composition of the primary cosmic rays.

2. Organization of the Auger analysis.

The date acquired at the Auger observatory are daily mirrored in sites, located in Lyon, Fermilab and Buenos Aires. Starting from these mirroring sites, the data are collected by the collaboration groups and they are used for reconstruction and analysis. At CNAF the data are daily transferred from Lyon allowing an easy access for the italian groups. The most challanging

task in term of CPU and SE allocation is the simulation process. This process can be divided in two steps: the simulation of the shower development in the atmosphere and the simulation of the shower interaction with the experimental apparatus. The two steps show completely different problematics and are fully separated, making use of different codes. For the shower development in the atmosphere, the code is based on the Corsika library[2]. This software is not a property of the Auger collaboration and it does not require external libraries (apart from FLUKA). For the detector simulation, the collaboration run a property code, based on Geant4 and needing several libraries as external. The shower simulation in the atmosphere requires the use of interaction hadronic models for simulating the interaction processes. These models are built starting from beam measurements taken at energies much lower then the ones of interest for Auger, and therefore can exhibits strong differences that must be evaluated in the systematics. The collaboration plans and defines through the simulation committee a massive production of the two simulation steps, that are executed under GRID environment. Concerning the second step, i.e. the simulation of the shower interaction with the experimental apparatus, the only GRID running environment is the so called *ideal detector* that does not consider during the simulation phase the uncertainties introduced by the data taking conditions.

3. Organization of the Italian Auger Computing

The national Auger cluster is located and active at CNAF since the end of 2010. The choice has allowed to use all the competences for the management and the GRID middleware of computing resources that are actually present among the CNAF staff. The cluster serves as Computing Element (CE) and Storage Element (SE) for all the Italian INFN groups. On the CE the standard version of reconstruction, simulation and analysis of Auger collaboration libraries are installed and updated, a copy of the data is kept, and the Databases, accounting for the different data taking conditions are up to date. The CE and part of the SE are included in the Auger production GRID for the simulation campaign. On the CE of CNAF the simulation and reconstruction mass productions are mainly driven from the specific requirements of the italian groups. On the remaing part of the SE, the simulated libraries, specific to the analysis of INFN group are kept. At CNAF there are two main running environments, corresponding to two different queues: *auger* and *auger_db*. The first is mainly used for mass production of Corsika simulation, and for the simulation of shower interaction with the atmosphere in condition independent from the environmental data. The second environment (*auger_db*) is an ad hoc configuration that allows the running of the offline in dependence with the running condition databases. CNAF is at present the only GRID infrastructure where this kind of environment can be run. The particular setup uses the WNodes environment with the Database accessed from the instantiated virtual machines. A specific configuration allows a suitable load to the DB servers.

4. The flux measurement of the Ultra High Energy Cosmic Rays

Given the very specific configuration for the Auger CNAF we restrict this section to the measurement that is performed at CNAF using *auger_db*, i.e. the flux measurement of the hybrid detector. The hybrid approach is based on the detection of showers observed by the FD in coincidence with at least one station of the SD array. Although a signal in a single station does not allow an independent trigger and reconstruction in SD, it is a sufficient condition for a very accurate determination of the shower geometry using the hybrid reconstruction. In order to determine the cosmic ray spectrum, a reliable estimate of the exposure is needed, and hence a strict event selection is performed [3]. A detailed simulation of the detector response has shown that for zenith angles below 60°, every FD event above 10^{18} eV passing all the selection criteria is triggered by at least one SD station, independent of the mass or direction of the incoming primary particle. The measurement of the flux of cosmic rays using hybrid events relies on the precise determination of the detector exposure that is influenced by several factors. The response

of the hybrid detector strongly depends on energy and distance from the relevant fluorescence telescopes, as well as atmospheric and data taking conditions. To properly take into account all of these configurations and their time variability, the exposure has been calculated using a sample of simulated events that reproduce the exact conditions of the experiment. The current hybrid exposure as a function of energy is shown in Figure 1 compared with the exposures of the surface detectors.

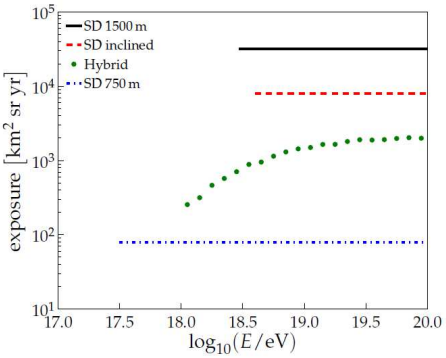


Figure 1. The integrated exposure of the different detectors at the Pierre Auger Observatory as a function of energy. The SD exposure in the three cases is at above the energy corresponding to full trigger efficiency for the surface arrays.

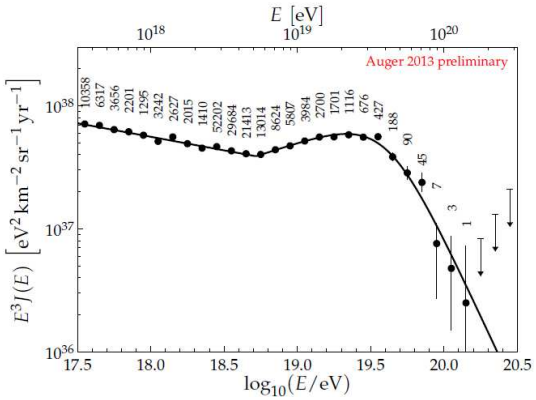


Figure 2. The combined energy spectrum as measured at the Pierre Auger Observatory. The numbers give the total number of events inside each bin. The last three arrows represent upper limits at 84% C.L.

Figure 2 shows the flux of cosmic rays above 3×10^{17}eV that has been measured combining data from surface and fluorescence detectors. There are two clear transitions: at 4×10^{18}eV there is a flattening of the spectrum (the ankle) and a strong flux suppression above 5×10^{19}eV. The physical origin of the ankle is still uncertain, being the main candidate scenarios to explain this feature those relating it to the transition from a dying galactic component to a harder extragalactic component becoming dominant, or alternatively the so-called dip-scenario [5], in which cosmic rays are assumed to be extragalactic protons down to energies below 10^{18}eV and the concave shape observed arises from the effect of energy losses by pair creation with cosmic microwave background (CMB) photons. The second feature mentioned, i.e. the suppression observed at the highest energies, is similar to the expectations from the so called GZK effect associated to the attenuation of extragalactic protons by photo pion production of CMB photons but might also be due to the maximum source energy. The exact physical explanation of the observed spectral features remains uncertain and the precise measurement of mass composition and of the flux at energies above 10^{17} eV is crucial for discriminating between different theoretical models.

References

[1] The Pierre Auger Collaboration 2010 *Nucl. Instr. and Methods in Physics Research* A **613** 29
[2] J. Knapp and D. Heck 1993 *Extensive Air Shower Simulation with CORSIKA, KFZ Karlsruhe* **KfK 5195B**
[3] The Pierre Auger Collaboration 2011 *Astropart. Phys.* **34** 368
[4] M. Tueros for the Pierre Auger Collaboration 2013 *Proc. 33rd ICRC, Rio de Janeiro, Brazil* **arXiv:1307.5059**
[5] V. S. Berezinsky and S. I. Grigorieva 1988 *Astron. and Astrophys.* **199** 1

The BaBar Experiment at the INFN CNAF Tier1

F. Bianchi
INFN and University of Torino, via Giuria 1, 10135 torino, Italy

E-mail: fabrizio.bianchi@to.infn.it

Abstract. The BaBar detector operated successfully at the PEP-II collider at the SLAC National Accelerator Laboratory from 1999 to 2008. This note describes its computing model with a focus on the role of CNAF as Tier1 centre.

1. Introduction

The BaBar detector took data at the PEP-II asymmetric e^+ e^- collider at the SLAC National Accelerator Laboratory from 1999 to 2008. The experiment was optimized for detailed studies of CP-violating asymmetries in the decay of B mesons, but it was well suited for a large variety of other studies, like precision measurements of decays of mesons with b and c quarks and of τ leptons, and searches for rare processes, including many not expected in the framework of the Standard Model of electroweak interactions.
The PEP-II collider operated in the center-of-mass energy between 9.99 GeV (just below the Y(2S) resonance) and 11.2 GeV, mostly at 10.58 GeV, corresponding to the mass of the Y(4S) resonance. This resonance decays exclusively to pairs of neutral and charged B mesons and thus provides an ideal laboratory for their study. At the Y(4S) resonance, the electron beam of 9.0 GeV collided head-on with the positron beam of 3.1 GeV resulting in a Lorentz boost to the Y(4S) resonance of $\beta\gamma = 0.56$. This boost made it possible to reconstruct the decay vertices of the B mesons, to determine their relative decay times, and to measure the time dependence of their decay rates, a feature that was critical for the observation of CP-violation in the B mesons system.

2. Computing Model

The BaBar experiment collected around one Petabyte of “raw data” that have been permanently stored on tape, calibrated, and reconstructed usually within 48 hours of the actual data taking. Reconstructed data have been permanently stored in a compact format suitable for subsequent physics analysis, the so-called “microDST”.
Many samples of Monte Carlo events, corresponding to different sets of physics channels, have been generated and reconstructed. In addition to the physics triggers, the data acquisition also recorded random triggers that have been used to create “background frames” that have been superimposed on the generated Monte Carlo events to account for the effects of the machine background and of electronic noise, before the reconstruction step.
Detector and Monte Carlo data have been centrally “skimmed” to produce subsets of selected events, the “skims”, designed for a specific area of analysis. Skims are very convenient for physics analysis,

but they increase the storage requirements because the same event can be present in more than one skim.
The quality of the detector data and of the simulated events has been monitored through all the steps of processing. From time to time, as improvements in detector calibration constants and/or in the code were implemented, the detector data have been reprocessed and new samples of simulated data generated. When sets of new skims become available, an additional skim cycle was run on all the events.
The total amount of data produced by BaBar was over six Petabyte.
BaBar has been one of the first experiments to adopt the C++ programming language to write offline and online software. In the mid-nineties, when this decision was taken, the dominant language in the High Energy Physics (HEP) community was Fortran 77. However, problems and limitations associated with this language were becoming very clear and BaBar chose early to commit to the C++ technology because there was the perception that the HEP computing model was a very good match to an object-oriented design.
BaBar was the first HEP experiment to effectively use geographically distributed resources, because the amount of computing needed to satisfy the production and analysis requirements exceeded what was possible at SLAC. Grid computing tools became available when the experiment was already running and BaBar solved the problem by assigning specific production tasks and datasets to different computing centers. In the last period of Babar life it was possible to produce 20-30% of Monte Carlo data using Grid resources with specific software tools.

3. The Role of CNAF

Until the end of 2013, CNAF hosted the full (Detector and Monte Carlo) event sample in “microDST” format and a significant fraction of the skimmed data. In addition it hosted also the user areas for the physics studies. Dedicated front-end machines and pledged CPU resources were also available.
In 2014 and in the following years the amount of CNAF resources used by the BaBar experiment is expected to decrease with the reduced number of physical analysis in progress.

4. Conclusions

CNAF has played a central role in the computing of the BaBar experiment. Even now, six year after the end of the data taking, it still provides services to support the on-going studies.

The Belle II Experiment at the INFN CNAF Tier1

F. Bianchi
INFN and University of Torino, via Giuria 1, 10135 torino, Italy

E-mail: fabrizio.bianchi@to.infn.it

Abstract. The Belle II experiment will collect data at the very high luminosity SuperKEKB asymmetric e^+ e^- collider, currently under construction in the KEK laboratory in Tsukuba, Japan. Its computing model is described with a special focus on the role of CNAF as Tier1 center.

1. Introduction

The BaBar and Belle experiments at the energy-asymmetric $e^+ e^-$ B factories PEPII and KEKB have observed CP violation in the neutral B meson system. The result was in good agreement with the predictions of the model of CP violation described by the Cabibbo-Kobayashi-Maskawa (CKM) matrix. We are now confident that the CKM phase is the dominant source of CP violation. The parameters of the Unitarity Triangle have been determined with a precision of O(10)% or better. Various other quantities in B meson decays have also become accessible in the era of the B factories. In particular, the first observation of direct CP violation in charmless B decays has been obtained. Over the past thirty years, the success of the Standard Model, which incorporates the CKM mechanism, has become increasingly firm. This strongly indicates that the Standard Model is the effective low-energy description of Nature. Yet there are several reasons to believe that physics beyond the Standard Model should exist.
Direct production of new particles at the LHC, or at another high-energy frontier collider, will be a distinctive signature of physics beyond the Standard Model. A complementary approach is to search for deviations from the Standard Model in flavor physics, and more importantly, to distinguish between different new physics models by a close examination of the flavor structure in the clean $e^+ e^-$ environment of an high luminosity asymmetric $e^+ e^-$ collider.
These are the primary motivation for the Belle II experiment that will collect data at the SuperKEKB collider currently under construction in the KEK laboratory in Tsukuba, Japan.

2. Computing Model

The Belle II computing model will be a distributed one and will use hardware resources in data centers located in the different participating countries.
Raw data coming from the detector will be permanently stored on tape at KEK and a full second copy will be stored in a different site. Sites hosting the raw data will have also the responsibility of processing them immediately after the data taking and of reprocessing them when a reprocess becomes necessary because of an update of the reconstruction software and/or of the detector constants.

The reconstructed data, the so-called mini-DST, will be stored in multiple copies in different sites to facilitate user access for physics analysis.
Monte Carlo events will be generated and reconstructed with the same code used for the detector data and also stored in multiple copies in different sites.
Data centers will have well-defined responsibilities that can coexist in the same physical site:

1. Raw Data Centers where the raw data will be stored and processed. "Raw Data Centers" will also serve as "Regional Data Centers" and "MC Production Sites".
    a. KEK Data Center: KEK is the host laboratory where the raw data will be recorded from the experiment and processed.
    b. PNNL Data Center: where the second copy of the raw data will be stored and (re)processed in parallel with KEK. If PNNL will not be able to store and reprocess the full second copy of the raw data, we plan to share this responsibility with GridKa and CNAF.
2. Regional Data Centers (GridKa, CNAF, etc.) where a copy of the mini-DST data will be stored. "Regional Data Centers" also serve as "MC Production Site".
3. MC Production Sites where a fraction of the MC production/reconstruction and physics analysis will be performed. All the other computing sites belong to this category and can be classified into three types according to used technology:
    a. GRID sites that operate with a standard GRID middleware (e.g. EMI, OSG).
    b. Cloud sites that operate with a standard Cloud infrastructure.
    c. Computing cluster sites: These sites are a standalone computer cluster which is accessible with the ssh protocol from the internet and available through a batch system such as LSF or TORQUE.
4. Local computing resources in institutes and universities, which will be used for ntuple-level user analysis.

For the time being the LCG grid technology is used for data storage, access, and transfer and for job handling. Work is in progress to implement the capability of using also sites with a Cloud infrastructure.

3. The Role of CNAF

INFN has officially joined the Belle II collaboration only in July 2013. Since then the Belle virtual organisation has already access to data centers in CNAF, Frascati, Legnaro, Napoli, Pisa, and Torino.
The INFN contribution to Belle II computing will consist of pledged resources located at CNAF and in Napoli (taking advantage of the ReCaS infrastructure) and eventually in other sites. CNAF will have the role of Regional Data Center while the other sites will be MC Production sites.
Hosting micro-DST format, detector and Monte Carlo data and ntuple-level user data at CNAF is already planned. In addition the possibility of having at CNAF also a fraction of the raw data and of the reprocessing is under discussion and a decision will be taken by the end of 2014.

4. Conclusions

CNAF will play a major role in the INFN contribution to the computing Belle II experiment. Even now, at this early stage of the experiment, a significant fraction of the Monte Carlo production is successfully performed here and in other INFN sites.

The Borexino experiment: recent scientific results and CNAF resources usage

The Borexino Collaboration

Abstract. Since 2012 the Borexino experiment, aimed at studying the solar neutrinos at LNGS, is supported by CNAF computing center for data storage, reprocessing and backup, interactive sessions and batch jobs . Details about resources usage and recent scientific results are given.

Borexino has been running since May 2007 at the LNGS with the primary goal of studying solar neutrinos [1]. The detector is a large, unsegmented liquid the scintillator calorimeter characterized by unprecedented low levels of intrinsic radioactivity and it is optimized for the study of the lower energy part of the solar neutrino spectrum.

The conceptual design is based on the principle of graded shielding with set of concentric shells of increasing radio-purity moving inwards surrounds the inner scintillator core. The core is made of ∼280 ton of scintillator, contained in a nylon Inner Vessel (IV) with a radius of 4.25 m and shielded from external radiation by 890 ton of inactive buffer fluid. Both the active and inactive layers are contained in a 13.7 m diameter Stainless Steel Sphere (SSS) equipped with 2212 8" PMTs (Inner Detector). A cylindrical dome with diameter of 18 m and height of 16.9 m encloses the SSS. It is filled with 2.4 kton of ultra-pure water viewed by 208 PMT's defining the Outer Detector. The external water serves both as a passive shield against external background sources, mostly neutrons and gammas, and also as an active Cherenkov veto system tagging the residual cosmic muons crossing the detector.

During the Phase-I (2007-2010) Borexino first detected [2, 3] and then precisely measured the flux of the ^7Be solar neutrinos [4], ruled out any signicant day-night asymmetry of their interaction rate [5], made the first direct observation of the *pep* neutrinos [6], and set the tightest upper limit on the flux of CNO neutrinos [6].

In 2010-2011 six purification campaigns were performed to further improve the detector performances and in October 2011, the Phase 2 data taking period was started.

Borexino is now in a mature operational condition, characterized on one hand by a very smooth and stable data taking and on the other by a refined and comprehensive analysis effort on the accumulated data.

Since 2012, data storage, reprocessing, simulation and analyses has been moved from a Gran Sasso devoted mini-cluster to INFN CNAF distributed calculus facility. Borexino data are classified into three types: raw-data, root files and dst's. Raw-data are compressed binary files with a typical size of ∼ 800 Mb, corresponding to a data taking time of ∼ 6h. Root files are reconstructed events files each organized in a number of Root TTrees. Their typical dimension is ∼ 2 GByte. Dst's contain only selected events for high level analyses. The whole statistics is reprocessed several times so the experiment needs a disk space increase of ∼ 10 Tb/year for data only. A similar disk space is required for the simulations output. CPU's usage is needed for

root-files production, whole statistics reprocessing, interactive and batch jobs and simulations. Peak usage up to $\sim$ 500 cnodes for a few week per year are necessary for a fast reprocessing while in normal conditions the possibility to run 150 parallel processes is adequate.

Among the most important scientific results obtained by Borexino after the migration to CNAF we remind:

- the release of the new geo-neutrino flux measurement [9], corresponding to a doubled statistics and reporting the first attempt to disentangle the individual U and Th contributions and the mantle signal;
- the identification of the annual modulation signature in the ^7Be neutrino flux due to the yearly variation of the Earth-Sun distance [10];
- new limits on heavy sterile neutrino mixing in the ^{8}B decay in the Sun [11];
- the measurement of cosmogenic backgrounds and muon-induced neutrons production in the Borexino detector [12].

Borexino continues in a rich solar neutrino program, including two even more challenging targets: *pp* and possibly CNO neutrinos. In parallel, the Borexino detector will be used in the SOX project, a short baseline experiment, aiming at investigation of the sterile-neutrino hypothesis [13].

References

[1] G. Alimonti et al., "A large-scale low-background liquid scintillation detector: The counting test facility at Gran Sasso," *Nuclear Instruments and Methods in Physics Research Section A*, vol. 406, pp. 411-426, 1998.

[2] C. Arpesella et al., (Borexino Collaboration), "First real time detection of ^{7}Be solar neutrinos by Borexino," *Physics Letters B*, vol. 658, pp. 101-108, 2008.

[3] C. Arpesella et al., (Borexino Collaboration), "Direct Measurement of the ^{7}Be Solar Neutrino Flux with 192 Days of Borexino Data," *Physical Review Letters* , vol. 101, Article ID 091302, 6 pages, 2008.

[4] G. Bellini et al., (Borexino Collaboration), "Precision Measurement of the ^{7}Be Solar Neutrino Interaction Rate in Borexino," *Physical Review Letters*, vol. 107, no. 14, Article ID 141302, 5 pages, 2011.

[5] G. Bellini et al., (Borexino Collaboration), "Absence of a day-night asymmetry in the ^{7}Be solar neutrino rate in Borexino," *Physics Letters B*, vol. 707, no. 1, pp. 22-26 , 2012.

[6] G. Bellini et al., (Borexino Collaboration), "First Evidence of pep Solar Neutrinos by Direct Detection in Borexino," *Physical Review Letters*, vol. 108, Article ID 051302, 6 pages, 2012.

[7] G. Bellini et al., (Borexino Collaboration), "Measurement of the solar ^{8}B neutrino rate with a liquid scintillator target and 3 MeV energy threshold in the Borexino detector," *Physical Review D*, vol. 82, Article ID 033006, 10 pages, 2010.

[8] G. Bellini et al., (Borexino Collaboration), "Observation of geo-neutrinos," *Physics Letters B*, vol. 687, pp. 299-304, 2010.

[9] G. Bellini et al., (Borexino Collaboration), "Measurement of geo-neutrinos from 1353 days of Borexino," *Physics Letters B*, vol. 722, pp. 295-300, 2013.

[10] G. Testera (Borexino Collaboration), Talk at the Neutrino Telescopes 2013 Conference, Venice, Italy.

[11] G. Bellini et al., (Borexino Collaboration), "New limits on heavy sterile neutrino mixing in 8B decay obtained with the Borexino detector", Physical Review, vol. D 88, Article ID 072010, 2013.

[12] G. Bellini et al., (Borexino Collaboration), "Cosmogenic Backgrounds in Borexino at 3800 m water-equivalent depth", Journal of Cosmology and Astroparticle Physics, vol. 2031, p. 049 (2013).

[13] G. Bellini et al., (Borexino Collaboration), "SOX:Short distance neutrino Oscillations with Borexino," arXiv:1304.7721 accepted for publication by Journal of High Energy Physics.

CDF computing at CNAF

Silvia Amerio

University of Padova and INFN, via Marzolo 8, 35131 Padova (Italy)

E-mail: silvia.amerio@pd.infn.it

Abstract. CNAF has played a key role in CDF computing model and is now contributing to the long term future preservation of CDF data and analysis capabilities. In this report, after a brief introduction on CDF most recent physics results, we will describe CDF computing model, the services currently hosted at CNAF and the long term future data preservation project under development.

1. The CDF experiment and recent results

The CDF experiment is a high-energy physics experiment which took data between 1986 and 2011. It detected and studied collisions of protons and anti-protons accelerated up to an energy of 2 TeV by the Tevatron accelerator complex, located at Fermilab (Batavia, US). The discovery of top quark in 1995 [1], the observation of $B_S^0 - \overline{B_S^0}$ oscillations [2] and of single top [3], are only few of the fundamental results obtained by the experiment.

CDF ended its data taking in 2011. More than 70 papers have been published since then, and many analysis are still ongoing on its $10 fb^{-1}$ data sample. In the top sector, new measurements of $t\bar{t}$ forward-backward asymmetry [4] and of single top cross sections on the full available statistics were published in 2013. A Tevatron top mass combination was also published and was recently combined with ATLAS and CMS results [5]. In the B sector, recent results include the measurements of new resonances [6] and the observation of $D^0 - \overline{D^0}$ mixing [7]. The energy dependence of minimum-bias and underlying event in $p - \overline{p}$ collision is being studied thanks to unique samples of data collected at different center-of-mass energies. In the electroweak sector the most recent published papers include full statistics results on $sin^2\theta_W$ measurements [8] and on ZZ cross section [9].

2. CDF Computing Model

During its operations, CDF computing architecture evolved from the initial dedicated farms to using Grid-based resources. Cu rrent CDF computing model is based on a central farm located in Fermilab and accessed through the *CDFGrid* portal, and Open Science Grid (OSG) and LHC Computing Grid (LCG) resources accessed through the dedicated portals *NamGrid* and *Eurogrid* respectively. An experiment specific package, the Central Analysis Farm (CAF) software, provides the users with a uniform interface to resources on the different Grid sites [10]. The three portals are based on glideinWMS [11]. Authentication is based on Kerberos [12]. Total collision and simulation data amount to about 10 PB, of which 4 PB are raw and ntuple-level data. Data handling is based on SAM (Sequential data Access via Metadata) [13] and dCache [14]. CDF reconstruction and analysis code is written in C, C++ and Python languages

and is preserved in frozen releases in CVS repositories. The latest version of CDF code runs on SL5 operating system; a SL6 legacy release is being prepared and will be ready by mid 2014.

3. CDF computing at CNAF

CNAF has been one of the major contributors to CDF computing outside Fermilab in the past and it now maintains a leading role in the data preservation effort. CDF has dedicated resources at CNAF: 8000 HS06 of computing power, 400 TB of disk and a set of machines for data access and analysis services.

3.1. Eurogrid portal

Eurogrid portal allows CDF users to access a dedicated farm at CNAF Tier-1 and additional LCG computing resources at different Tier-2 sites in Italy, Spain, Germany and France. As other CDF portals, Eurogrid uses a pilot based workload management system, the *glideinWMS*. GlideinWMS comprises two main elements: the *Factory* and the *Frontend*; the first knows the details of various Grid sites and properly configures the entry points for the pilot jobs (glideins); additionally, the Factory also does the submissions. The Frontend is the user inteface to the glideinWMS: it looks for user jobs and asks the Factory to provide glideins, if needed; the Frontend knows nothing about the glideins or Grid sites; it only has to match user job requirements to the sites attributes published by the Factory. In Eurogrid, CDF frontend interacts with the Factory at University of California San Diego (UCSD). Eurogrid is official since June 2011. Since the beginning, users response has been very good: during the first two years of usage on average 400 jobs/day were running on the portal, and peaks of 3000 jobs have been reached during the most intense periods of data analysis. While CDF data taking was still ongoing, about 40 CDF users regularly submitted their jobs on Eurogrid, on average 5 users/day. Eurogrid is the most used job submission portal outside Fermilab. Eurogrid has demonstrated to be very reliable: during its first years of operations, it has experienced a negligible rate of failures due to site related errors. Recently the load on the portal has greatly decreased due to the shrinking of the collaboration; nevertheless, as can be seen in fig. 1, the activity was still high up to February 2014.

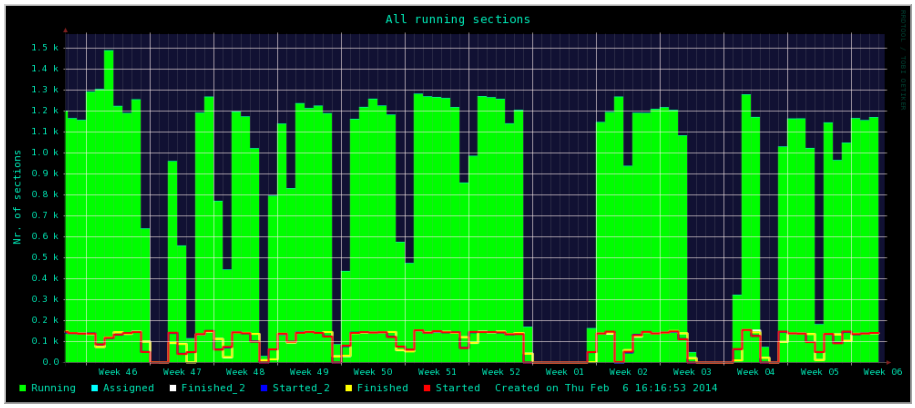


Figure 1. CDF jobs running on Eurogrid portal in the period November 2013 - February 2014

Besides the Eurogrid portal and dedicated working nodes to run CDF jobs, other services hosted at CNAF are a SAM station to access CDF data stored at CNAF, and a disk area to store user's job outputs, accessible via a dedicated machine. All these services are implemented on virtual machines and will be part of the long term future analysis framework under development.

3.2. Long Term Data Preservation

A project for the long term future preservation of CDF data is being implemented at CNAF computing center, in collaboration with CDF experiment and within the DPHEP collaboration. This is the first project funded by INFN on long term data preservation and will serve as a prototype for other experiments hosting their data at CNAF and other INFN sites. The goal is to copy all CDF raw and user-level data files (4 PB) from Fermilab to CNAF storage system, and provide users with tools to access and analyze it in the long term future. A mechanism able to copy the data at 5 Gb/s rate and store it in CNAF tape library has been setup in the second half of 2013 and successfully tested. A dedicated 10Gb/s link and a reserved network allow to manage and monitor CDF data movement independently from CNAF Tier 1 network resources and to have a secure high speed channel always available for data transfers, with no sharing of any resource. The storage layout consists of a pool of disks managed by GPFS, a tape library infrastructure for the archive back-end and an integration system to transfer data from disk to tape and vice versa. The CNAF Tier 1 storage solution is GEMSS [15], an integration of GPFS, TSM and StoRM. A single 10 Gb/s GridFtp Server, connected directly through the Storage Area Network (SAN) to the CDF GPFS file system disks and to the CNAF Tier 1 network switch 10 Gb backbone is used for the data copy. This allows a plain method for transferring data from Fermilab to CNAF through a single point. In fig. 2 shows the data transfer rate during the first months.

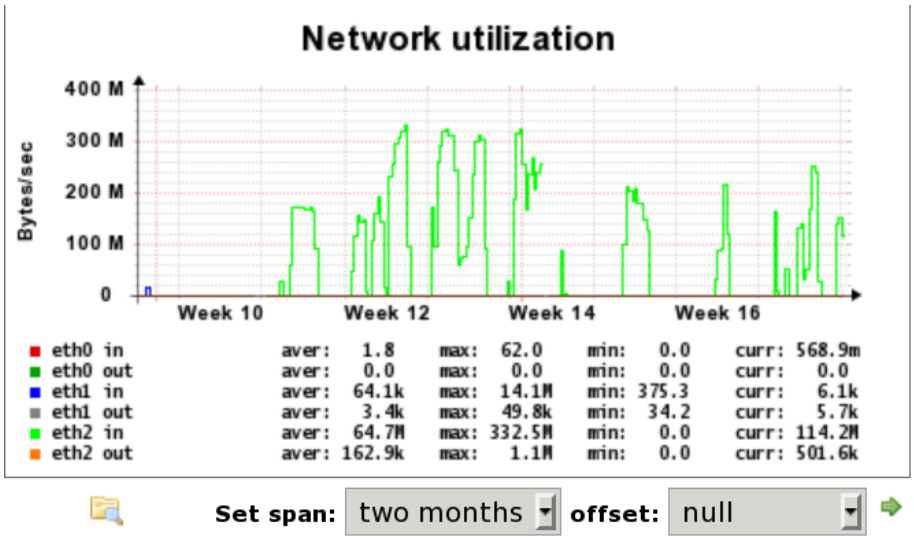


Figure 2. Data transfer rate from Fermilab to CNAF for the CDF long term data preservation project.

For the analysis, an infrastructure based on virtualization is being developed, allowing the CDF analysis services to be instantiated on demand in a controlled environment. Services used to access CDF data will be eventually migrated to a dynamic virtual infrastructure. This infrastructure will be implemented so that CDF services can be instantiated on-demand on pre-packaged virtual machines in a controlled environment, where in- and out-bound access to these services and connection to storage data is administratively controlled. The set-up will be such that, when authorized access to CDF data is requested, instantiation of the virtual services will happen automatically and the virtual machines will be placed into a suitably isolated network infrastructure.

3.3. Conclusions

During CDF RunII (2001-2011) operations CNAF has been one of the major computing centers for the experiment. A portal (*Eurogrid*) to access CNAF Tier-1 and other LCG resources is hosted at CNAF, together with dedicated data processing and storage resources. *Eurogrid* has been the most used job submission portal outside Fermilab in the last years. Now, two years after the end of data taking, the analysis activity is facing a natural decrease. CDF is entering the data preservation phase and CNAF will playing a fundamental role in it. A long term data preservation project is being implemented at CNAF: a complete copy of all CDF data will be available at CNAF, together with all the necessary services to access and analyze it. Currently more that 1PB of CDF data is already duplicated at CNAF, and a long term future analysis framework is being setup. The plan is to complete the copy (4PB) and have a first prototype of the analysis framework by the end of 2014.

References

[1] Abe F *et al.* (CDF Collaboration) 1995 *Phys.Rev.Lett.* **74** 2626–2631 (*Preprint* `hep-ex/9503002`)
[2] Abulencia A *et al.* (CDF Collaboration) 2006 *Phys.Rev.Lett.* **97** 242003 (*Preprint* `hep-ex/0609040`)
[3] Aaltonen T *et al.* (CDF Collaboration) 2009 *Phys.Rev.Lett.* **103** 092002 (*Preprint* `0903.0885`)
[4] Aaltonen T A *et al.* (CDF Collaboration) 2014 (*Preprint* `1404.3698`)
[5] ATLAS, CDF, CMS and D0 Collaborations (*Preprint* `1404.3698`)
[6] Aaltonen T A *et al.* (CDF Collaboration) 2013 *PhyS.Rev.D* (*Preprint* `1309.5961`)
[7] Aaltonen T A *et al.* (CDF Collaboration) 2013 *Phys.Rev.Lett.* **111** 231802 (*Preprint* `1309.4078`)
[8] Aaltonen T A *et al.* (CDF Collaboration) 2014 *Phys.Rev.* **D89** 072005 (*Preprint* `1402.2239`)
[9] Aaltonen T A *et al.* (CDF Collaboration) 2014 *Phys.Rev.D* (*Preprint* `1403.2300`)
[10] Lucchesi D (CDF Collaboration) 2010 *J.Phys.Conf.Ser.* **219** 062017
[11] Sfiligoi I, Wurthwein F, Andrews W, Dost J, MacNeill I *et al.* 2011 *J.Phys.Conf.Ser.* **331** 072031
[12] Kerberos: The Network Authentication Protocol URL `http://web.mit.edu/kerberos/`
[13] Stonjek S, Baranovski A, Kreymer A, Lueking L, Ratnikov F *et al.* 2005 1052–1054
[14] dCache group URL `http://www.dcache.org/`
[15] Ricci P P, Bonacorsi D, Cavalli A, Dell'Agnello L, Gregori D *et al.* 2012 *J.Phys.Conf.Ser.* **396** 042051

The CMS Experiment at the INFN CNAF Tier1

T. Boccali
INFN Sezione di Pisa, L.go B.Pontecorvo 3, 56127 Pisa, Italy

E-mail: Tommaso.Boccali@cern.ch

Abstract. A brief description of the CMS Experiment is given, with particular focus on the computing aspects. The setup for CMS at the CNAF Tier1 centre is shown, highlighting the peculiar points with respect to the other sites. New developments and expected resource growth are also presented.

1. Introduction

The CMS Experiment at CERN collects and analyses data from the pp collisions in the LHC Collider. The first physics Run, at centre of mass energy of 7 -8TeV, started in late March 2010, and ended in February 2013; more than 25 fb-1 of collisions were collected during the Run.
The CMS Experiment is designed as a general purpose detector, and hence is interested in a huge list of physics subjects; however, given the new energy regime the LHC can probe, the main expectations were on one side on the completion of the Standard Model, with the discovery of a Higgs-like boson, on the other side on the discovery if physics beyond the Standard Model, where multiple models were to be probed (Super-symmetry in all the possible incarnations, Extra dimensions, and all the sorts of more exotic models).
More than 300 physics papers were produced from Run 1 data, including the now renowned paper on the Observation of a 126 GeV Higgs Boson, which sets the final cornerstone to the Standard Model.

2. The CMS Computing Model

CMS trigger rates, exceeding 1 kHz in the last months of the Run (less than 500 Hz averaged on the Run, though), combined with large event sizes and computational needs, have a big impact on CMS Computing Model. CMS uses a derivative of the MONARC Hierarchical Model, based on GRID Middleware, where a Tier0, 7 Tier1 and roughly 50 Tier2 sites share the computational load. One of the Tier1s resides at CNAF, in Bologna, Italy.
The CNAF Tier1 has been used during Run1 to fulfil a series of tasks:

- Custody of a fraction of the raw and processed data and simulation,
- Simulation of the Monte Carlo events needed for analyses,
- Processing and reprocessing of both data and simulated events.

The resources CMS has deployed at CNAF constitute the 13% of the total Tier1 resources, the fraction being equal to the fraction of the Italian component in CMS; they amount (2013 numbers) to

- 22.75 kHS06 computational power;
- 6500 TB of tape;
- 3400 TB of disk frontend.

Due to the very specific nature of CNAF, which serves all the LHC Collaborations and other less demanding experiments, CMS has actually been able to use large CPU over pledges quite constantly over time, consistently resulting as the second Tier1 as number of processed hours after the US Tier1; the same holds for total number of processed jobs, as shown in Figure 1.
The tape resource has been used at levels exceeding 90%, resulting in CNAF as the Tier1 holding more custodial data, after the US Tier1. The disk resource has been used up to mid 2013 as a cache frontend to the tape, and thus maintained always full at 90%.

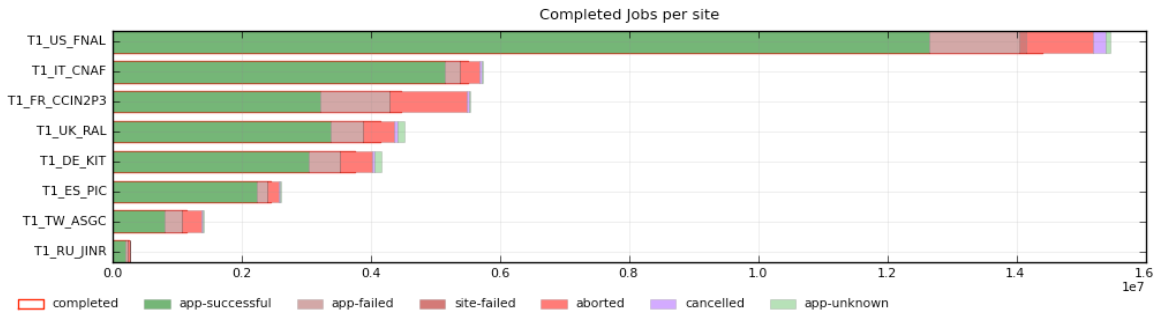


Figure 1. Jobs processed at CMS Tier1s during 2013.

The specific setup chosen at CNAF for CMS is unique among CMS Tier1 centres. CNAF is the only site that uses as storage technology Storm over GPFS, which on its turn offers a TSM tape backend. Storm is a lightweight storage component, which offers SRM (and Http) access layers, but not disk aggregation capabilities. The latter is instead delegated to a commercial GPFS installation, which encapsulates also TSM tape backend. The solution has proven as appropriate for CMS, and the Storm solution is being investigated or implemented at a number of CMS Tier2 sites.

Starting from 2013, the CMS storage setup has evolved at CNAF.
Access to the files has been granted from remote locations via the Xrootd. An initial installation has been realized in the first 6 months of the year, allowing for access only to those files already present on the disk backend; this has been deemed as necessary to protect the tape drivers from chaotic file recalls from remote analysis activities.
Later in the year, the overall storage system has undergone a deep transformation. The disk has been split into a smaller tape cache, and a proper disk area directly managed by the experiment. The Xrootd servers have been directed only to this latter resource, thus protecting by design the tape area.
The new setup for the CPU + the disk area reduces significantly the differences between a Tier1 and a Tier2; indeed, in late 2013 CNAF has opened the batch queues also for the standard analysis jobs, which is currently ramping from virtually zero to some % level (the limit being the fact that Monte Carlo simulation and reprocessing have higher priorities).

3. New developments

The new flexible setup has allowed in the last months interesting operating modes which on one side have allowed for greater reliability to hardware problems and scheduled interventions, on the other have allowed for tests relevant for the planning of next generation computing models.
A few examples are listed here:

- A full reprocessing of a 30 TB sized dataset has been performed on CNAF CPUs, reading data directly (in streaming) from FNAL;
- A full reprocessing of a similar dataset has been performed at RAL (UK) Tier1 reading raw data directly from CNAF Xrootd servers;

- Usually, when CNAF storage is down for scheduled maintenance, all local CMS activity are stopped. In principle, local analysis activities accessing remote data via Xrootd can still work. This was tested on a 4 day storage downtime, where more 5000 analysis jobs were running simultaneously at CNAF. The overall job CPU efficiency has been exceeding 80%, comparable with local access, and additional failures due to the remote operation mode have been at the % level if not less. This has resulted in a near saturation of the CNAF LHCOPN 20 Gbps line.

Furthermore CNAF provides experiments with testing environments for new technologies such as multi-core queues and many-cores systems, which CMS is actively using to evolve its computing and software frameworks.

4. Expected resource growth

The LHC collider is at the moment (March 2014) off for upgrade, and is expected to go back in operations in the first months of 2015, with a new centre of mass energy of 13 TeV. The new Run, called Run2, will last up to 2018, with an instantaneous luminosity roughly doubled with respect to Run1. CMS expects to carry on an extensive study of the Higgs boson properties during Run2, while performing searches for new physics at the newly available energy.

The former aspect needs an increased trigger, which should stably collect 1 kHz of more complex events. Present estimates require, even in presence of drastic optimizations, roughly a factor 2 in Tier1 CPU resources. For 2015, the only year currently under precise scrutiny, CPU requests are expected to ramp up by 70%, with a much smaller impact on Disk and Tape needs. These figures will be discussed in the coming months, but already predict a steep increase of resources. In the next years, though, the CMS Collaboration has agreed to live within the envelope of a stable Computing Budget, which by current estimates means a year-by-year increase of roughly 20%.

5. Conclusions

The CMS Collaboration expressed in many occasions its praise to the CMS CNAF Tier1, which has consistently been in the top 2 Tier1 sites for resource utilization, resource pledges and availability. CNAF represents an important asset for the CMS Collaboration, and all the expectations are towards an even greater role inside the CMS Computing.

The Cherenkov Telescope Array

Ciro Bigongiari
INAF - Osservatorio Astrofisico di Torino & INFN Torino,
Via Pietro Giuria 1, 10125 Torino, IT

E-mail: ciro.bigongiari@to.infn.it

Abstract.
The Cherenkov Telescope Array (CTA) is a project to build a new generation ground-based gamma-ray instrument operating in the energy range extending from some tens of GeV to above 100 TeV. It will serve as an open observatory to a wide astrophysics community and will provide a deep insight into the non-thermal high-energy universe.

1. Introduction

Radiation at gamma-ray energies cannot conceivably be generated by thermal emission from hot celestial objects. In a bottom-up fashion, gamma-rays can be generated when highly relativistic particles accelerated for example in the gigantic shock waves of stellar explosions collide with ambient gas, or interact with photons and magnetic fields. High-energy gamma-rays can also be produced in a top-down fashion by decays of heavy particles such as hypothetical dark matter particles or cosmic strings. Present Imaging Air Cherenkov Telescopes (IACTs, like H.E.S.S. MAGIC and VERITAS) have clearly demonstrated the capacity of the imaging technique discovering tens of cosmic sources of high energy gamma-rays. CTA intends to improve the flux sensitivity of the current generation of IACTs by an order of magnitude, lower the energy threshold, improve the angular resolution, and extend the energy coverage. Such improvements will be achieved with an array composed by tens of Cherenkov telescopes of different sizes: some very large telescopes which will achieve a few tens of GeV energy threshold, about 30 telescopes with size comparable to present IACTs which will increase the array effective area, improve the angular resolution and sensitivity, and finally about 60 small size telescopes which will extend the energy range upper limit to beyond 100 TeV. Actually the CTA consortium plans to operate from one site in the southern and one in the northern hemisphere, allowing full-sky coverage. The southern site will cover the central part of the galactic plane and see most of the galactic sources and will therefore be designed to have sensitivity over the full energy range. The northern site will be optimized for extragalactic astronomy, and will not require coverage of the highest energies. The better sensitivity and the lower energy threshold will allow the detection of many more sources and the study of short transient emissions while the better angular resolution will allow detailed studies of emission regions. Thanks to the extended energy range the emission spectrum of some sources could be measured over an unprecedented energy span providing strict constraints on acceleration mechanisms. Moreover the energy band above few tens of TeV will be explored for the first time. CTA will explore a wide variety of particle accelerators in the Universe, from the nearby pulsars, micro-quasars, stellar winds and supernova to active galactic nuclei, gamma-ray bursts and clusters of galaxies. High-energy gamma-rays can be used

moreover to trace the populations of high-energy particles, thus providing insightful information about the sources of high energy cosmic rays. The project is currently in its preparatory phase. The construction is planned to begin in 2016 and will be completed around 2019-2020. The current status of the project and its expected performance are described in a dedicated volume of the Astroparticle Physics journal [1].

2. Computing requirements

Presently the CTA consortium is testing and developing the technologies which will be used for the construction of the various classes of telescopes. Meanwhile detailed Monte Carlo simulation of the entire array are ongoing to estimate its performance and to optimize some parameters like the distance between telescopes or the trigger strategy. In particular, the selection of the CTA sites (North and South) has a significant impact on the final sensitivity of the instrument, as performance depends on the altitude of the site, its atmospheric conditions, geomagnetic field and night-sky background. The CTA MC working group is studying the impact of these various parameters by means of detailed MC simulations of the detector response to extensive air showers. Luckily the imaging air Cherenkov technique is very effective in rejecting background events due to charged cosmic ray particles. On the other hand this means that a huge amount of background events need to be simulated to achieve reliable estimates of the array sensitivity. About 10^{10} cosmic ray induced atmospheric showers for each site are needed to properly estimate the array performance, requiring extensive computing needs, disk space and CPU power. Currently the use of the EGI grid infrastructure and of the DIRAC (Distributed Infrastructure with Remote Agent Control) framework is adopted. The EGI CTA Virtual Organization (CTA VO) was created in 2008 and today is supported by 19 EGI grid sites spread in 7 countries, with resources of the order of 10k of available cores and more than 800 TB of dedicated storage. Within the CTA VO, we distinguish two main activities: MC production, centralized and performed by users associated to the production role, in charge of providing the CTA consortium with MC samples, and User analysis, performed potentially by any CTA user, which consists in processing the MC data samples. For the time being, the DIRAC framework is used to manage both the above mentioned activities of the CTA VO. The computing centers supporting the CTA VO are distinguished according two classes, i.e. analysis centers and production centers. The main difference between them is that while production centers provide only CPU, analysis centers are also providing support for long term storage of the data. With this difference in mind, analysis centers are employed for both MC production and analysis, while production centers are only used for MC production. The amount of CPU power and disk space needed for the user analysis is negligible with respect to the ones needed for MC production. Since MC datasets must be produced within delays of a few weeks, the CTA activity is not flat during the year. The peak value of the required computing power has thus been estimated considering 2 weeks-campaigns and assuming an average core performance = 10 HS06. On the basis of the above considerations a peak requirement of 6000 cores and an overall CPU power of 100 millions of HS06 hours have been estimated for 2014. The estimated disk space for 2014 is more or less the same used in 2013, about 900 TB. Due to a late agreement between the CNAF and the CTA VO and to some configuration problems the CNAF contribution to 2013 CTA needs has been quite limited, about 0.03 M HS06 hours, but the solution of the above mentioned problems allowed a quick growth of CNAF share. Actually in the first three months of 2014 the CNAF already provided about 2.3 M HS06 hours.

References

[1] Hinton J, Sarkar S, Torres D and Knapp J 2013 *Astroparticle Physics* **43** 1-356

The *Fermi*-LAT Grid Interface to CNAF

L Arrabito[1], J Bregeon[1], J Cohen-Tanugi[1], M Kuss[2], F Longo[3], F Piron[1], S Viscapi[1] and S Zimmer[4], on behalf of the *Fermi* LAT collaboration

[1] Laboratoire Univers et Particules, Université de Montpellier II Place Eugène Bataillon - CC 72, CNRS/IN2P3, F-34095 Montpellier, France
[2] Istituto Nazionale di Fisica Nucleare, Sezione di Pisa, I-56127 Pisa, Italy
[3] Department of Physics, University of Trieste, via Valerio 2, Trieste and INFN, Sezione di Trieste, via Valerio 2, Trieste, Italy
[4] The Oskar Klein Centre for Cosmoparticle Physics and Department of Physics, Stockholm University, AlbaNova, SE 106 91, Stockholm, Sweden

E-mail: `francesco.longo@ts.infn.it`

Abstract. The *Fermi* Large Area Telescope current generation experiment dedicated to gamma-ray astrophysics is massively using the CNAF resources to run its Monte-Carlo simulations through the Fermi-DIRAC interface on the grid under the virtual organization glast.org.

1. The *Fermi* LAT Experiment

The Large Area Telescope (LAT) is the primary instrument on the *Fermi Gamma-ray Space Telescope* mission, launched on June 11, 2008. It is the product of an international collaboration between DOE, NASA and academic US institutions as well as international partners in France, Italy, Japan and Sweden. The LAT is a pair-conversion detector of high-energy gamma rays covering the energy range from 20 MeV to more than 300 GeV [1]. It has been designed to achieve a good position resolution (<10 arcmin) and an energy resolution of ∼10 %. Thanks to its wide field of view (∼2.4 sr at 1 GeV), the LAT has been routinely monitoring the gamma-ray sky and has shed light on the extreme, non-thermal Universe. This includes gamma-ray sources such as active galactic nuclei, gamma-ray bursts, galactic pulsars and their environment, supernova remnants, solar flares, etc.. A brief and recent review of *Fermi*-LAT discoveries can be found in [2].

2. *Fermi* LAT Performance

The LAT response to gamma rays is parametrized by the so-called "instrument response functions" (IRFs), which together with the data from the instrument are provided to the scientific community[1]. As described in [3], IRFs are derived using Monte-Carlo (MC) simulations and are also corrected for discrepancies observed between flight and simulated data, as the LAT team gains insight into the in-flight performance of the instrument. In the near future, major improvements are expected from the new "Pass 8" data, such as an increased effective area with

[1] Data release and software maintenance is done via the Fermi Science Support Center http://fermi.gsfc.nasa.gov.

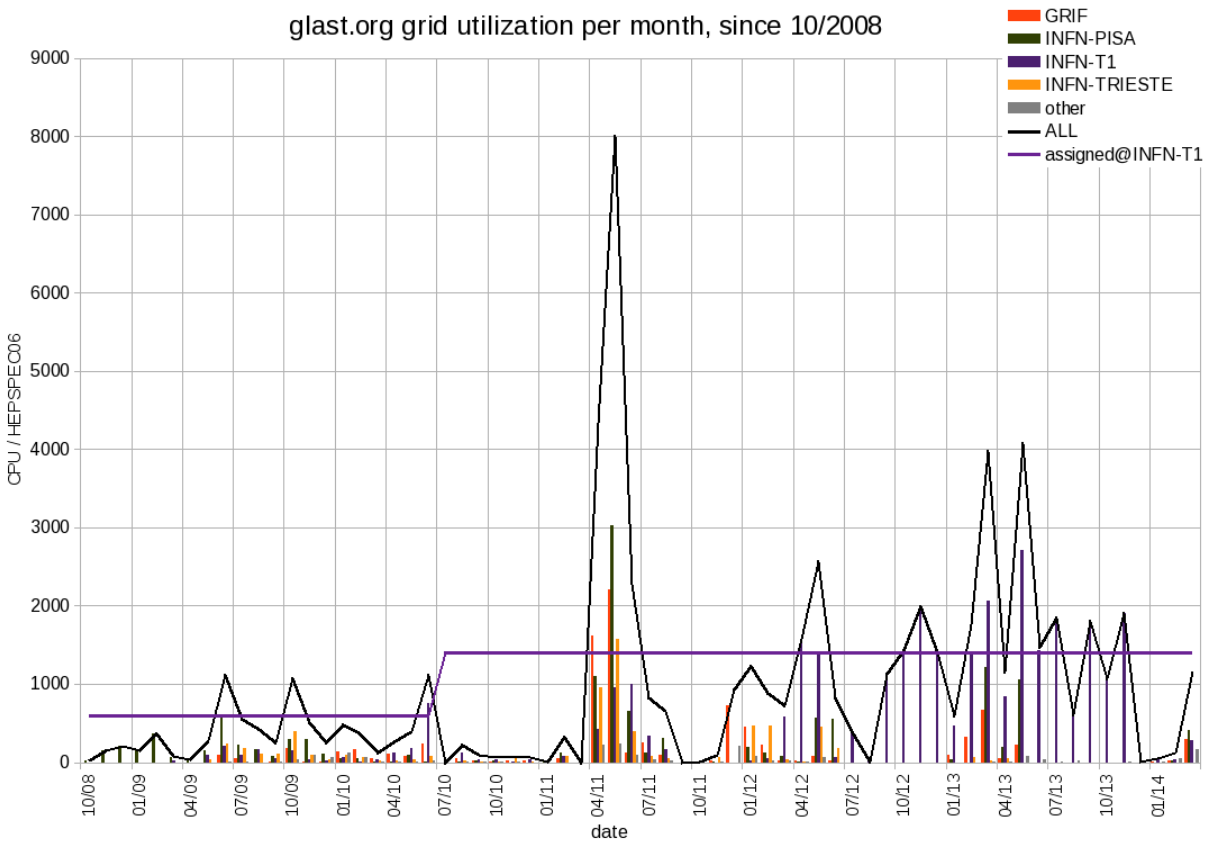


Figure 1. Usage of grid sites by the VO glast.org since its creation

respect to the current "Pass 7" public data [4]. These improvements correspond to a radical revision of the LAT event-level analysis. The optimization of the event reconstruction and of the background rejection, and the full characterization of the new IRFs, require the production of large simulated data sets including gamma rays and charged cosmic backgrounds (protons, heavy ions, electrons). These simulations are also fundamental for high-level analyses which require a proper evaluation of the residual backgrounds (e.g., the extragalactic diffuse emission [5] or the cosmic electron-positron spectra [6]).

3. The *Fermi* LAT Virtual Organization

The *Fermi* LAT virtual organization (VO) glast.org was created in 2008 to start using distributed resources mainly at INFN sites, CNAF, Pisa and Trieste, but then over the years we got activated at several more sites. We ported the *Fermi* LAT MC and data analysis framework Gleam [7], part of a larger collection of packages called GlastRelease, to the grid. We installed various GlastRelease versions on the grid and performed several campaigns to produce specific MC data sets. This allowed the *Fermi* LAT collaboration to acquire grid usage experience and to prove the possibility to run MC simulation over distributed resources. Figure 1 shows the usage of grid resources since the creation of the virtual organization. The majority of the jobs were run at the INFN Tier 1 at CNAF as shown by Fig. 2.

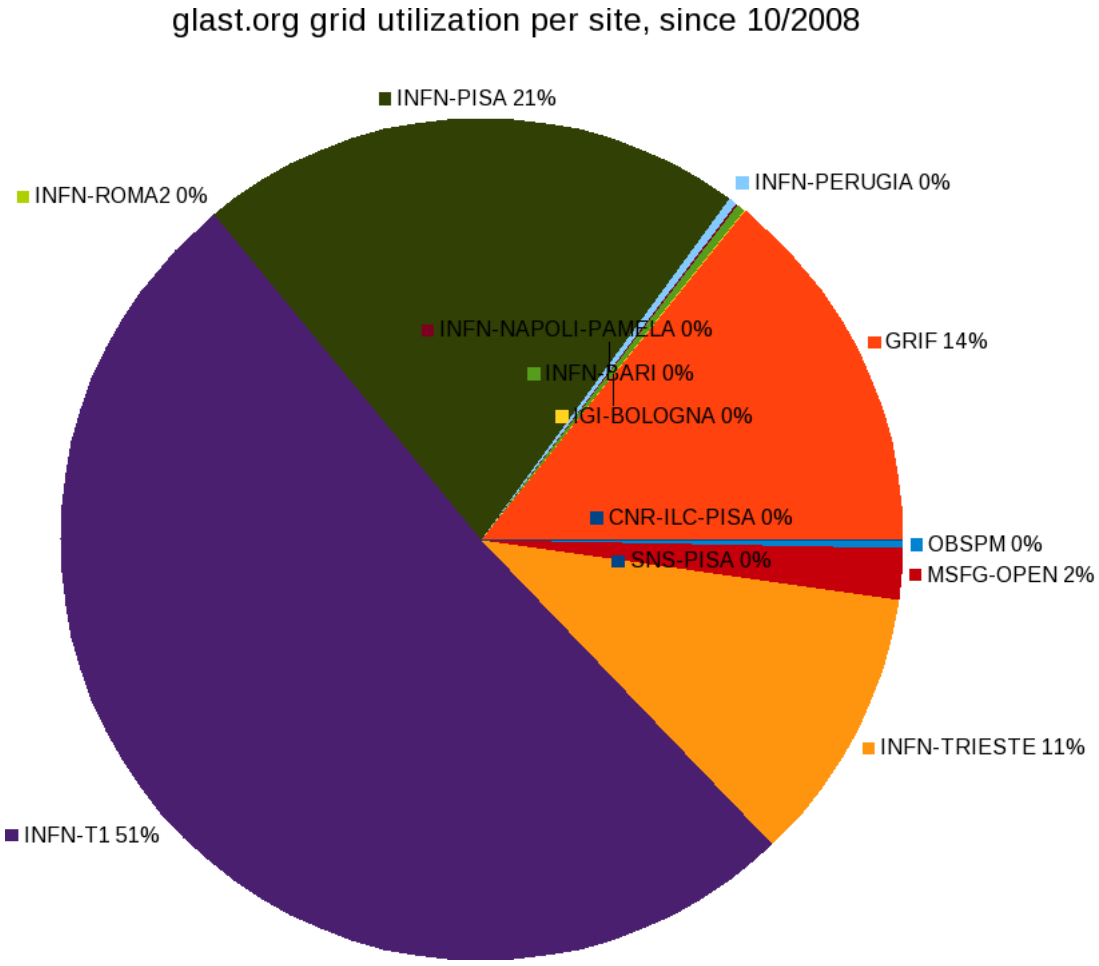


Figure 2. Cumulative usage of grid sites by the glast.org VO

4. *Fermi* LAT Monte-Carlo Production Runs on the Grid

The *Fermi*-LAT offline processing system is hosted by the LAT ISOC (Instrument Science Operations Center) based at the SLAC National Accelerator Laboratory in California. The *Fermi*-LAT data processing pipeline (e.g. see [8] for a detailed technical description) was designed with the focus on allowing the management of arbitrarily complex work flows and handling multiple tasks simultaneously (e.g., prompt data processing, data reprocessing, MC production, and science analysis).

Since the *Fermi* mission has recently entered (after 5 years of operations) into its extended phase, we are planning to move the bulk of MC production to grid resources.

To move to a more productive procedure to generate large MC datasets, and to fulfil present and future requirements for massive MC simulations and data analysis, we started to use grid resources under the VO glast.org through the recently developed DIRAC (Distributed Infrastructure with Remote Agent Control) [9] interface to the LAT data pipeline [10].

We decided to ramp up with DIRAC to exploit the full potential of the grid because for the need of larger MC production campaigns in the near future.

A first working prototype of the grid/pipeline interface with DIRAC was completed in summer 2013. The first MC production on the grid using this interface consisted of 4000 streams, each one simulating 5k gamma-ray events [11]. This first test was limited to four EGI grid sites supporting glast.org: CNAF, MSFG, OBSPM and BARI (sorted by resource usage). They

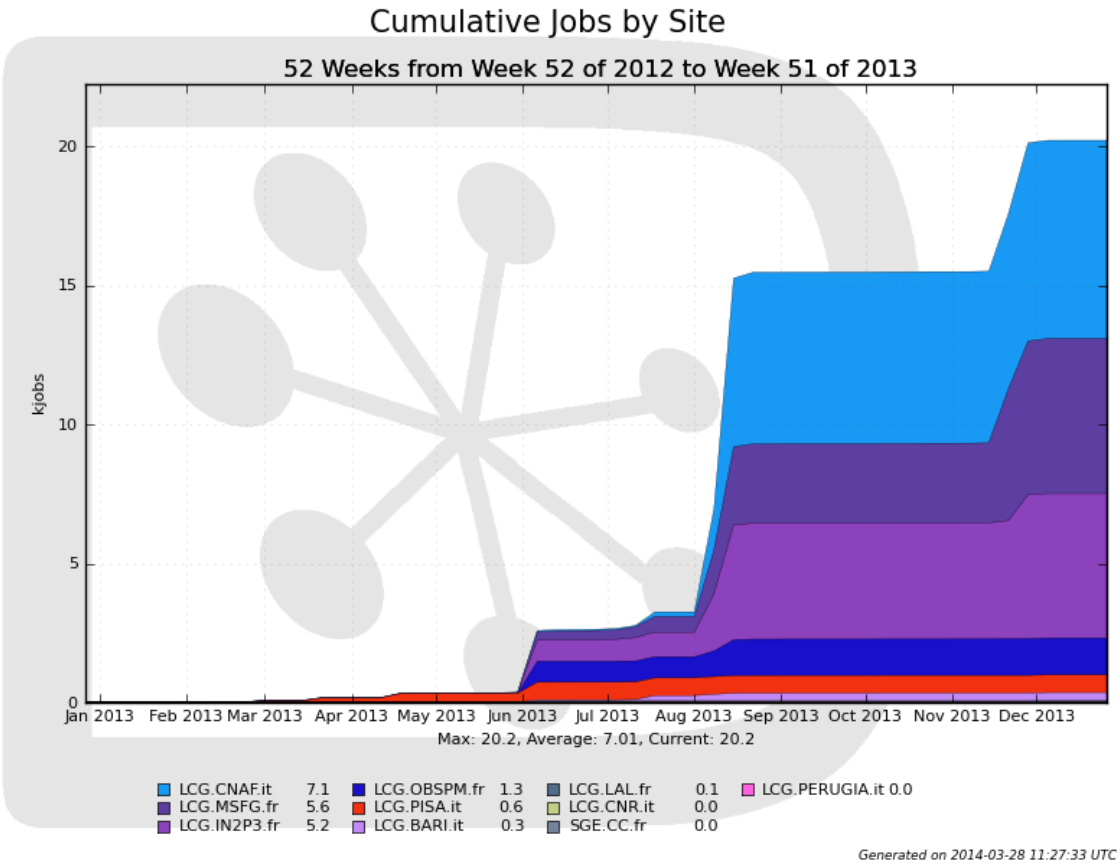


Figure 3. Jobs at the various sites launched using the DIRAC interface

represent about one third of the VO's resources. The production used between 100 and 300 cores simultaneously, and it ran smoothly with a 97 % success rate.

Since the first test with which we started to use DIRAC, the number of jobs ran at CNAF increased persistently. Figure 3 shows the current number of jobs that we ran during 2013 on the various sites, showing that CNAF is the largest resource used so far.

5. Conclusions and Perspectives

In order to exploit the grid resources for massive MC simulations and data analysis, we have deployed a prototype setup, based on the DIRAC framework. The whole production chain of the *Fermi*-DIRAC setup has been extensively tested, confirming that the DIRAC solution fulfils all the requirements imposed by the *Fermi*-LAT pipeline. In the medium-term, we also plan to learn how the overall system behaves under stress through scalability tests, and to optimize the resource usage before entering production mode in view of the massive simulation of "Pass 8" backgrounds in fall 2014.

References

[1] Atwood W B et al. 2009 *The Astrophysical Journal* **697** 1071
[2] Thompson D J 2013 arXiv:1308.1870
[3] Ackermann M et al. 2012 *The Astrophysical Journal Supplement Series* **203** 4
[4] Atwood W B et al. 2013 *2012 Fermi Symposium: eConf Proceedings C121028* arXiv:1303.3514
[5] Abdo A A et al. 2010 *Physical Review Letters* **104** 101101
[6] Ackermann M et al. 2012 *Physical Review Letters* **108** 011103

[7] Boinee P et al 2003 *Proceedings of the first workshop on Science with the new generation of high energy gamma-ray experiments : between astrophysics and astroparticle physics.*. Edited by S. Ciprini, A. De Angelis, P. Lubrano, and O. Mansutti, 141
[8] Dubois R 2009 *ASP Conference Series* **411** 189
[9] Tsaregorodtsev A et al. 2008 *Journal of Physics: Conference Series* **119** 062048
[10] Zimmer S et al. 2012 *Journal of Physics: Conference Series* **396** 032121
[11] Arrabito L et al. 2013 CHEP 2013 conference proceedings arXiv:1403.7221

The ICARUS Experiment

M. Antonello on behalf of the ICARUS Collaboration
INFN - Laboratori Nazionali del Gran Sasso, Assergi, Italy
E-mail: maddalena.antonello@lngs.infn.it

1. ICARUS experiment scientific program

In 1977 C.Rubbia [1] conceived the idea of the LAr-TPC (Liquid Argon Time Projection Chamber). The ICARUS T600 cryogenic detector is the largest LAr-TPC ever built and operated. Installed in the Hall B of the Gran Sasso underground laboratory and exposed to the CNGS neutrino beam, on June 26^{th} 2013 ICARUS has completed 3 years of continuous data taking, collecting about 3000 CNGS neutrino events but also cosmic rays and other self triggered events, and showing optimal overall detector performance and stability. The ICARUS T600 addresses a wide physics program including atmospheric and solar neutrino interactions and also charged and neutral current neutrino interactions associated with the CNGS neutrino beam, focusing on neutrino oscillation studies in the $\nu_\mu \rightarrow \nu_e$, $\nu_\mu \rightarrow \nu_\tau$ and ν_μ disappearance channels. It can also search for rare and up to now unobserved events like the long sought for proton decay with zero background in one of its $3{\times}10^{32}$ nucleons (in particular into exotic channels).

1.1. Experimental results

Recently, ICARUS has published updated results on the search for anomalous "LSND-like" $\nu_\mu \rightarrow \nu_e$ transitions [2, 3]. The outcome allowed to strongly limit the window of open options for the LSND anomaly in a well defined parameter region centered around $(\Delta m^2, sin^2(2\theta)) = (0.5eV^2, 0.005)$ where there is an over-all agreement (90% CL) between the present ICARUS limit, the published limits of KARMEN [4] and the published positive signals of LSND [5] and MiniBooNE [6] collaborations.

In the last years, ICARUS T600 also gave a prompt contribution in the rejection of the so called "superluminal neutrino" hypothesis through a search for the analogue to Cherenkov radiation in the CNGS neutrino events [7], as predicted by Cohen and Glashow for superluminal neutrinos, and performing a precision measurement of the neutrino velocity using the CNGS bunched neutrino beam [8, 9].

The successful operation of the ICARUS T600 experiment at LNGS has conclusively demonstrated that the LAr-TPC is the leading technology for the future short and long baseline accelerator-driven neutrino physics. A technical paper on the cryogenic system performance is under submission to JINST and a paper on the trigger system performance is at final stage of preparation.

2. ICARUS T600 detector

The ICARUS T600 detector consists of a large cryostat split into two identical, adjacent half-modules with internal dimensions 3.6 × 3.9 × 19.6 m^3 and filled with a total of 760 tons of ultra-pure LAr. Each half-module houses two TPCs separated by a common cathode, with a

drift length of 1.5 m. Ionization electrons, produced by charged particles along their path, are drifted under uniform electric field (E_D = 500 V/cm) towards the TPC anode, made of three parallel wire planes, facing the LAr active volume. A total of about 54000 wires are deployed, with a 3 mm pitch, oriented on each plane at different angles with respect to the horizontal direction (0^0, + 60^0, - 60^0). The drift time of each ionization charge signal, combined with the electron drift velocity information (v_D = 1.55 mm/s), provides the position of the track along the drift coordinate. Combining the wire coordinate on each plane at a given drift time, a three-dimensional image of the ionizing event can be reconstructed [10].

2.1. Computing resources

The ICARUS collaboration counts many INFN sections (LNGS, Milano, Padova, Pavia, Naples) but also a variety of foreign institution, mainly from Poland. The computing resources of the experiment have thus been dimensioned and organized in order to guarantee an easy and safe data accessibility from all the collaboration sites. An "on-line" storage system is located in the underground ICARUS control room. It hosts a 40 TB storage element, upgraded with an additional 50 TB element during the 2012 beam stop. The DAQ dedicated disks belong to a SAN (over a fiber channel network) and are directly mounted on every online machine, including the host installed in the external LNGS computing centre. The "off-line" local resources at the outside LNGS buildings count four 24TB raid-5 storage units, for a total of 72 TB, available online on a cluster of 6 processing units (8 CPU cores each). One of the processing units serves as access point to the Grid and installs the User Interface package. The offline system is completed by a 24-slot LTO4 Tape Library. As additional resources for data analysis, one processing unit and one storage unit of the same type are replicated at the INFN sections of Milano, Padova, Pavia and Naples. The ICARUS experiment is represented, on the Italian Grid Infrastructure, by a Virtual Organization named *icarus-exp.org*. A buffer of about 150 TB is reserved online for ICARUS at the Tier1, where a D0T1 sevice is provided. Wide band, easy and safe accessibility of data is guaranteed for all the groups in the ICARUS collaboration involved in the analysis.

2.2. Data management

An ICARUS event is defined as the whole of the information coming from the two T600 half-modules: the event size is of about 100 MB and the global average trigger rate[1] is about 350 mHz, driving to a total throughput of 87 TB per month. A big effort has been taken in order to reduce the total throughput to a much lower value. At each trigger, performed exploiting both the LAr scintillation properties and the local ionization charge deposition inside the detector, one waveform per wire (i.e. per electronic channel) is collected; waveforms are packed in four different streams, one per TPC. Data are initially stored on the online storage system in the underground control room. The ICARUS data flow has been structured in five sequencial steps: (1) Data Quality Monitor; (2) data streams merging and zipping; (3) empty events filtering through the Cosmic Event software Filter (CEF); (4) data transfer to the offline storage system; (5) long term data storage on tape. The Data Quality Monitor checks the compatibility, in time, of the four data streams tagged with the same event number by the DAQ. If the event passes this first filter, the four data streams are copied and contextually merged and zipped with a lossless compression factor of about 56%. Original streams are preserved for redundancy: from this stage on, a double copy of each event is guaranteed by the data-flow configuration, up to the final destination of data. After an initial validation phase, during which all events where anyway preserved up to the final stage of the data flow, the CEF was finally applied for an effective data reduction: events classified as empty by the CEF were deleted[2], with a reduction

[1] The global rate includes the cosmic event trigger, the CNGS neutrino trigger and trigger based on local charge deposition inside the detector.

[2] One empty event every thirty was anyway kept as checking sample.

to about the 15% of the total amount of data. After empty events filtering, the global trigger rate gets to about 50 mHz, being dominated by cosmic ray events. Considering the lossless data compression, the total data throughput is of 7 TB per month. The data transfer from the online to the offline ICARUS storage system were performed through the 8 km optical fiber connecting the underground Hall B and the LNGS outside buildings. Data were continuously transferred from the offline storage to the tape filesystem facility at CNAF Tier1, at a maximum rate of 30 MB/s. In order to ensure a proper level of redundancy, data were also copied on LTO4 tapes at LNGS. The final double copy on tape of the ICARUS data is the green light for data deletion from the online and the offline storage buffers, thus freeing disk space for further data taking. In this sense, the continuous availability of CNAF sevices, including the technical support to the collaboration, can be regarded as an important ingredient of a continuos and successful data taking.

References

[1] C. Rubbia, CERN-EP/77-08 (1977).
[2] M. Antonello et al. [ICARUS Coll.], Search for anomalies in the ν_e appearance from a ν_μ beam , Eur.Phys.J. C73 (2013) 2599.
[3] M. Antonello et al. [ICARUS Coll.], Experimental search for the LSND anomaly with the ICARUS LAr TPC detector in the CNGS beam, Eur. Phys. J. C, 73:2345 (2013).
[4] B. Armbruster et al. (KARMEN), Phys. Rev. D 65 (2002) 112001, hep-ex/0203021.
[5] A.Aguilar et al. (LSND), Phys. Rev. D 64, 112007 (2001).
[6] A.Aguilar et al. (MiniBooNE), Phys. Rev. Lett. 110, 161801 (2013).
[7] M. Antonello et al. [ICARUS Coll.], A search for the analogue to Cherenkov radiation by high energy neutrinos at superluminal speeds in ICARUS, Physics Letters B 711 (2012) 270-275.
[8] M. Antonello et al. [ICARUS Coll.], Precision measurement of the neutrino velocity with the ICARUS detector in the CNGS beam, Journal of High Energy Physics, JHEP 11 (2012) 049.
[9] M. Antonello et al. [ICARUS Coll.], Measurement of the neutrino velocity with the ICARUS detector at the CNGS beam, Physics Letters B 713 (2012) 17-22.
[10] M. Antonello et al. [ICARUS Coll.], Precise 3D track reconstruction algorithm for the ICARUS T600 liquid argon time projection chamber detector, Adv. High Energy Phys. (2013) 260820.

Kloe data management at CNAF

Stefano Dal Pra[1]**,Francesco Sborzacchi**[2]
[1] INFN-CNAF, viale Berti-Pichat 6/2, 40127 Bologna, Italy
[2] INFN-LNF, Via Enrico Fermi 40, 00044 Frascati, Italy

E-mail: stefano.dalpra@cnaf.infn.it

Abstract. The Kloe experiment, built at the DAFNE factory in Frascati performs CP-violation studies of the order of 10^4 events/s, corresponding to a total throughput of 50 MB/s. A local storage tape library (IBM 3494) was deployed to host the collected raw data together with an onsite computing farm based on AIX–UNIX running on IBM Power processor architecture. During June 2011 the system was updated to handle an off–site copy toward the storage at INFN–T1 of the locally produced data, to prevent losses in case of disaster. The Kloe data management model had to be redesigned and adapted for remote access, while keeping the change transparent to the end user, and bridging between different architectures such as GNU Linux on x86 vs IBM AIX on PPC. The implemented model had proven reliable and effective respect to the constraints. We describe in this report how the data migration was implemented and performed. Since the only standard Grid tools available on AIX platform are provided by the `globus` suite, a number of standard procedures had to be customized and adapted to deal with the special Kloe needs.

1. Introduction

This report describes the work done to adatpt the KLOE [1] data management model [2] in order to migrate its data, approximatively 500TB, and relay on the INFN–T1 storage infrastructure [3]. To achieve this, the following constraints had to be dealt with:

Transparent to the end user The computational model must not change. The activity of the researcher prosecutes as usual. The actual physical location of the data being processed doesn't affect the way he/she works.

Avoid or minimize downtimes It should be possible to perform ordinary computational activities while the migration process is in progress.

DAQ Compliant Further data acquisition from the KLOE revelator have to be permanently stored at INFN-T1. An adequate steady bandwith (100 MB/s) must be guaranteed for the transfer of the generated data flow to be sustainable.

Architecture compliant The usual datatransfer adopted at INFN–T1 is based on a GRID middleware running on RHEL GNU/Linux platforms, whereas the KLOE's computational infrastructure is IBM/AIX based. The data transfer must be compatible with these architectures, in a bidirectional fashion.

The operation of archiving on tape the disk copy of a file is called *migration*. The inverse operation is called *recall*. In later sections we data migration from the KLOE storage system to the one at INFN–T1 will be consider, as also the recall of data stored at INFN–T1 back to the KLOE disk storage.

2. The Kloe storage model

Events collected on data–tacking from the DAQ are stored raw datafiles on YBOS format [4] on a local disk buffer area (6TB). The Reconstructed datafiles are built from raw datafiles analisys and stored on another disk buffer area (26TB). Both these disk area are accessible through NFS mountpoint to the I/O servers, automatically taking care of the archiving on tape. When needed, files can be recalled from tape back on a third disk area for offline analisys (40TB). Status manegement of the file resources is kept on a IBM DB2 database. A file resource is identified by a URI, whose status is retrieved through a SQL request.

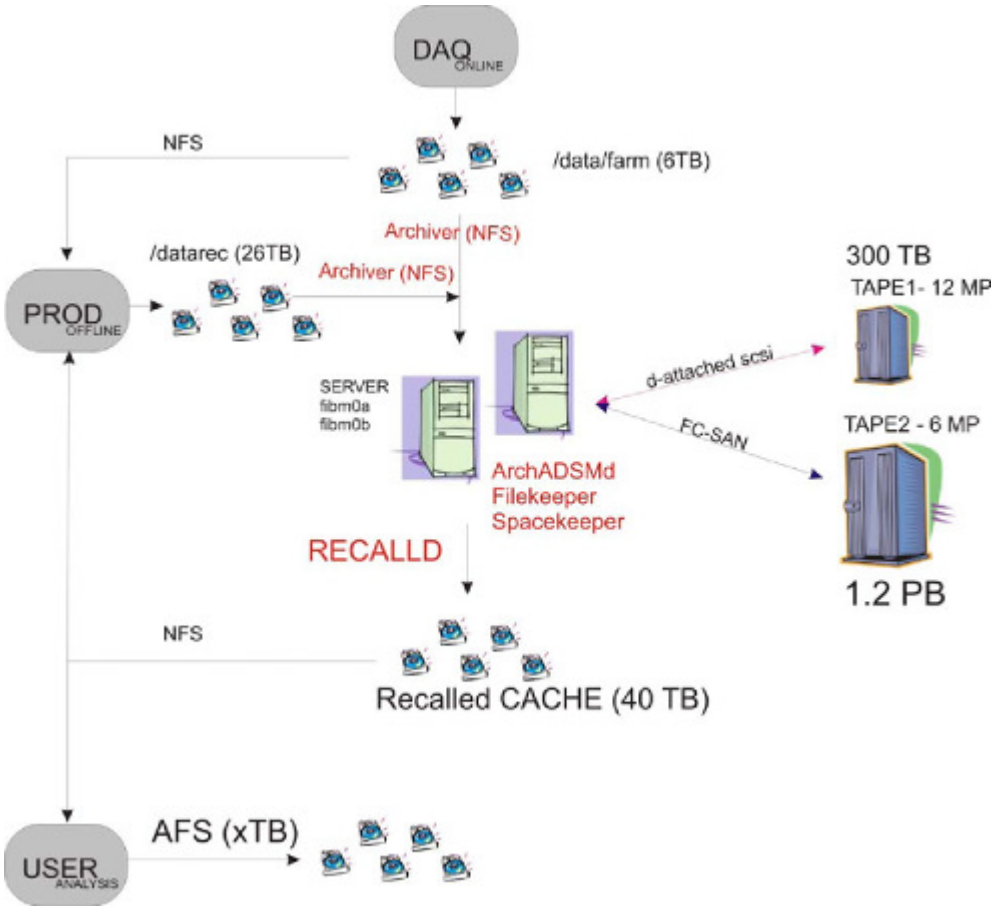


Figure 1. The KLOE storage model

3. The INFN–T1 storage model

The Massive Storage System selected to host KLOE data works in a D0T1 modality: every file must have a copy on Tape; a physical copy on disk may or may not be present. The Disk space is on a GPFS filesystem and tape libraries are managed by TSM. Interaction between the two subsystems is made possible by the GEMSS module, developed at INFN–T1. Each file resource is identified by its position in the filesystem, through its absolute path. When the file *content* is only available on tape, the file is replaced on the GPFS filesystem by a *stub file*. Attempting to access it triggers the recall from tape. Remote management access is offered by the StoRM service [5]. Securely authenticated read write access [6] is provided by gridftp servers.

4. Design and implementation
Details of the implemented data transfer mechanism are provided. The main tools used have been: python and MySQL for scripting and catalog at INFN–T1 side, python and the globus tools for scripting and file transfer.

4.1. File transfer
The selected instrument, available for AIX is the `globus-url-copy` command, providing a `gridftp` client. This also implies secure authentication for both client and server. To enforce reliability of the copy to INFN–T1, an `adler32` checksum is computed at source while reading the file from disk, and at destination. A file transfer is marked as succesfully completed only if the checksum at the destination, computed from the copy on disk, matches the one at the origin. The operation of copying a file from the original KLOE storage to the one at INFN–T1 is called *migration*. The backward copy process from INFN–T1 to the local KLOE disk area is called *recall*.

It is worth noticing the fact that `globus-url-copy` offers a nice `--recursive` option to create possibly missing directories on the remote side. This feature is however disabled in the `gridftp` endpoints at INFN–T1, for consistency with the `srm` protocol.

4.2. File catalog and status
A MySQL database on a User Interface host at INFN–T1 keeps track of the files to transfer, their progress status, both for migrations and recalls. The relational database provides a reliable queing mechanism and an effective serialization of concurrent requests. This prevents the risk of race conditions.

4.3. The migration process
Every file to be transferred is added by a script at source side to the `fileset` table, with references in the `tomigr` table on the database. The file is specified by its destination full path, and its `md5` hash which is stored as Primary Key and used for lookup purposes in the table. This also prevents the risk of inserting duplicates. An integer `status` field in the fileset table indicates the progress in the copy process. Each table has its `status` field. See the table below for a description of the various status.

The `tomigr` table is adopted to guarantee responsiveness even when the fileset table grows to huge size; this in fact only acts as a queue for files to be copied, whose entries are deleted after the process reach a terminal status.

At the end of a successful file migration, the corresponding entry in the `fileset` table reports fullpath, size, adler32 checksum at source and destination, as also initial and final time of transfer, thus providing useful data to estimate the overall transfer rate over any selected time interval.

4.4. The overall mechanism
4.4.1. Migration process At source side, an agent inserts file entries on the `fileset` and `tomigr` tables. Then, one or more other agents keep selecting one file for transfer, by updating its status in the `tomigr` table. This is an atomic operation, so there cannot be two agents booking the same file. After that the agent copies the file with `globus-url-copy` while computing its `adler32` checksum. Upon success, the `fileset.status` goes to 100 or -100 in case of failure, and the checksum is updated in the `fileset` table.

At destination side, an agent checks for entries in the `tomigr` table with a matching entry having `fileset.status=100`. It then queries the filesystem for the `adler32` checksum, which is stored as an extended attribute. On postitive match, `fileset.status` is updated to 200 and the entry in `tomigr` is deleted.

`fileset.status`	
0	to be transferred
100	transfer succesfull
−100	transfer failed
200	checksum match
−200	checksum mismatch
`tomigr.status`	
0	to be transferred
1	booked for transfer
−1	transfer failed
`torecall.status`	
0	to be recalled from tape
1	booked for recall
100	recall requested
200	file is on disk
−200	recall failed

Table 1. Relevant statuses for file, file migration, and file recall

4.4.2. Recall process At destination side (KLOE), an agent inserts entries in the `torecall` table, each one with initial status set to zero. Meanwhile, one or more agents keep selecting one entry to copy by updating to 1 the status of one row having status to zero. The agent then polls for the status to reach the status 100. When this happens the file is on disk and the agent can retrieve the remote file with `globus-url-copy`. Upon successful transfer the record in the `torecall` table is deleted.

At source side (INFN-T1) an agent polls the `torecall` table for entries with status set to zero. The corresponding files are then checked on the filesystem. Those having a copy on disk gets their status updated to 100. The other ones are recalled to disk by issuing the GEMSS command `yamsRecall`. The script then keeps reading the `YAMSS_STAT/recall` file, where the YAMSS system notifies the recall events. When the successful recall of a requested file appears, `torecall.status` is updated to 100.

5. The migration

After test and setup, the migration begun on June 2011. The implemented mechanism proved to be effective and fulfilling the initial requirements. The maximum transfer rate was near to $100MB/sec$, with up to 8.3TB transferred per day.

Year–Month	Migrated files	size (GB)
2012–01	18520	12788
2012–02	26565	20800
2012–03	10500	13613
2012–04	3432	2743
2012–05	140548	96233
2012–06	75889	57448

Table 2. Files transferred over a six month period.

Migration and recall activities can go on simultaneously and independently.

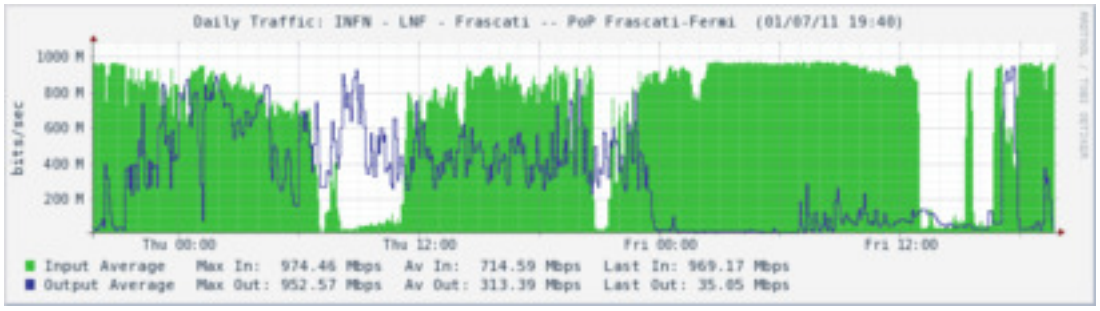


Figure 2. Simultaneous recall (blue) and migration (green) activity.

6. Integration with the computational model

Once the file transfer from/to INFN–T1 works as expected, the computational model must be adapted to be integrated with it.

The standard usecase, before migration A set of files is selected for analysis. Their location on the local tape library (or in the buffer disk area) is retrieved through a local IBM/DB2 database; the files are recalled from tape to disk, then the computation begins. Output files are possibly archived on tape.

Adapting the usecase Every *recall file_X from local tape library* operation must be converted to *recall file_X from INFN–T1* if the file is not found locally, and every *archive outfile_Y* operation must be converted to *migrate outfile_Y to INFN–T1*.

To make the adaptation, first of all it must be possible to map a file to its full pathname on the new remote storage and add a pointer to it in the database schema. When this is `NULL`, the file is a candidate for migration. When this is present, the local tape copy may be freed. Then each attempt to access the tape library is converted to the corresponding operation on the remote storage if needed, or in the local one if a copy is available.

7. Conclusion

A mechanism to migrate custodial data acquired by the KLOE revelator and archived on the local tape library to the D0T1 storage system at INFN–T1 without affecting nor interrupting or altering the research activities depending on these data has been designed, implemented and effectively put on production. Although a number of improvements are possible, the core functionalities have been working without significative problems, unforeseen or undesired behaviours, over a period of three years now, thus confirming the validity of the initial design and its implementation.

This work also demonstrates how a local data and computational management model can be integrated in a distributed paradigm even when only a bare minimal subset of the ordinary Grid middleware can be taken into use.

References

[1] http://www.lnf.infn.it/kloe/
[2] I. Sfiligoi for the KLOE collaboration "KID – KLOE Integrated Dataflow", Jul 13, CHEP 2001
[3] A. Cavalli, S. Dal Pra, L. dell'Agnello, A. Fella, D. Gregori, L. Li Gioi, B. Martelli, A. Prosperini, P.P. Ricci, V. Sapunenko, V. Vagnoni "Experience with Hierarchical Storage Management based on GPFS and TSM at INFN-CNAF", PIK, Vol. 14, pp. 1–6, March 2010
[4] http://www-cdf.fnal.gov/upgrades/computing/projects/run2mc/murat/doc/ybos/ybos.01.html
[5] E. Corso et al. "StoRM: A SRM Solution on Disk Based Storage System", Proceedings of the Cracow Grid Workshop 2006 (CGW2006), Cracow, Poland, October 15–18, 2006.
[6] http://toolkit.globus.org/toolkit/docs/3.2/security.html

The KM3NeT detector

A. Margiotta

Dipartimento di Fisica e Astronomia – Università Bologna e Sezione INFN Bologna

margiotta@bo.infn.it

1. The physics case

The KM3NeT is a deep-sea research infrastructure being constructed in the Mediterranean Sea. It will host the next generation Cherenkov neutrino telescope and nodes for a deep-sea multidisciplinary observatory that will provide oceanographers, marine biologists, and geophysicists with real time measurements [1].

One of the main goals of the KM3NeT neutrino telescope is the discovery of high-energy neutrino Galactic sources. A detector with an effective volume larger than a cubic kilometre is required because of the low fluxes expected from possible sources and of the small interaction neutrino's cross section. Its location, in the Mediterranean Sea, is optimal to look at the Galactic plane and the Galactic centre. These regions of the sky host several TeV gamma-ray sources that could emit measurable neutrino fluxes between a few TeV and a few 10 TeV. Actually, KM3NeT's sensitivity to neutrinos extends to a wider energy range (~100 GeV to ~100 PeV).

The excellent optical properties of water (long scattering and absorption lengths) allow the reconstruction of muon tracks with an angular resolution better than one degree. This makes the charged current interactions of muon neutrinos in the vicinity of the detector the golden channel for source identification. Nevertheless, other channels will be explored as well. In the case of neutral currents and electron-neutrino interactions, which produce shower-like events, the angular resolution is expected to be around 10-15 degrees.

2. The detector

The detector will have a modular structure with a total active volume of about 3 km^3. It will complement the IceCube telescope, at the South Pole, in its field of view, exceeding it in sensitivity. It will be distributed at three sites: KM3NeT-Fr, offshore Toulon, France, KM3NeT-It, offshore Portopalo di Capo Passero, Sicily (Italy) and KM3NeT-Gr, offshore Pylos, Peloponnese, Greece. In each site, one or more modules, designed according to the same technology, data handling and operation control, will be installed. Each module, called a Building Block, consists of about one hundred Detection Units (DUs). A DU is a flexible string, 700 m long, anchored at the sea bottom and kept taut by a system of buoys, carrying 18 storeys, with a vertical spacing of 36 m.

A major innovation concerns the design of the active part of a neutrino telescope, the Digital Optical Module (DOM). It consists of a 17" glass sphere, resistant to the high pressure present at the sea bottom and houses 31 3" photomultiplier tubes (PMT) and the active bases for power. A multiPMT DOM has several advantages if compared to the traditional, large cathode single-PMT optical module used in ANTARES, Baikal and IceCube detectors, like, for example, a larger (three to four times) total photocathode area and a better discrimination of single vs multi photoelectrons.

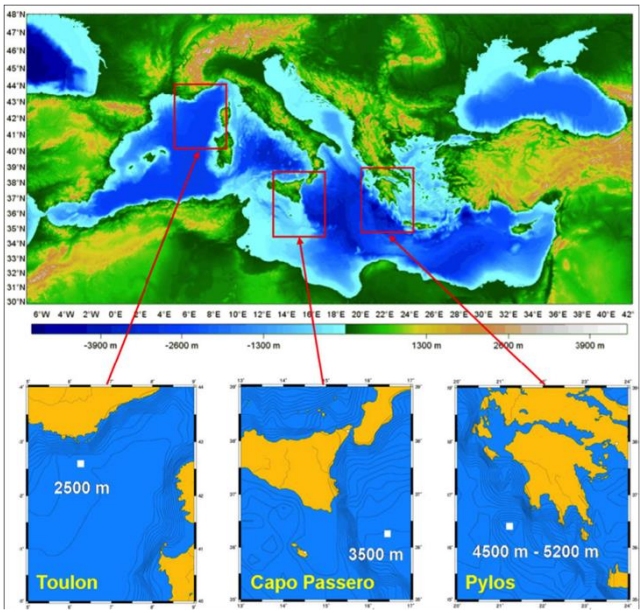


Fig. 1 Location of the three proposed sites for the construction of the KM3NeT neutrino telescope in the Mediterranean Sea.

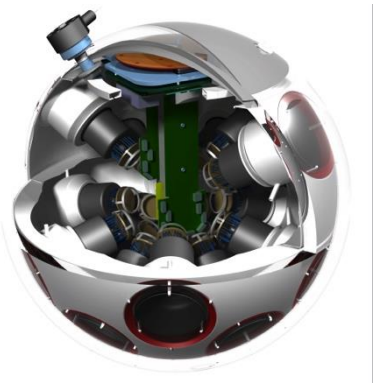


Fig. 2 - Internal structure of a KM3NeT DOM.

Each DOM acts independently as a remote node of the network connecting the off-shore detector with the computing resources on-shore. All DOMs are synchronized to the subnanosecond level using a clock signal broadcast from shore.

The readout electronic boards (Central Logic Board, CLB) control the data acquisition and communications with the shore station and are also hosted in the DOM. The signal from a PMT consists of the arrival time and the width of the pulse, measured as the time-over-threshold (ToT), typically set at 0.3 p.e., digitized and sent via a network of optical fibers to shore. Long range transmissions exploit DWDM techniques at 50 GHz spacing. The "all-data-to-shore" concept is applied to the readout of the detector, following the experience of the ANTARES detector. On shore, the physics events are filtered from the background using a dedicated software and stored on disk. The maximum throughput from each DOM is 200 Mb/s, taking into account the possibility of persistent bursting activity due to bioluminescent organisms. The corresponding rate can vary depending on the site. The first bunch of strings will be deployed at the beginning of 2015. With the presently available

funding, the construction of the first phase of the project has started (KM3NeT-phase1). The Capo Passero site will host a set of 8 detection units constructed according to a previous design (NEMO tower) and 24 string-like DUs, starting at the end of 2014.

3. Collaboration with CNAF

The "all-data-to-shore" approach requires a particular effort in the definition of the triggering algorithms. A strong reduction of the huge amount of data arriving at the shore station must be accompanied with an efficient preservation of all potentially interesting events.
A strict collaboration between the KM3NeT Bologna group, which has the coordination of the data acquisition system, and CNAF software engineers has started on this item [2].

The KM3NeT collaboration will exploit the CNAF infrastructures for the permanent storage of reduced data. The computing model of the KM3NeT collaboration is still being defined in its details and needs the approval of financial requests from the CSN2 of INFN, but the framework has been already fixed.

Hereafter, the general scheme for treatment of data collected at the Italian site (Capo Passero) is shortly described. It refers to the installation of the 8 NEMO-towers. The main features of this scheme can be easily extended to a larger detector (the 24 strings that will be deployed during 2015 and possibly an entire Building Block) including others computing resource providers (for example, the Lyon CC).
Data collected by the telescope and transmitted via a 100 km long electro-optical cable are processed and filtered at the on-shore station using a dedicated computer farm with specific algorithms. Filtered raw data are temporarily stored at Laboratori Nazionali del Sud and then transferred to CNAF for permanent storage and to RECAS sites for processing (data calibration and event reconstruction, essentially) and a 1-year rolling storage capacity. When a larger detector will be taking data, a mirror permanent data repository will be considered at other European centres, for example CC-Lyon. During the first year of data taking a tuning of the computing KM3NeT necessities is foreseen.

References

[1] A. Margiotta: Status of the KM3NeT project, JINST 9 (2014) C04020
[2] T. Chiarusi, F. Giacomini, M. Manzali and C. Pellegrino, "The Trigger and Data Acquisition system of the KM3NeT-Italy detector", this report

LHCb Computing at CNAF

V. Vagnoni
INFN Sezione di Bologna, via Irnerio 46, 40126 Bologna, Italy

E-mail: Vincenzo.Vagnoni@bo.infn.it

Abstract. A quick overview of the LHCb computing activities is given, including the latest evolutions of the computing model. An analysis of the usage of CPU, tape and disk resources in 2013 is presented, emphasising the achievements of the INFN Tier-1 at CNAF. The expected growth of computing resources in the years to come is also briefly discussed.

1. Introduction

The Large Hadron Collider beauty (LHCb) experiment [1] is one of the four main particle physics experiments collecting data at the Large Hadron Collider accelerator at CERN. LHCb is a specialized *c*- and *b*-physics experiment, that is measuring rare decays and *CP* violation of hadrons containing *charm* and *beauty* quarks. The detector is also able to perform measurements of production cross sections and electroweak physics in the forward region. Approximately the LHCb collaboration is composed of 800 people from 60 institutes, representing 15 countries.

The experiment has a wide physics programme covering many important aspects of heavy flavour, electroweak and QCD physics. The core LHCb physics measurements notably include the branching ratio of the rare $B_s \to \mu^+\mu^-$ decay [2], the forward-backward asymmetry of the muon pair in the flavour changing neutral current $B^0 \to K^{*0}\mu^+\mu^-$ decay [3], the *CP* violating phase in the decay $B_s \to J/\psi\phi$ [4], the properties of radiative *B* decays [5], the determination of the unitarity triangle angle γ [6], and charmless charged two-body *B* decays [7]. More than 180 physics papers have been heretofore produced.

The LHCb detector is a single-arm forward spectrometer covering the pseudorapidity range between 2 and 5. The detector includes a high-precision tracking system consisting of a silicon-strip vertex detector surrounding the *pp* interaction region, a large-area silicon-strip detector located upstream of a dipole magnet with a bending power of about 4 Tm, and three stations of silicon-strip detectors and straw drift tubes placed downstream. The combined tracking system provides a momentum measurement with relative uncertainty that varies from 0.4% at 5 GeV/*c* to 0.6% at 100 GeV/*c*, and impact parameter resolution of 20 μm for tracks with high transverse momenta. Charged hadrons are identified using two ring-imaging Cherenkov detectors. Photon, electron and hadron candidates are

identified by a calorimeter system consisting of scintillating-pad and preshower detectors, an electromagnetic calorimeter and a hadronic calorimeter. Muons are identified by a system composed of alternating layers of iron and multiwire proportional chambers. The trigger consists of a hardware stage, based on information from the calorimeter and muon systems, followed by a software stage, which applies a full event reconstruction. A sketch of the LHCb detector is given in Fig. 1.

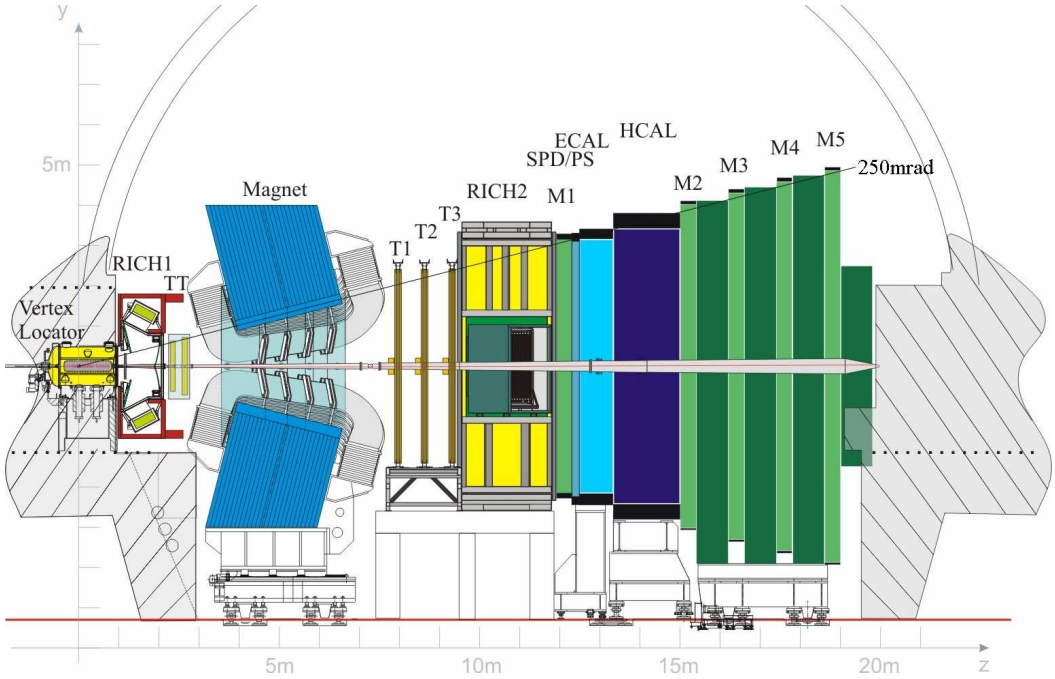


Fig. 1: Sketch of the LHCb detector.

2. Recent evolutions of the LHCb computing model

In the initial LHCb computing model, described in the LHCb Computing TDR [8], it was foreseen that all production and analysis activities, except simulations, had to be performed at the Tier-0 and Tier-1s, whereas the Tier-2s were devoted exclusively to run Monte Carlo jobs. However, this model had a number of shortcomings that made it expensive on resources. As reprocessing could only run at CERN and Tier-1s and had to be completed within two months, large peaks in the CPU power were required at Tier-1s, increasing with accumulated luminosity. Another limitation was due to the fact that the jobs were required to run at sites holding the input data. Since disk was located only at CERN and Tier-1s, this caused inflexible use of CPU resources, with only simulation allowed to run at Tier-2 sites. In addition, to allow all Tier-1 sites to be treated equally in the job matching, each site was holding one complete disk copy of all active analysis datasets, which was very demanding on storage space.

The limitation of running reprocessing on Tier-1s was relaxed in 2011. Reprocessing jobs were executed on a selection of Tier-2s by downloading the input RAW files from Tier-1 storage, running the job, and uploading the reconstruction output to the same storage. This was generalised in 2012 when 45% of reconstruction CPU time was provided outside CERN and Tier-1s. In 2012, only 30% of the RAW data was processed by the first pass reconstruction and used mainly for monitoring and calibration. The reprocessing was run continuously on the full dataset once calibration constants were available, within 2-4 weeks of data taking, removing the need for end of year reprocessing.

The DIRAC framework [9] for distributed computing has allowed to easily integrate non WLCG resources in the LHCb production system. In 2013, 20% of CPU resources were routinely provided by the LHCb HLT farm, and a further 6.5% by the Yandex company. Small scale production use has been made of virtual machines on clouds and other infrastructures, including volunteer computing through the BOINC framework. LHCb is therefore in a good position to make production use of opportunistic resources.

The use of storage resources has been optimised by reducing the number of disk-resident copies of the analysis data. Recently, the possibility has been introduced to provide disk also at certain large Tier-2s, which are therefore opened to analysis jobs, further blurring the functional distinction between Tier-1 and Tier-2 sites in the LHCb computing model.

The LHCb data model is illustrated in Fig. 2, which shows the various data formats and the processing stages between them. The acronyms used in the figure are defined in Tab. 1. All RAW data from the pit are transferred to CERN Castor and written to CERN Tape (3 GB files). A second, distributed, copy of the RAW files is made on Tier-1 tape, shared according to share of total tape pledge. The FULL.DST is not available for end-user analysis. The Stripping step is a data reduction step that selects events of interest for specific physics analyses from the FULL.DST files; the selected events are streamed to DST or MDST output files that are made available to end users.

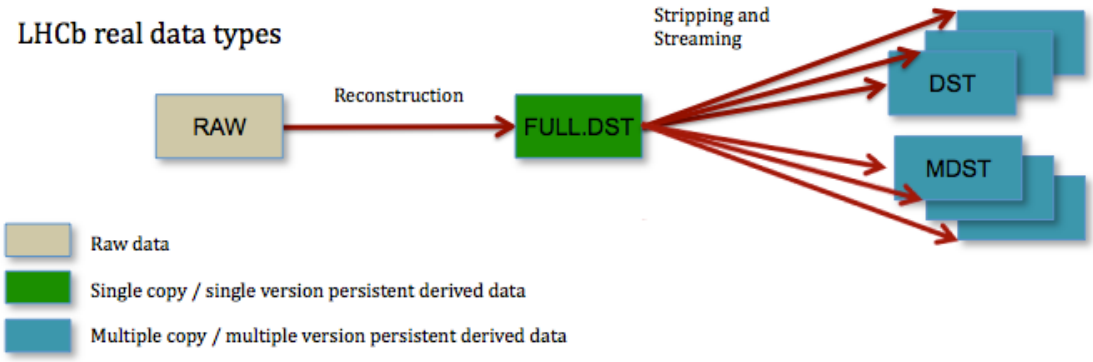


Fig. 2: The LHCb data model.

RAW	Raw data: all events passing the trigger. Input to reconstruction (prompt or reprocessing).
FULL.DST	Complete reconstruction output for all physics events, plus a copy of the Raw event. Input to stripping (could also be used for reprocessing). Persistent (tape only, one copy), to allow restripping.
DST	Output of stripping: events selected by physics criteria, complete copy of reconstructed event plus particle decay tree(s) that triggered selection. Self-contained, input to user analysis. Persistent, multiple copies on disk.
MDST	MicroDST, same event model as DST, but containing only subset of event (tracks, PID) that triggered the selection, and minimal Raw data (mainly trigger information). Self-contained, input to user analysis. Content defined on per stream basis. Persistent, multiple copies on disk.

Tab. 1: LHCb data formats.

Currently, simulation workflows consist of a number of steps that are run in sequence in the same job. Because of the amount of CPU required for the event generation and GEANT tracking steps, the number of events produced per job is small (a few hundred), resulting in output files of ~100-250MB. Because they have no input data, simulation workflows can run on any computing element that is compatible with the LHCb production platforms (currently SLC5 and SLC6 on x86 64-bit architectures), including unpledged resources such as the HLT farm or non-WLCG sites.

In all LHCb production workflows, input data files are copied from a Grid SE to the local disk of the worker node when a job starts, and output data files are uploaded to a Grid SE at the end of the job. It has been found that this model leads to a greater overall processing efficiency, due to the non-negligible probability of failure in opening a remote file by a running job. Production jobs are configured such that all input and output data fits into a 20 GB disk allocation on the worker node. In data analysis activities (user jobs), with sparse data access on large numbers of 5 GB input files, the above model cannot work and data are accessed remotely, in general via the file or xrootd protocols (though other protocols are also supported by the application software). A mechanism is being implemented to access different replicas in turn if a job fails to open the chosen replica of a file.

As a file catalog to pinpoint active file replicas, LHCb is currently using a central LCG File Catalog (LFC) instance located at CERN, with separated read-write and read-only services. The performance of the LFC is adequate for the foreseeable future. However LHCb is considering the replacement of the LFC by the integrated DIRAC File Catalog (DFC) that is better adapted within the DIRAC Data Management System, as it provides natively functionalities that are not available in the LFC.

3. Resource usage in 2013

Table 2 shows the resources pledged for LHCb at the various tier levels for the 2013 period.

2013	CPU	Disk	Tape
	(kHS06)	(PB)	(PB)
Tier0	34	4.0	6.5
Tier1	92	7.0	9.5
Tier2	52		

Tab. 2: LHCb 2013 pledges.

The usage of WLCG CPU resources by LHCb is obtained from the different views provided by the EGI Accounting portal. The CPU usage for Tier-0 and Tier-1s is presented in Fig. 3. The same data is presented in tabular form in Tab. 3. It must be emphasised that CNAF has provided more CPU power than any other centre, including CERN. This has been possible owing to great stability, in particular of the storage system, leading to maximal efficiency in the overall exploitation of the resources.

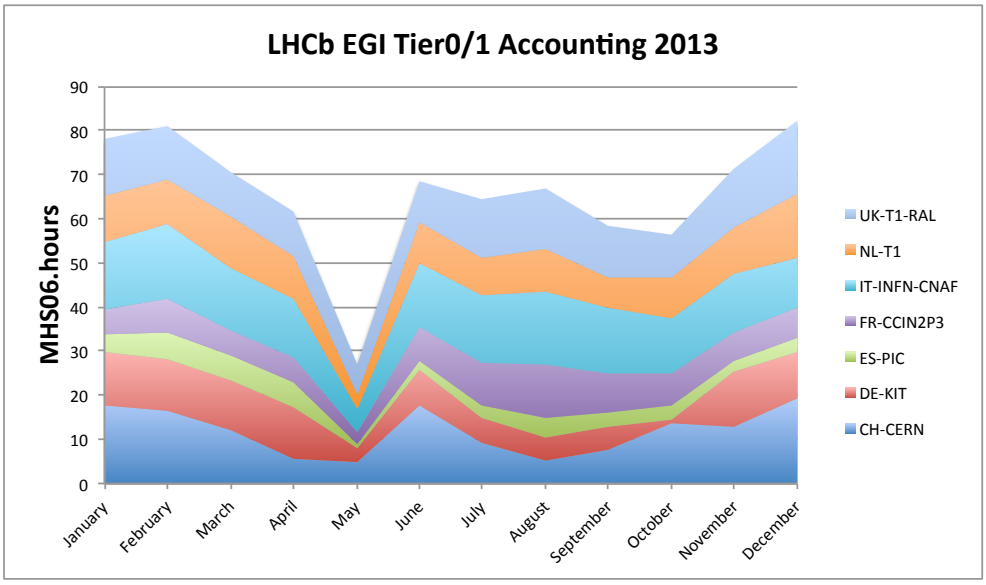


Fig. 3: Monthly CPU work provided by the Tier-0 and Tier-1s to LHCb during 2013.

<Power>	Used (kHS06)	Pledge (kHS06)
CH-CERN	16.4	34.0
DE-KIT	11.3	19.2
ES-PIC	5.1	5.6
FR-CCIN2P3	9.9	16.5
IT-INFN-CNAF	18.9	16.5
NL-T1	13.1	13.8
UK-T1-RAL	16.2	20.5
Total	91.1	126.1

Tab. 3: Average CPU power provided by the Tier-0 and the Tier-1s to LHCb during 2013.

The number of running jobs at Tier-0 and Tier-1s is detailed in Fig. 4. As seen in the top figure, except for about a month, LHCb has also been running simulation at Tier-1s.

The usage of the Storage is the most complex part of the LHCb computing operations. Not much new tape storage was necessary in 2013, as there was only few new data coming in. The rate of data, resulting from stripping and simulation production, archived on tape is 100 TB/month. The used tape space at CERN and Tier-1s at the end of March 2014 does not exceed the 2013 pledges. Figure 5 shows that the incremental stripping in autumn 2013 was completed within 1 month. The amount of data recalled from tape for this stripping campaign was 3.75 PB. Table 4 reports the staging throughput achieved at Tier-1s in the autumn stripping campaign, and the number of days during which the site was staging at that rate. It is apparent that CNAF has shown the best performance in terms of staging throughput, even better that CERN, owing to the home-grown GPFS-GEMSS-StoRM tape management system. Table 5 shows the situation of disk storage resources at the Tier-0 and Tier-1s at the end of January 2014. Despite the lower disk pledges, CNAF has been the second Tier-1 in terms of disk storage made available to LHCb.

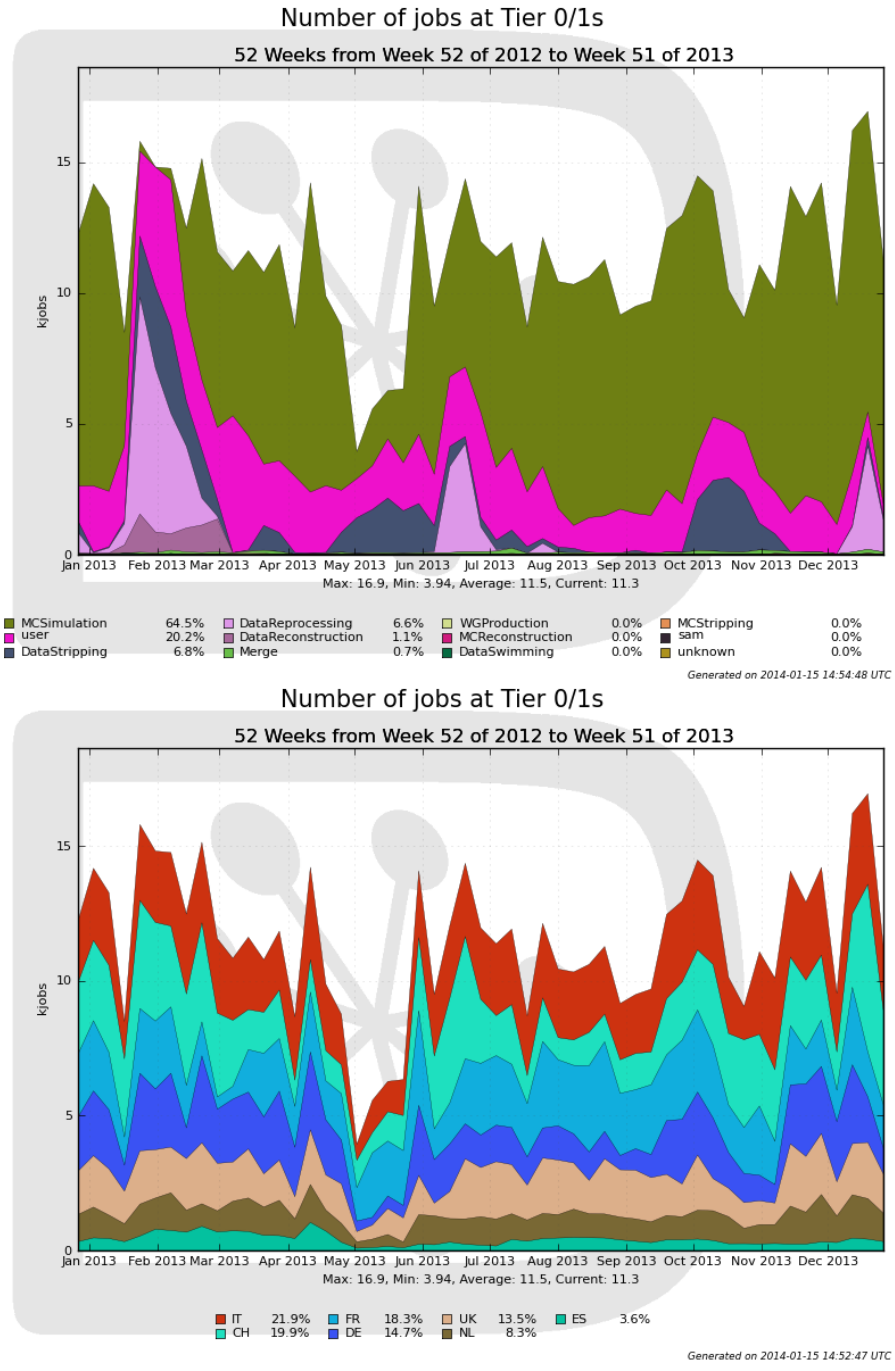


Fig. 4: Usage of LHCb resources at Tier-0 and Tier-1s during 2013. The top plot shows the usage of resources for the various activities, whereas the bottom plot shows the contributions from the different countries.

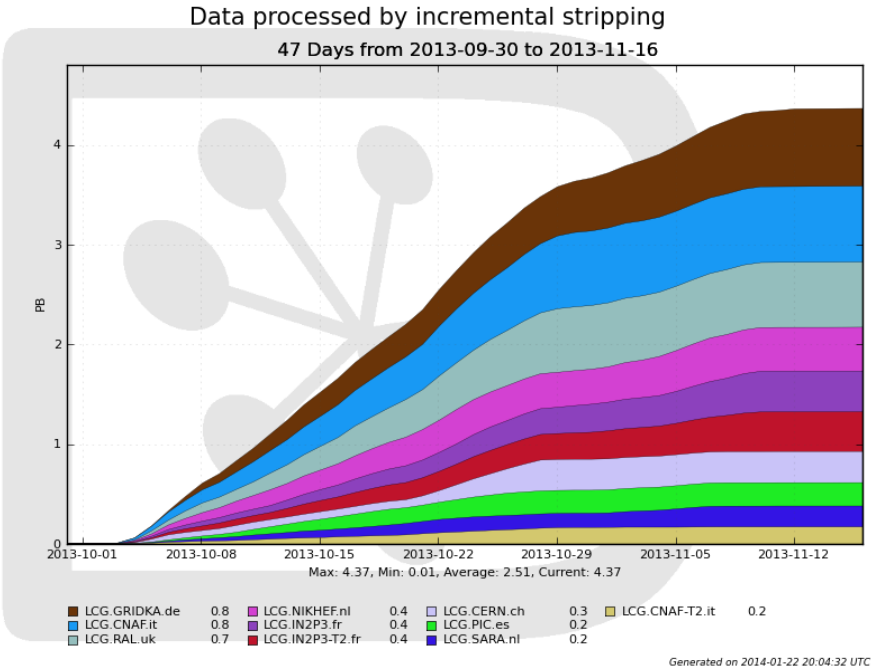


Fig. 5: Data processed by the autumn 2013 incremental stripping.

Site	CERN	CNAF	GridKa	IN2P3	PIC	RAL	SARA
Throughput (MB/s)	334	369	255	231	106	274	276
Number of days	6	21	34	35	21	21	17

Tab. 4: Staging throughput achieved at Tier-1s during the autumn 2013 stripping campaign.

Disk (PB)	CERN	CNAF	GRIDKA	IN2P3	PIC	RAL	SARA	Tier-1s
LHCb accounting	**2.95**	**1.34**	**1.26**	**1.01**	**0.60**	**1.65**	**0.86**	**6.71**
Disk used	2.98	1.33	1.26	1.01	0.60	1,76	0.86	6.82
Disk free	0.41	0.23	0.08	0.19	0.05	0,38	0.06	0.99
Stage area (used+free)	0.44	0.35	0.13	0.03	0.01	0.20	0.03	0.75
Disk total	**3.84**	**1.91**	**1.47**	**1.14**	**0.66**	**2.34**	**0.95**	**8.47**
Pledge 2013	***4.00***	***1.30***	***1.45***	***1.20***	***0.44***	***1.60***	***1.01***	***7.00***

Tab. 5: Situation of disk storage resource usage at the end of January 2014, available and installed capacity, and 2013 pledge.

In summary, the usage of the computing resources during 2013 has been quite smooth for LHCb. The reprocessing of 2011/2012 datasets was completed early in the year and an incremental stripping was successfully run in late spring. The main concern there was the rate of recall from tape at some Tier-1s, but measures have been taken in order to improve the performance for the following incremental stripping, which took place during autumn.

Simulation has been running at almost full speed using all available resources. The amount of CPU work achieved during 2013 has slightly exceeded the expectation. The sharing between the various

WLCG Tiers was somewhat different with respect to the estimated needs. The usage of less resources than anticipated at the Tier-0 and Tier-1s was compensated by the possibility of utilizing more resources at Tier1s than pledged by the funding agencies.

4. Expected resource growth

In terms of CPU requirements, Tab. 6 presents for the different activities the CPU work estimates for 2015, 2016, 2017. Note that in this table there are no efficiency factors applied: these are resource requirements assuming 100% efficiency in using the available CPU. The last row shows the power averaged over the year required to provide this work, after applying the standard CPU efficiency factors.

LHCb CPU Work in WLCG year (kHS06.years)	2015	2016	2017
Prompt Reconstruction	19	31	43
First pass Stripping	8	13	9
Full Restripping	8	20	9
Incremental Restripping	0	4	10
Simulation	134	153	198
User Analysis	17	17	17
Total Work (kHS06.years)	**185**	**238**	**286**
Efficiency corrected average power (kHS06)	**220**	**283**	**339**

Tab. 6: Estimated CPU work needed for the various LHCb activities in the years to come.

The required resources are apportioned between the different Tiers taking into account the computing model constraints and also capacities that are already installed. This results in the requests shown in Tab. 7. The table also shows resources available to LHCb from sites that do not pledge resources through WLCG.

Power (kHS06)	Request 2015	Forecast 2016	Forecast 2017
Tier 0	44	53	63
Tier 1	123	148	177
Tier 2	52	62	74
Total WLCG	**219**	**263**	**315**
HLT farm	10	10	10
Yandex	10	10	10
Total non-WLCG	**20**	**20**	**20**

Tab. 7: CPU power requested at the different Tiers in the years to come.

Tables 8 and 9 present, for the different data classes, the forecast total disk and tape space usage at the end of the years 2015-2017. These disk and tape estimates are then broken down into fractions to be provided by the different Tiers. These numbers are shown in Tables 10 and 11. As can be seen the increase in disk storage can be managed to fit inside a reasonable growth envelope by adjustments in the details of the processing strategy. On the other hand, the growth in the tape storage requirement is more challenging but largely incompressible: in Tab. 9 one can see that the major part of the increase is due to raw data that, if not recorded, is lost.

LHCb Disk storage usage forecast (PB)	2015	2016	2017
Stripped Real Data	7.3	13.1	14.7
Simulated Data	8.2	8.8	12.0
User Data	0.9	1.0	1.1
ALL.DST	1.5	1.9	
FULL.DST	3.3		
RAW buffer	0.4	0.5	0.3
Other	0.2	0.2	0.2
Total	**21.7**	**25.4**	**28.2**

Tab. 8: Breakdown of estimated disk storage usage for different categories of LHCb data.

LHCb Tape storage usage forecast (PB)	2015	2016	2017
Raw Data	12.6	21.7	34.5
FULL.DST	8.7	15.2	19.7
ALL.DST	1.8	5.2	7.7
Archive	8.6	11.5	14.7
Total	**31.7**	**53.7**	**76.6**

Tab. 9: Breakdown of estimated tape storage usage for different categories of LHCb data.

LHCb Disk (PB)	2015 Request	2016 Forecast	2017 Forecast
Tier0	6.7	8.3	9.5
Tier1	12.5	14.2	15.4
Tier2	2.5	2.9	3.3
Total	**21.7**	**25.5**	**28.3**

Tab. 10: LHCb disk request for each Tier level. Countries hosting a Tier-1 can decide what is the most effective policy for allocating the total Tier-1+Tier-2 disk pledge.

LHCb Tape (PB)	2015 Request	2016 Forecast	2017 Forecast
Tier0	10.4	15.9	21.6
Tier1	21.3	37.8	55.0
Total	**31.7**	**53.7**	**76.6**

Tab. 11: LHCb tape request for each Tier level.

5. Conclusions

A description of the LHCb computing activities has been given, with particular emphasis on the evolutions of the computing model, on the usage of resources and on the forecasts of resource needs during the next three years. It has been shown that CNAF has been in 2013 the most important LHCb computing centre in terms of CPU power made available to the collaboration. This is a great achievement, which has been possible due to the hard work of the CNAF Tier-1 staff, to the overall stability of the centre and to the friendly collaboration between CNAF and LHCb people. The importance of CNAF within the LHCb distributed computing infrastructure has been recognised by the LHCb computing management in many occasions.

References

[1] A. A. Alves Jr. *et al*. [LHCb collaboration], *JINST* **3** (2008) S08005.
[2] R. Aaij *et al*. [LHCb collaboration], *Phys. Rev. Lett*. **110** (2013) 021801; R. Aaij *et al*. [LHCb collaboration], Phys. Rev. Lett. **111** (2013) 101805.
[3] R. Aaij *et al*. [LHCb collaboration], JHEP **08** (2013) 131.
[4] R. Aaij *et al*. [LHCb collaboration], Phys. Rev. D **87** (2013) 11.
[5] R. Aaij *et al*. [LHCb collaboration], Nucl. Phys. B **867** (2013) 1.
[6] R. Aaij *et al*. [LHCb collaboration], Phys. Lett. B **726** (2013) 151.
[7] R. Aaij *et al*. [LHCb collaboration], Phys. Rev. Lett. **110** (2013) 22; R. Aaij *et al*. [LHCb collaboration], JHEP **1310** (2013) 183.
[8] [LHCb collaboration], CERN-LHCC-2005-019.
[9] F. Stagni *et al*., J. Phys. Conf. Ser. **368** (2012) 012010.

The PAMELA experiment

M Bongi[1,2], F S Cafagna[3], E Mocchiutti[4] for the PAMELA Collaboration

[1] Department of Physics, University of Florence, I-50019 Sesto Fiorentino, Florence, Italy
[2] INFN, Sezione di Florence, I-50019 Sesto Fiorentino, Florence, Italy
[3] INFN, Sezione di Bari, I-70126 Bari, Italy
[4] INFN, Sezione di Trieste, I-34149 Trieste, Italy

E-mail: Massimo.Bongi@fi.infn.it

Abstract. PAMELA has been in orbit studying cosmic rays for more 7 years. Its operation time will continue in 2014. In this paper we will present some of the latest results obtained by PAMELA in flight, describe the data handling and processing procedures and the role of CNAF in this context.

1. Introduction

PAMELA is a satellite-borne instrument designed and built to study the antimatter component of cosmic rays from tens of MeV up to hundreds of GeV and with a significant increase in statistics with respect to previous experiments. The apparatus, installed on board the Russian Resurs-DK1 satellite in a semi-polar Low-Earth orbit, is taking data since June 2006.

PAMELA has provided important results on the antiproton [1] and positron fraction [2, 3] in galactic cosmic rays. The high-resolution magnetic spectrometer allowed hydrogen and helium spectral measurements up to 1.2 TV [4], the highest limit ever achieved by this kind of experiment.

2. Results obtained in 2013

Antiparticles measurement is the major goal of this experiment. Previous PAMELA measurements of the positron fraction between 1.5 and 100 GeV [2, 3] revealed the first clear deviation from secondary production models, showing indeed an unpredictable increase of the positron fraction above 10 GeV. New PAMELA data concerning the pure positron flux were presented in [5]. The positron data shows an increase in the positron flux at high energy, excluding the possibility that the raise in the ratio could be due to a decrease in the electron flux.

Hydrogen and helium isotopes in cosmic rays are generally believed to be of secondary origin. These isotopes can be used to study and constrain parameters in propagation models for galactic cosmic rays. The results presented in [6] are based on the data set collected by PAMELA between July 2006 and December 2007. About 5 million hydrogen nuclei were selected in the energy interval between 100 and 600 MeV/n and almost two billions helium nuclei between 100 and 900 MeV/n. PAMELA results are the most precise to date. Considering the relatively large spread in the existing data, PAMELA results agree with previous measurements.

Protons are the most abundant species in cosmic rays. Cosmic-ray particles reaching the Earth are affected by the solar wind and the solar magnetic field that modify their energy spectra (*solar modulation*). Solar activity varies with a 11 years cycle. Protons with rigidities up to at least 30 GV are affected and the effect becomes progressively larger as the rigidity decreases. Precise measurements of cosmic-ray spectra over a wide rigidity range, from a few hundred MeV to tens of GeV over an extended period of time can be used to study the effect of solar modulation in greater detail. The analysis presented in [7] is based on data collected by PAMELA between 2006 July and 2009 December. The analyzed period covers the unusually long most recent solar minimum. The large proton statistics collected by the instrument allowed the proton flux to be measured with unprecedent detail. From the data it is clear that protons reached maximum intensities at the end of 2009.

3. PAMELA data handling and processing

The radio link of the Resurs-DK1 satellite can transmit data about 2-3 times a day to the ground segment of the Russian Space Agency (Roskosmos) located at the Research Center for Earth Operative Monitoring (NTs OMZ) in Moscow. The average volume of data transmitted during a single downlink is about 6 GBytes, giving an average of 15 GBytes/day. In NTs OMZ the quality of data received by PAMELA is verified and faulty downlink sessions can be assigned for retransmission up to several days after the initial downlink. As soon as downlinked data are available they are automatically processed on a dedicated server in order to extract "QuickLook" information used to monitor the status of the various detector subsystems. In case some anomaly emerges, suitable commands can be sent from NTs OMZ to the satellite to change acquisition parameters, switch on/off part of the detectors, reboot the on-board CPU, etc.

After this preliminary data analysis, raw data are copied through a standard internet line to a storage centre in the Moscow Engineering Physics Institute (MePhI). From here, *Grid* infrastructure is used to transfer raw data to the main storage and analysis centre of the PAMELA Collaboration, located at CNAF. In CNAF raw data are written to magnetic tape for long-term storage and an automated "real-time" data reduction procedure takes place. The first step comprises a software for the extraction of the single packets associated to the different PAMELA subdetectors from the data stream: they are unpacked, organized inside ROOT structures and written on files. These files are afterwards scanned by a second program in order to identify "runs", i.e. groups of consecutive events acquired with a fixed trigger and detector configuration, which can correspond to acquisition times ranging from some minutes to about 1.5 hours. This step is necessary since the order of events inside data files is not strictly chronological, due to the possible delayed retransmission of faulty downlink sessions. Along all the described processing procedure, some information about data (e.g. the timestamps of the runs, the association between each run and its calibration data, the location of the files on disk, the satellite position and orientation data, etc.) is stored in a MySQL database hosted on an a dedicated server in CNAF. This database is then used in the final and most time consuming stage of the data reduction in which physical information for the particles registered in each event is calculated, all the events belonging to each run are fully reconstructed, calibration corrections are applied, and single runs are merged together to form larger files containing 24 hours time periods.

The aim of the real-time data reduction at CNAF is twofold: to make available as soon as possible reconstructed events for the analysis of interesting transient phenomena, such as solar flares, and to provide processed files that can be used to extract improved calibration information for the full data reduction. This longer procedure is performed periodically, usually once every 1-2 years, and takes place both in CNAF and in the computing farms of some of the INFN sections (Firenze, Napoli, Trieste) and of other institutions participating to the PAMELA experiment, where part of the raw data are periodically copied to.

References

[1] Adriani O et al, PRL, 102, 051101 (2009).
[2] Adriani O et al, Nature, 458, 607 (2009).
[3] Adriani O et al, Astropart. Phys. 34, 1 (2010).
[4] Adriani O et al, Science, 332, 69 (2011).
[5] Adriani O et al, PRL, 111, 081102 (2013).
[6] Adriani O et al, ApJ, 770, 2 (2013).
[7] Adriani O et al, ApJ, 765, 91 (2013).

The SuperB project at the INFN CNAF Tier1

F. Bianchi
INFN and University of Torino, via Giuria 1, 10135 torino, Italy

E-mail: fabrizio.bianchi@to.infn.it

Abstract. The SuperB project is outlined and its computing model is described with a focus on the role of CNAF as Tier1 centre.

1. Introduction

The SuperB project aimed at the design and at the construction of an asymmetric $e^+ e^-$ collider with a luminosity in excess of $L = 10^{36}$ $cm^{-2}s^{-1}$ and of an high performance detector optimized for discovering New Physics effects at the high luminosity frontier. The "Cabibbo Lab", near the campus of the University of Roma II in Tor Vergata, was selected to host the collider.
The SuperB detector was expected to produce in excess of 500 PB of raw data in 5 years of data taking. The event reconstruction step and the Monte Carlo production would have resulted in 300 additional PB of data. Predictable progress in computing technology would have provided the performance increase to cope with those data volumes. In addition, effective exploitation of computing resources on the Grid, that has become well established in the LHC era, would have made possible to access a huge pool of world wide distributed resources.
The SuperB project was terminated at the end of 2012.

2. Computing Model

The data processing strategy of the BaBar experiment has proven to be quite successful in handling the data volume generated by a flavor factory in the $L = 10^{34}$ $cm^{-2}s^{-1}$ luminosity regime. A similar strategy is expected to work well also for SuperB and can be summarized as follows.
The "raw data" coming from the detector would have been permanently stored, and reconstructed in a two-step process:

- a "prompt calibration" pass performed on a subset of the events to determine various calibration constants.
- a full "event reconstruction" pass runned on all the events using the constants derived in the previous step.

Reconstructed data also would have been permanently stored and data quality monitored at each step of the process. In addition to the physics triggers, the data acquisition would have recorded random triggers used to create "background frames".

Monte Carlo simulated data, incorporating the calibration constants and the background frames would have been generated and processed in the same way as the detector data. The amount of Monte Carlo simulated data was expected to be of the same order of the detector data.
Reconstructed data, both from the detector and from the simulation, would have been stored in two different formats:

- the Mini, that contains reconstructed tracks and energy clusters in the calorimeters as well as detector information. It could be a relatively compact format, through noise suppression and efficient packing of data.
- the Micro, that contains only information essential for physics analysis.

Detector and simulated data would have been made available for physics analysis in a convenient form through the process of "skimming". This involves the production of selected subsets of the data, the "skims", designed for different areas of analysis.

SuperB has exploited distributed computing resources using Grid technology. At the end of 2012, the SuperB virtual organization was enabled on 27 sites in 6 different countries and actively using pledged resources and non-pledged ones in parasitic mode.
In Italy, the pledged resources were located at the CNAF computing center in Bologna and in 4 new centers in Bari, Catania, Cosenza, and Napoli funded by the PON ReCaS. As a consequence, only limited computing resources would have been needed at the experiment site for data acquisition and for the calibration pass of the raw data processing that cold have provided a quick feedback on the detector status. Raw data would have been sent to CNAF and to another center for permanent storage on tape, and to the ReCaS centers for the full reconstruction step. Disk buffers would have been needed at the Cabibbo Lab to cope with network failures and at the ReCaS sites to host the raw data for the first processing and, after the first year of data taking, for the reprocessing of the data collected in the previous years. In this scheme, Cabibbo Lab, CNAF and the ReCaS centers would have had the functionalities of a distributed Tier0.
Monte Carlo production would have been performed at all classes of Tier sites, while skimming would have been done only at Tier0 and Tier1.
Multiple copies of real and Monte Carlo events in Mini and Micro format together with the collections of skimmed events would have been stored on disk at Tier1 and Tier2 for physics analysis.

3. The Role of CNAF

Until the end of the project, CNAF hosted the largest fraction of the storage (up to 300 TB) and of CPU pledged for SuperB. It hosted also the user areas for the physics and detector studies. The collaborative tool (web portal, wiki, Alfresco document repository) and the software svn repository were distributed between Padova, Ferrara and CNAF.

In 2013 the collaborative tools have been consolidated at CNAF and in March 2014 also the svn repository has been ported to CNAF. The user areas have been stored on tapes and the subset requested by the users will stay available on disk until needed.

4. Conclusions

CNAF has played a central role in the computing of the SuperB project. Even now, one year after the end of the project, it still provides services to support the remaining legacy studies.

The Virgo experiment. Report for the CNAF

Pia Astone, Data analysis coordinator of the Virgo experiment and Alberto Colla, Virgo contact person at CNAF.

Virgo experiment: `https://wwwcascina.virgo.infn.it/`

E-mail: `pia.astone@roma1.infn.it,alberto.colla@roma1.infn.it`

Abstract. We give here a general description of the computing strategy of the Advanced Virgo (AdV) project, based on the experience and work done for the Virgo project. We will focus in particular on the computing aspects related to the use of resources at CNAF, to show how the usage of CNAF has and will be important for our searches.

1. Introduction: description of the project

The Virgo experiment is based on a kilometer-scale gravitational wave detector, which has taken scientific data from May 2007 to Sept. 2011, over four runs (VSR1, 2, 3, 4). It is part of a wide network of detectors, so far with the two LIGO detectors at Hanford and Livingston. In the future the network of advanced detectors will include also the Japanese Kagra detector and farther on the horizon the IndIGO Indian detector. The first generation runs have been completed and the detectors have been shut down to undergo major upgrades aimed at reaching a much improved sensitivity. The new project is named "Advanced Virgo" (AdV) detector and the expected sensitivity, together with prospects for the detection of gravitational waves (GW), have been reported in [3]. The analysis of the first generation detector data for many different searches has been done and results published, while other analysis are still ongoing. We are now taking the opportunity of the transition from first to second generation experiments (AdV and aLIGO) to improve the computing framework of the experiment, when needed. The new Computing Model (CM) has been written with the experience done with Virgo and trying to enhance the usage of our internal resources.

2. Major past accomplishments

As said, many analysis on the data of first generation detector have been completed and some (not less important) are still ongoing. In particular the searches for continuous persistent signals, like those emitted from unknown isolated neutron stars, are still ongoing as they require long integration time in order to increase the signal-to-noise ratio and huge computational resources to explore a large parameter space. So far, we have not detected GW., as expected given the sensitivities, run time and theoretical predictions [3]. But the results we have obtained are indeed important, not only as we have gained experience by learning how to deal with real data and how to optimize data analysis algorithms, but also as the upper limits we have put in some cases do constraint the physical parameters of possible sources. The scientific goals have been divided into four main groups, reflecting the features of expected signals: CBC (search for compact binary coalescences), Burst (search for unmodeled transients), CW (search for

continuous waves), STOCH (search for stochastic background of cosmological or astrophysical origin). In the following there is a selection of some recent results, to give an idea of the different GW searches accomplished. The titles of the papers are quite explicative of the scientific goal behind the search.

- "Search for Gravitational Waves from Low Mass Compact Binary Coalescence in LIGO's Sixth Science Run and Virgo's Science Runs 2 and 3" Phys. Rev. D 85, 082002 (2012)
- " Parameter Estimation for Compact Binary Coalescence Signals with the First Generation Gravitational-Wave Detector Network ". Phys. Rev. D 88, 062001 (2013)
- "Search for Gravitational Waves from Binary Black Hole Inspiral, Merger and Ringdown in LIGO-Virgo data from 2009-2010 ". Phys. Rev. D 87, 022002 (2013)
- "All-sky Search for Gravitational-Wave Bursts in the Second Joint LIGO-Virgo Run " Phys. Rev. D 85, 122007 (2012)
- "Search for long-lived gravitational-wave transients coincident with long gamma-ray bursts" Phys. Rev. D 88, 122004 (2013)
- "First Searches for Optical Counterparts to Gravitational-Wave Candidate Events " Astrophys. J. Supp. 211, 7 (2014)
- "Constraints on cosmic (super)strings from the LIGO-Virgo gravitational-wave detectors " to appear in PRL
- " Gravitational waves from known pulsars: results from the initial detector era " To appear in Astrophysica Journal.
- "Implementation of an F-statistic all-sky search for continuous gravitational waves in Virgo VSR1 data " Submitted to CQG

3. The overall computing strategy

The overall computing strategy for Advanced Virgo has been described in the AdV Computing Model [1] and details on the technical implementation are being described in the Implementation Plan [2]. Virgo (and AdV) has a hierarchical model for data production and distribution: different kinds of data are produced by the detector and firstly stored at the EGO site in Cascina, where the detector is. There is no permanent data storage in Cascina but we foresee to install a disk buffer of 6 months of data acquisition for local access. The external CCs receive a copy of the data and provide storage resources for permanent data archiving. They must guarantee fast data access and computing resources for off-line analyses. Finally, they must provide the network links to the other AdV computing resources. For this goal a robust data distribution and access framework (based on file and metadata catalogs) is a crucial point.

The collaboration manages also smaller CCs used to run part of some analyses, simulations or for software developments and tests.

During science runs the Cascina facility is dedicated to data production and to different detector characterization and commissioning analysis, which have the need to run "on-line" (with a very short latency, from seconds to minutes, to give rapid information on the quality of the data) or "in-time" (with a higher latency, even hours, but which again produce information on the quality of the data within a well defined time scale). The detector characterization activity gives support to both commissioning and science analysis.

Science analyses are carried out only off-line at the external CCs, with the only exception of the low-latency searches. Low-latency searches are run with a very small latency after the data taking, of the order of minutes, to provide alerts to EM partners. Thus this analysis runs in Cascina, at the EGO site. The same analysis can be then repeated offline, using CNAF or CCIN2P3 resources, to refine the explored parameter space or to use a new version of calibrated

data of the detector. Some analyses, due to the fact that we analyze data jointly with aLIGO for many searches, are carried in LSC CCs.

To face the huge computational demands of GW searches in the advanced detectors era (ADE), there will be the need to gather the resources of many CCs into a homogeneous distributed environment (like Grids and/or Clouds) and to adapt the science pipelines to run under such distributed environment.

Another very important need for ADE is to provide a Grid-enabled, aLIGO-compatible Condor cluster for AdV people. In fact one bottleneck that has been identified in the past years was the difficulty to run at CNAF and CCIN2P3, which have been the Virgo external Computing Centers (CCs) from the very beginning, pipelines that were initially developed to run on LSC clusters and had shown a tight dependency on their architecture (Condor based submissions). The first attempt (2011/2012) to face these problems was the implementation of a submission system at CNAF based on the pilot job framework, which creates a virtual Condor cluster on a Virgo farm to which jobs can be submitted from a user interface. Another alternative which we have more recently (2013) succesfully tested at CNAF is the Pegasus Workflow Management System [http://pegasus.isi.edu/], which provides a layer for the job submission in different Grid environments.

Another important task, which we started to face, is the possibility in ADE to run some search pipelines in GPU clusters.

Most GW searches require the use of a network of detectors (at least AdV and aLIGO). As a consequence, these search pipelines must be able to run either in AdV or aLIGO CCs. It is therefore important to develop pipelines adaptable to different environments or interfaces which hide the different technologies to the users.

Thus the most important issues of the AdV CM may be summarized as follows:

- guarantee adequate storage and computing resources at Cascina, for commissioning, detector characterization and low-latency searches;
- guarantee fast communications between Virgo applications at Cascina and aLIGO CCs/other detectors for low-latency searches;
- guarantee reliable storage and computing resources for off-line analyses in the AdV CCs;
- push towards the use of geographically distributed resources (Grid/Cloud), in external CCs and whenever appropriate;
- push towards a homogeneus model for data distribution, bookkeping and access.

Figure 1 gives a big picture of the data workflow for what concerns scientific data analysis (DA) and detector characterization (Detchar) activities for AdV. In the picture CC2 indicates the CNAF, CC1 indicates CCIN2P3. Possible additional CCs have also been indicated, as a resource to perform intensive data analysis computation on the most important scientific data channels (which amounts to a really negligible storage need/year).

4. The role of CNAF

The computing usage and needs at CNAF is described in internal documentation which we prepare at the end of each year, to plan the needed resource for the next year[4, 5].

Over the last year CNAF has mainly been used for:

- Parameter estimation and General Relativity tests by the CBC group
- Science data preconditioning work by the Burst and Noise studies group
- All-sky searches for unknown isolated neutron stars by the CW group
- Narrow-band searches for isolated neutron stars by the CW group
- Optimazion studies for the CBC low-latency pipeline

Most of these use the GRID.

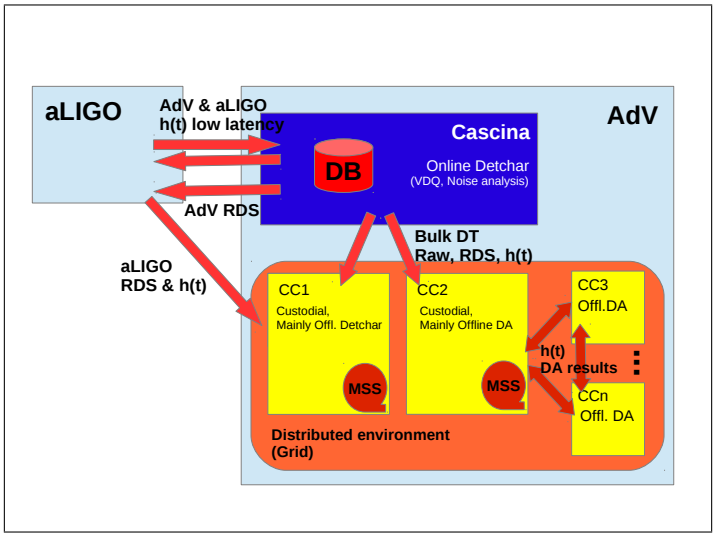


Figure 1. Data workflow for DA and Detchar activities in AdV. CC2 indicates the CNAF. CC1 indicates CCIN2P3.

4.1. Storage

Table 1 shows the storage at CNAF by the year 2009 up to the end of 2013.

Year CNAF	gpfs4 [TB] used / available Virgo	gpfs3 [TB] used / available Virgo	Castor or GEMSS [TB]	Castor disk [TB] used / available all exp.
2009	190 / 256	9 / 16	145 (Castor)	(+)
2010 (Oct. 1)	261 / (256+186)=442	16 / 16	163 (Castor)	17 / 36
2011	345 / 384	26 / 32	750	0
2012 (Oct. 29)	325 / 368	33 / 48	826	0
2013 (Nov. 18)	254/ 379	67 / 48	826	0

Table 1. Storage at CNAF since 2009. (+) means that we don't know the exact number. In 2011 data from Castor have migrated to GEMSS, which uses gpfs_virgo4 as cache disk.

4.2. Computing

CNAF accounting system is providing[1] information in wct_hep_day and cpt_hep_day[2]. Current computing consumption at CCIN2P3 are reported[3] in HSE06.hours (CPU time not wall clock time). Current consumption for 2013 is given below, while Table 2 shows the evolution since 2007 of the CPU consumptions. We mainly use "Wall-clock time" as this is the quantity we use to account to CNAF.

- **CNAF (date: January, 1st - November, 20 2013)**

```
Wall clock time (kHS06.day):  770 (Total)
```

[1] http://tier1.cnaf.infn.it/monitor
[2] 1 hep_day = 1 HS06.day
[3] http://cctools.in2p3.fr/mrtguser/info_sge_rqs.php?group=virgo

year	CNAF (WCT) [kHSE06.day]
2007	60
2008	240
2009	453
2010	162
2011	674
2012	669
2013	770

Table 2. Evolution since 2007 of the CPU used at the CNAF

The most intensive usage of CPU resources has been done by all-sky CW search and by the CBC Parameter Estimation and General Relativity work, in preparation of advanced detector era (ADE).

Some detector characterization work is also performed at CNAF on VSR2 & VSR3 raw-data set. Very recently some resources are being used at CNAF to develop new improved features of the CBC low-latency CBC pipeline.

During the year 2013, an important work has been done to do the porting of CBC pipelines from an LSC related submission method to an architecture complaint also with our CCs and in particular with GRID. This has solved the limit to run CBC analyses only on LSC clusters, opening new possibilities in particular in view of ADE, to both the Virgo and LSC collaborations. Tests on real data have begun in September 2013 and to allow them to run soon CNAF has granted to Virgo a number of cores O(1000) since September which was the minimum needed to prepare the CBC analysis in view of ADE and to run some new CW analysis (on VSR2/VSR4 data) enlarging the parameter space covered so far. To continue this analysis and to run new searches in the CW group we have then made an official request for 1000 cores to be assigned to Virgo in the year 2014.

5. General remarks

The CNAF support to the VIRGO experiment is not limited to the technical access of the CC facilities. The CNAF expertise was crucial for driving the discussion in the collaboration to define our computing model and focus on right solutions. This kind of support is even more important than the access to their hardware infrastructure. In addition, CNAF support will be fundamental for testing the porting on the GPUs located at CNAF some of most computing demanding searches of gravitational wave signals. Finally, we note that in the near future of the Advanced detector era, our computing needs will increase: we expect by the year 2018 a need for a continuous power O(100) kHS06, as we detailed in the CM. This implies that our impact on the CNAF infrastructure will be still below that of the HEP experiments at LHC, but not marginal any more.

References

[1] The Virgo collaboration, 2013 *The AdV Computing Model*
Virgo TDS num:"VIR-129E-13"
URL = https://tds.ego-gw.it/ql/?c=10325

[2] In preparation, The Virgo collaboration, 2014 *The AdV Implementation Plan.* TDS number: to be assigned

[3] The LSC and VIRGO collaboration, 2014 *Prospects for Localization of Gravitational Wave Transients by the Advanced LIGO and Advanced Virgo Observatories* arXiv:1304.0670

[4] The Virgo Collaboration, Virgo computing status and needs for 2013, "VIR-0413A-12", url =https://tds.ego-gw.it/itf/tds/file.php?callFile=VIR-0413A-12.pdf

[5] The Virgo Collaboration, Virgo computing status and needs for 2014, "VIR-0505A-13", url =https://tds.ego-gw.it/itf/tds/file.php?callFile=VIR-0505A-13.pdf

Xenon computing activities

G. Sartorelli, R. Persiani
INFN e Università di Bologna
E-mail: Gabriella.Sartorelli@bo.infn.it

1. The Xenon experiment

Astronomical and cosmological observations indicate that a large amount of the content of the Universe is made of Dark Matter. Although its presence is well established, its nature is still unknown. A possible candidate particle that arises in theories beyond the Standard Model is the Weakly Interacting Massive Particle (WIMP). The search for these particles is performed with a wide variety of experimental approaches.

The XENON dark matter project searches for the direct detection of nuclear recoils from WIMPs scattering off Xenon nuclei. Within the XENON program, we are operating and developing double-phase time projection chambers (TPCs) using liquid xenon (LXe) as target material with increasingly larger mass and lower background. A particle interacting with the target generates scintillation light and ionization electrons. The primary light (S1) is detected immediately by two photomultiplier arrays above and below the target medium. Ionization electrons are drifted upwards across the TPC by an electric field to the liquid-gas interface. A second electric field extracts such electrons from the liquid into the gas phase where they generate, by multiple interactions with Xenon atoms, a very localized secondary scintillation light (S2) proportional to the initial number of electrons. The S2 signal is used to determine the horizontal position of the interaction vertex, while the time difference between S1 and S2 signals gives the depth of the interaction in the TPC. The ability to localize events within millimeter resolution enables the selection of a fiducial volume in which the radioactive background is minimized. The simultaneous measurement of charge and light provides a powerful discrimination between nuclear and electron recoils through the ratio S2/S1.

The current phase, XENON100 installed in the interferometer tunnel of LNGS, is in science mode since the beginning of 2010. The detector is filled with a total of 161 kg of ultra pure liquid Xenon divided into two concentric cylindrical volumes. The inner target volume is a two-phase TPC containing a mass of 62 kg. It is optically separated from an instrumented active veto made of 99 kg of LXe. The data show that the detector is working extremely well in stable conditions. The data show also that the design goal of a very low background compared to the first prototype, XENON10, has been met and in 2012 we accumulated the exposure needed to explore, for the first time, a sensitivity to spin-independent WIMP nucleon cross section of $2 \times 10^{-45} cm^2$. Nowadays this limit is one of the best limits ever reached by a direct detection experiment.

In parallel to the successful operations of XENON100, we have already started to build the next generation detector: XENON1T. The detector is based on a LXe TPC with a total mass of 3.3 tons of ultra pure LXe and a fiducial mass of about 1.3 tons, viewed by low radioactivity 3 photomultiplier tubes and housed in a water Cherenkov muon veto. It is located in Hall B at LNGS. The experiment aims to reach an unprecedented sensitivity of $2 \times 10^{-47} cm^2$ and

for such achievement the background expected from all sources must be reduced by a factor 100 with respect to XENON100. For that purpose all detector components and materials are being screened and several Monte Carlo simulation are running to optimize the detector design and to evaluate the expected background level. Indeed the water shield and the LXe itself are very effective in reducing the external background; for the same reason the fraction of events that reach the fiducial volume of the experiment is very small. So, in order to reproduce the background behavior with sufficient statistics a huge computing power is needed: we are currently taking advantage of the GRID facility for our MC simulations.

2. Xenon computing

The current XENON100 computing infrastructure involves a dual core DAQ machine for data acquisition. It has a storage buffer of 1.1 TB to continue data taking in case of network issues. That machine is connected to the XENON computing facility at surface where data are transferred. Here the raw data are processed by a processing server with 32 CPU and then both raw and processed data are stored on 3 disks server with a total capacity of about 85 TB. Raw data are also stored in tapes as backup copy. Two machines with 8 cores each are dedicated to data analysis while another machine with 4 cores and 2.1 TB of disk space is devoted to host the web server, the Monte Carlo repository on SVN, the database and the XENON wiki. In the latest published scientific run (2011- 2012), XENON100 collected 225 days of Dark Matter search (light-weight data), 41 days of gamma calibration (heavy-weight data) and 7 days of neutron calibration (light weight data but needing prompt answers). Considering the data flow of 0.9 MB/s, 15 MB/s and 1.7 MB/s respectively for dark matter search, gamma calibration and neutron calibration, we produced 78 GB/day, 1.3 TB/day and 150 GB/day of data. For the latest scientific run, that means a total amount of data of 17 TB, 53 TB and 1 TB for dark matter search, gamma and neutron calibrations. The CPU-hours used in that run are divided as: 11k CPU-hours to process dark matter raw data, 61k CPU-hours for processing gamma calibration data and 3.5k CPU-hours for processing neutron calibration data. The total amount of resources used so far at LNGS are: 71 TB of raw data, 4.5 TB of processed data and 76k CPU-hours.

Scaling from XENON100 to XENON1T we expect to have a DAQ rate with higher data flow (up to 300 MB/s) and so we will need several hundreds of TB for the storage. The processing resource for the single event will be the same, but we have to scale to a much higher number of events; thus we will need hundreds of thousands of CPU-hours for data processing. More in details, we foresee to produce during one year of data taking: 50 TB of data for dark matter search and 500 TB for calibration (the last value could varying depending on the calibration sources we will adopt). For what concern the CPUh for each year, we expect to need about 20 kCPUh for dark matter data processing, 600 kCPUh for processing calibration data, 80 kCPUh for Monte Carlo studies and 44 kCPUh for data analysis, for a total of about 0.7 MCPUh.

XENON1T will start taking data in 2015 and for that time the computing facilities must be ready and well tested before the commissioning. The collaboration is moving to finalize the computing model for XENON1T. We definitely need a local facility underground, close to the detector, to handle data flow during limited times (network interruption or remote facility issues). For what concern remote facilities that will be used for most of the heavy work (processing, re-processing, analysis) the Collaboration has not yet made a final decision and several alternatives are under study. A very tempting possibility is to store all the processed calibration data on GRID and to move there all the analysis related to that kind of data. Given the large amount of data foreseen for calibrations, the over mentioned solution would be feasible only if a reasonable bandwidth to connect LNGS to GRID could be guaranteed (we are investigating the possibility to use the high-speed telecommunication network provided by the GARR consortium). All Monte Carlo simulations are currently running on GRID: during the first months of 2014 we

already used about 500 HS06 per day (the whole pledged computing power) and produced about 10 TB of data for the optimization of the detector design and the background evaluation. This item will last up to summer 2014. In the second part of the year we expect to run simulation for modeling the detector response, with similar needs in computing power.

The INFN-Tier1 Center and National ICT Services

The INFN-Tier1: a general introduction

L. dell'Agnello
Tier-1 coordinator

E-mail: luca.dellagnello@cnaf.infn.it

1. **Introduction**

Since 2003 CNAF hosts the main INFN computing centre, the so-called Tier1. It was initially designed to host the Italian Tier1 for the LHC experiments (Alice, Atlas, CMS and LHCb) but it has become the reference for the computing activities of a steadily increasing number of INFN experiments, thanks also to the complete renewal of the facility power and Heating, Ventilation and Air Conditioning (HVAC) systems in 2008. Currently, over 20 scientific collaborations use computing and storage resources hosted at Tier1, including experiments at accelerator facilities (the above mentioned LHC experiments, BABAR, CDF, AGATA, KLOE, LHCf and, formerly, SUPERB), astro-particle physics experiments (AMS, ARGO, Auger, Borexino, FERMI/GLAST, Gerda, ICARUS, MAGIC, PAMELA, Xenon100, VIRGO) and contacts are on going with others. The rapid growth of CPU and storage capacity in the last five years has been mainly driven by the start up of LHC, as the following table clearly shows.

Table 1 Tier1 computational and storage resources in 2008-2013

Year	2008	2009	2010	2011	2012	2013
CPU	[MHS06]	[MHS06]	[MHS06]	[MHS06]	[MHS06]	[MHS06]
LHC	11.3	11.3	42.5	60.7	85.0	88.1
CSN1	9.9	9.9	14.0	14.4	20.1	22.4
CSN2	1.6	1.6	9.7	12.0	15.8	18.3
Total	22.8	22.8	66.1	87.1	120.8	128.7
Disk storage	[PB]	[PB]	[PB]	[PB]	[PB]	[PB]
LHC	1.4	1.4	5.1	6.9	8.6	10.1
CSN1	0.5	0.5	0.6	0.8	1.1	1.1
CSN2	0.3	0.5	0.8	1.0	1.4	2.0
Total	2.1	2.4	6.6	8.7	11.0	13.2
Tape Storage	[PB]	[PB]	[PB]	[PB]	[PB]	[PB]
LHC	1.9	1.9	5.5	12.4	14.1	15.8
CSN1	0.0	0.0	0.0	0.0	0.6	0.6
CSN2	0.5	0.5	1.1	1.3	2.3	2.7

The INFN Tier1 has been designed to host the resources for the LHC experiment during their entire lifecycle, according with what is foreseen by the approved computing plans. The storage capacity, both disk (currently ~13 PB-N) and tape (currently ~17 PB), will be further increased in 2014 and in the following years, while for the computing resources (currently ~135 KHS06 corresponding to ~13000 job slots), after a large amount of replacements in 2013-2014 (a number of servers corresponding to about 40% of the currently installed CPU power), a sensible increase will happen in 2015.

An important goal of the CNAF computing centre is also to support all other INFN experiments fostering the adoption of standard interfaces also in collaborations with less computing experience.

In fact, the INFN Tier1 as part of the WLCG collaboration (Worldwide LHC Computing Grid [1], the computing infrastructure based on Grid technologies supporting the LHC experiments) and of the Italian and European Grid Infrastructures (EGI [2]) offers access to the resources through standard grid interfaces based on EMI middleware. In this framework, CNAF supports, in an opportunistic way, also other scientific collaborations.

The following figures (fig. 1,2) show the computing and storage resources installed in 2012-2013 and those foreseen for the next two years: the planning of the resources is updated yearly according to the experiments requirements.

On the long term a huge increase of resources after the end of LHC Long Shutdown 2 (2019) is expected, but the real plans are quite vague at this stage.

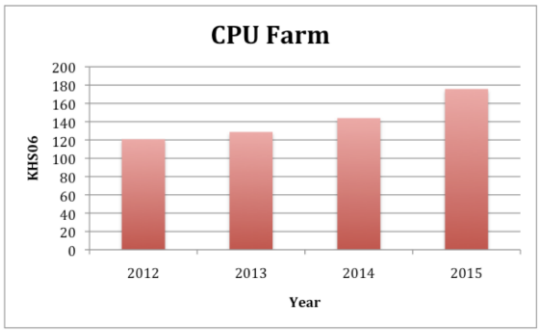


Figure 1 CPU farm power trend

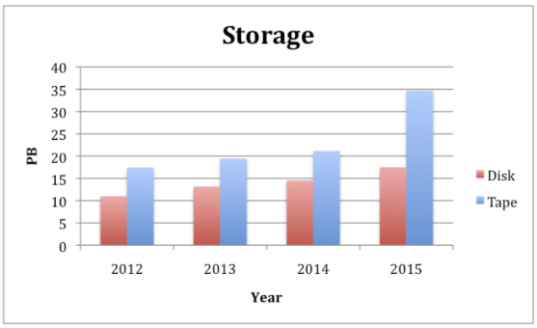


Figure 2 Disk and tape storage trends

In the datacentre, besides the Tier1 and a Tier2 for the LHCb experiment (logically separated but actually part of the Tier1 farm) also other resources are hosted: a Tier3 co-managed with the INFN Bologna unit, Italian Grid worker nodes and services as well as INFN national services managed by other CNAF groups.

2. Organization

The Tier1 operations are structured in 4 groups: the computing farm is managed by the Farming group, the Mass Storage System, the databases and the transfer service is managed by the Data Management group, the LAN and WAN connections of CNAF are managed by the Net group, and, finally, the facilities of the centre are managed by the Infrastructure group.

Overall 22 people (~ 21 FTE) have been working during 2013 at the INFN-Tier1 (see table 2).

Table 2 Manpower distribution at Tier1

	Permanent	Temporary	Total
Farming	**3**	**2**	**5**
Storage	**4**	**5**	**9**
Network	**2**	**1**	**3**

Infrastructure	**1**	**3**	**4**
Coordination	**1**	**0**	**1**
TOTAL	**11**	**11**	**22**

References

[1] WLCG – Worldwide LHC Computing Grid (http://wlcg.web.cern.ch/)
[2] EGI – European Grid Infrastructure (http://www.egi.eu/)

The INFN-Tier1: infrastructure facilities

L. dell'Agnello, A. Ferraro, A, Mazza, G. Bortolotti, M. Onofri

E-mail: luca.dellagnello@cnaf.infn.it

1. Data centre infrastructure

The computing centre is composed of 2 halls for IT resources hosting a total of 180 racks and one tape library (extra room is available for an additional tape library). The centre also hosts, in a dedicated third hall, the National Research Network (GARR) Point Of Presence (PoP).

The two main halls, based on the hot aisle containment system for rack cooling, have different characteristics:

- the Hall 1 hosts 71 racks, 59 for IT and 12 for in row air handlers, mainly grouped in 2 isles;
- the Hall 2 hosts 104 racks, 75 for IT and 32 for in row air handlers, grouped in 4 high-density isles.

The cooling system, so organized, can provide a cooling power of 200 kW for each isle in Hall 1 and up to 300 kW for each isle in the Hall 2, ensuring in any case a (N+1) redundancy per isle.

(5+1) free cooling chillers complete the cooling system: each one has a maximum thermal capacity of 320 kW (it is foreseen to add and additional chiller in case of need). At the present status, thanks also to a careful configuration and tuning of the chillers, no more than 4 chillers are in use at the same time, even during summer. The cooled water temperature is set to 15 °C (the highest temperature for these chillers): this choice allow us to benefit as much as possible of the free cooling during the cold seasons from one side (minimising though the work of compressors) and to guarantee a server room temperature of 23 °C yearlong, from the other side.

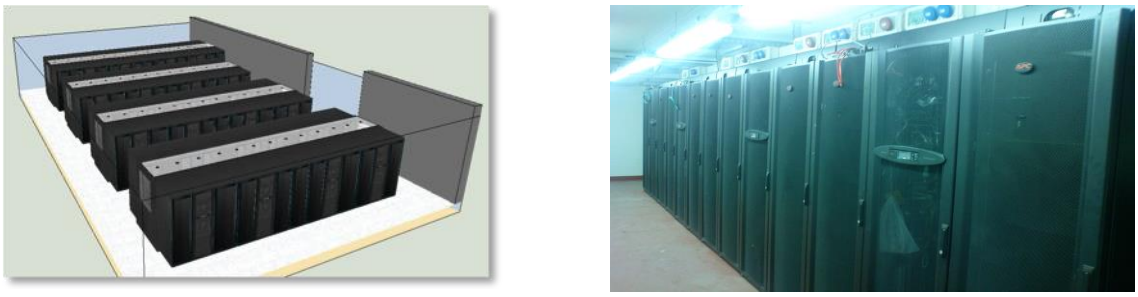


Figure 1. High-density racks in Hall 2: left a schematic view and, right, an external view

The electric power supply is also completely redundant: in first place the (2+1) transformers convert the electric tension from 15,000 V to 400 V and convey it through 2 independent power lines, each one able to provide up to 1.4 MW. Each power line is coupled to a rotary UPS as to guarantee an uninterrupted and harmonic filtered power supply, against both short and protract interruptions. During an electric interruption, after just few seconds, the diesel engine starts up able to take over all the electric load (see fig. 2).

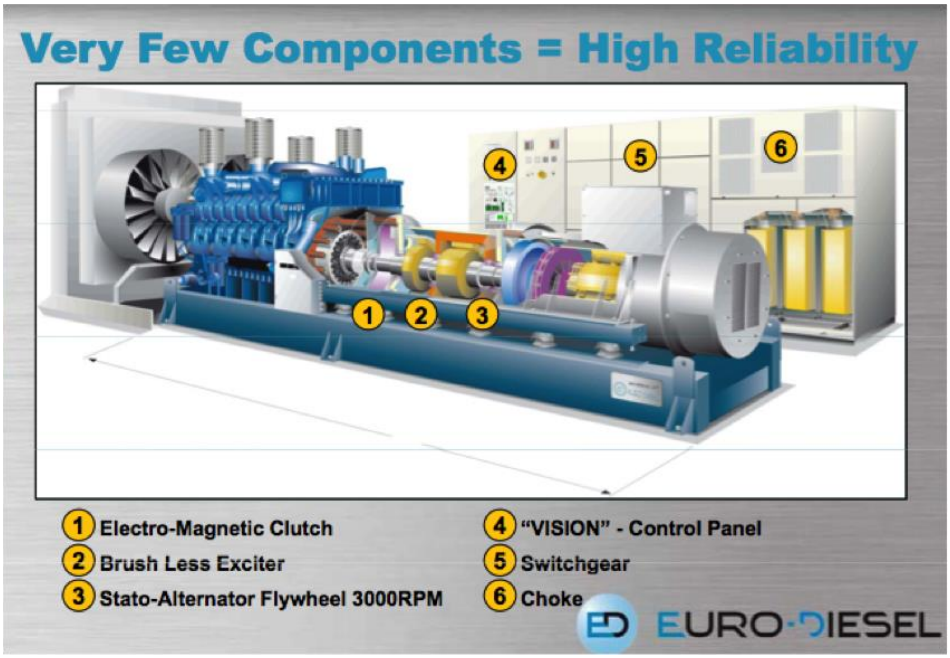


Figure 2. Rotary UPS and diesel engine group (one per each line)

The setup also includes an additional 1.2 MW diesel generator that could be used for supplying power to the chillers, if needed.

Both the power and the cooling systems can cope with significant future increase of the installed IT resources: at present the power breakdown shows ~700 kW for IT (out of 1400 kW, which is the maximum power that can be dissipated by the cooling system relying on N+1 redundancy), a peak of ~300 kW for the cooling system and ~130 kW for the rotary UPSs.

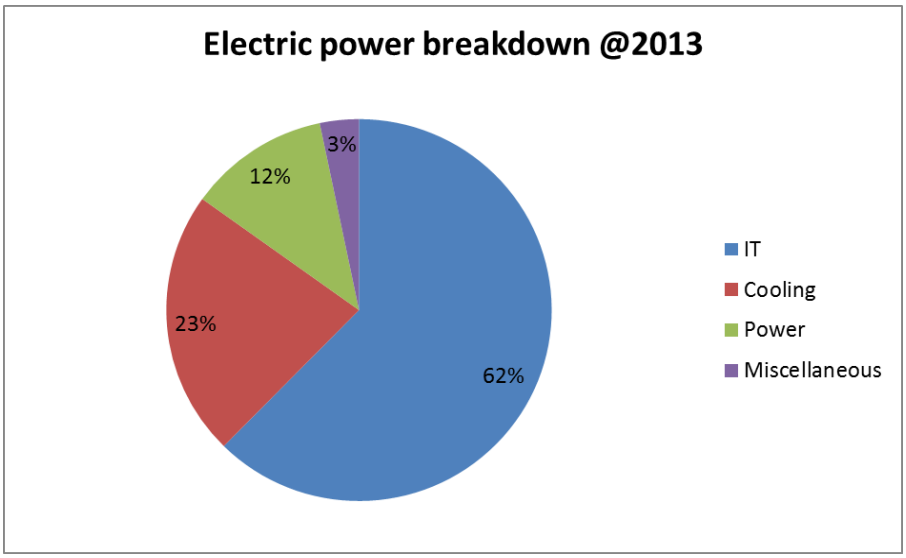


Figure 3. Tier1 Electric Power Breakdow for 2013.

From the rack usage point of view, with the natural phasing out of obsolete resources and the technological evolution of the servers (more computing power per consumed Watt), we will be able to cope also with a strong increase in capacity: at present there are 24 empty high-density racks in Hall 2

and 10 empty racks in Hall 1. Moreover, we are considering installing 8 additional racks (for a total dissipating power of ~120 kW) in an empty area characterized by high-energy efficiency (free cooling for at least 70% of the year).

The INFN-Tier1: Network

S. Zani, D. De Girolamo, L. Chiarelli, L. dell'Agnello

E-mail: stefano.zani@cnaf.infn.it

1. Introduction

The Network department manages the wide area and local area connections of CNAF, is responsible for the security of the centre and also contributes to the management of the local CNAF services (e.g., DNS, mailing, Windows domain etc.) and some of the main INFN national ICT services.

2. Wide Area Network

Inside CNAF datacentre is hosted the main PoP of GARR network, one of the first "Nodes" of the recent GARR-X evolution based on a fully managed dark fibre infrastructure.

CNAF is connected to the WAN via GARR/GEANT essentially with two physical links:

- General IP with a 10Gb/s connection via GARR and GEANT
- To WLCG destinations with a 20Gb/s link shared between the **LHC-OPN** Network for Tier0-Tier1 and Tier1-Tier1 traffic and **LHCONE** network for T2 and T3 traffic.

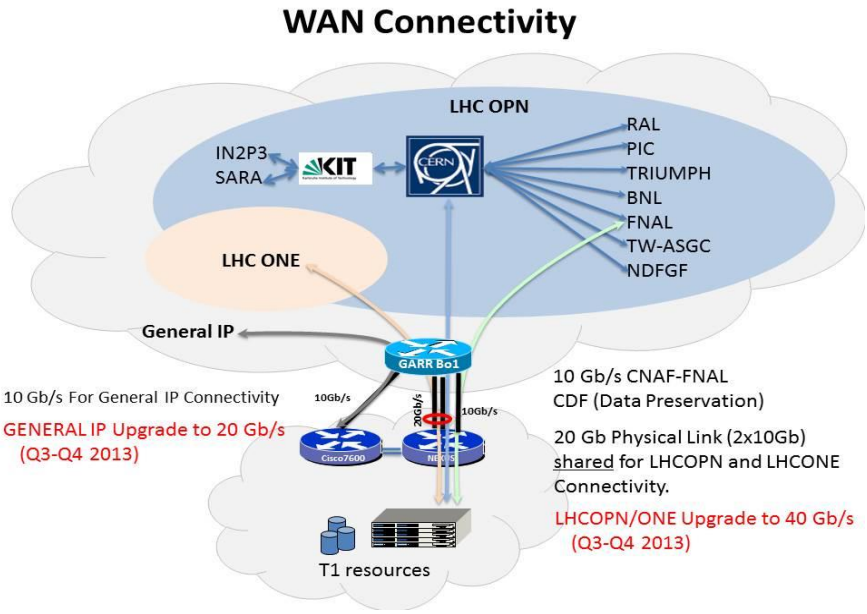


Figure 1 WAN connection schema

WLCG link is going to be upgraded to 40Gb/s (4x10Gb/s) at the beginning of 2014 and GARR has in its roadmap a 100Gb/s link between CNAF and CERN (end of 2014).

An additional temporary 10Gb/s link between FNAL and CNAF is used to transfer all the CDF data to CNAF for the long-term data preservation project.

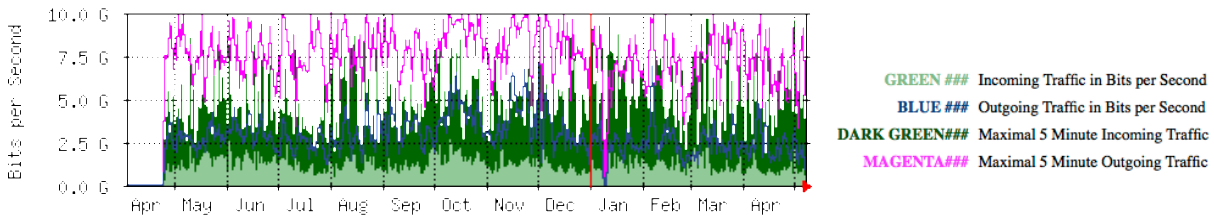


Figure 2 General IP bandwidth usage (yearly trend)

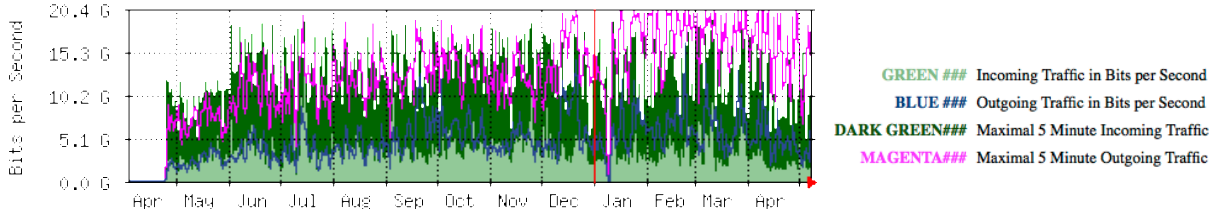


Figure 3 LHC-OPN/ONE bandwidth usage (yearly trend)

3. Local Area Network

The Tier1 LAN is essentially a star topology network based on a fully redundant Switch Router (Cisco Nexus 7018), used both as core-switch and Access Router for LHC-OPN and LHCONE networks, and more than 100 aggregation ("Top Of the Rack") switches with Gigabit Ethernet interfaces for the Worker Nodes of the farm and 10Gb Ethernet interfaces used as uplinks to the core switch.

Disk-servers and *gridftp* servers are directly connected to the core switch at 10Gb/s.

General Internet access, local connections to the offices and INFN national services provided by CNAF are managed by another network infrastructure based on a Cisco7606 Router, a Cisco Catalyst 6509 and an Extreme Networks Black Diamond 8810.

CNAF has an IPv4 B class (131.154.0.0/16) and a couple of C classes (for specific purposes): half of the B class is used for Tier1 resources and the other half is used for all the other services thus providing sufficient IP addresses. The private address classes are used for IPMI and other internal services.

Two /48 IPv6 prefixes are assigned to CNAF (2001:760:4204::/48 for CNAF General and 2001:760:4205::/48 for CNAF WLCG) and recently we have started the IPv6 implementation on LAN.

4. Network monitoring and security

In addition to the perfSONAR-PS and the perfSONAR-MDM [1] infrastructures required by WLCG, the monitoring system is based on several tools organized in the "Net-board", a dashboard realized at CNAF. The Net-board integrates MRTG[2], NetFlow Analyser [3] and Nagios **Error! Reference source not found.** with some scripts and web applications to give a complete view of the network usage and of possible problems.

The alarm system is based on Nagios.

The network security policies are mainly implemented as hardware based ACLs on the access router and on the core switches (with a dedicated ASICS on the devices).

The network group, in coordination with GARR-CERT and EGI-CSIRT, also takes care of security incidents at CNAF (both for compromised systems or credential and known vulnerability of software and grid middleware) cooperating with the involved parties.

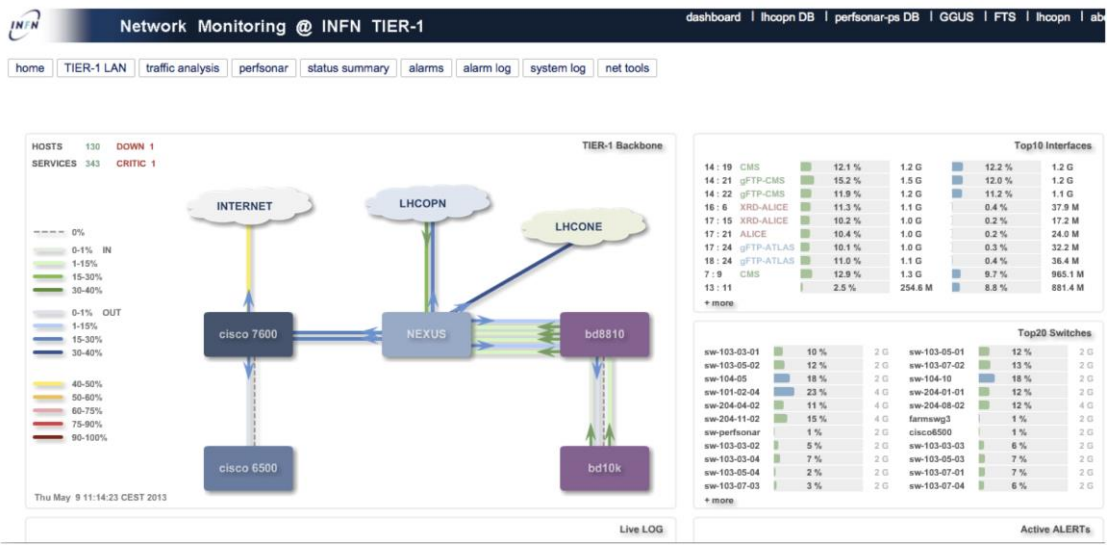


Figure 4 Net-board screenshot

References

[1] PerfSONAR (http://psps.perfsonar.net/)
[2] MRTG – Multi Router Traffic Grapher (http://it.wikipedia.org/wiki/Multi_Router_Traffic_Grapher)
[3] NetFlow (http://en.wikipedia.org/wiki/NetFlow)
[4] NAGIOS (http://www.nagios.org)

The INFN-Tier1: Data management

A. Cavalli, L. dell'Agnello, D. Gregori, A. Prosperini, P.P. Ricci. V. Sapunenko

E-mail: luca.dellagnello@cnaf.infn.it

1. Introduction

The Data Management group is responsible for the installation, configuration and operations of the Mass Storage System (MSS) including the Storage Area Network (SAN) infrastructure, the storage systems, the disk-servers (local, *gridftp* [1] and *xrootd* [2] servers) and the tape library. It is responsible for the INFN instance of FTS [3] (the WLCG [4] File Transfer System), and of administration of databases both for experiments and for other services. It also manages the SRM interfaces (StoRM [5] in our case) to the storage systems.

2. The storage system

The storage infrastructure is based on industry standards, both for connections (all disk servers and disk enclosures are interconnected through a dedicated Storage Area Network) and for data access (data is hosted on parallel file systems, typically one per major experiment). This allowed the implementation of a completely redundant data access system from a hardware point of view and capable of very high performances. Currently the aggregate bandwidth between the computing farm and the storage exceeds 70 GB/s, making it possible to extend the usage of the centre also to the user analysis that is typically more demanding for what concerns data access with respect to the organized productions.

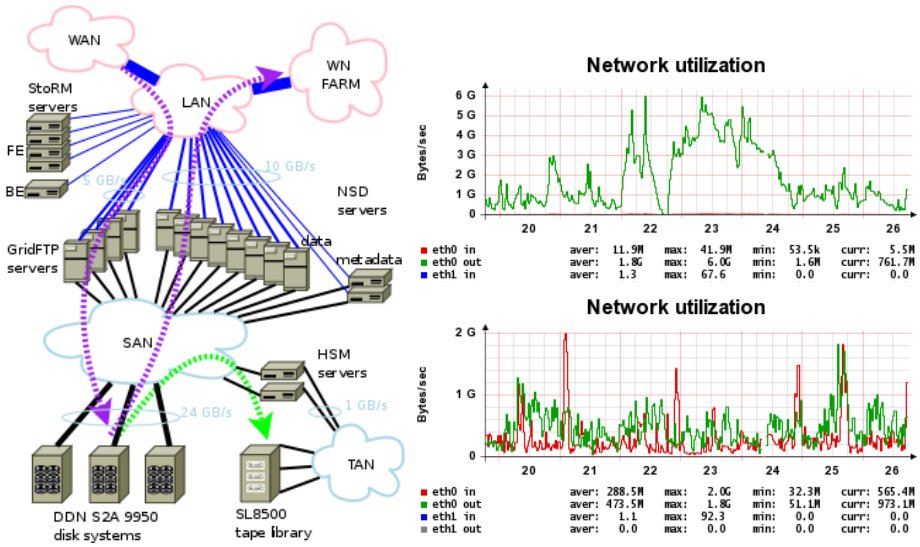


Figure 1 The Atlas case: (left) the storage system and (right) the performance on LAN (upper graph) and WAN (lower graph) over one week

Currently the Tier1 hosts 13 PB of net disk space and more than 17 PB on tapes.

The MSS used at Tier1, with all the functionalities of a Hierarchical Storage Manager (HSM), is **GEMSS** (Grid Enabled Mass Storage System) (see section 5).

Storage resources are accessible through standard protocols according to interfaces defined by the WLCG and EGI **Error! Reference source not found.** projects (i.e. SRM). Legacy interfaces, used mainly by small collaborations, are still supported.

3. The hardware components

The current amount of 13 PB of net disk space is based on several hardware systems with SATA2 disks. The oldest generation (~1.9 PB of net disk space in RAID 5), to be decommissioned at the end of this year, is composed by 7 **EMC^2** CX3-80 and 1 **EMC^2** CX4-960 served by 100 servers (each one with 2x1 Gbps connection to the LAN). The remaining part of the storage in production (~11.1 PB of net disk space in RAID 6) is composed by 7 **DDN** S2A 9950, 1 DDN SFA 10000 and 1 DDN SFA 12000 served by ~75 servers (each one connected to the core switch with a 10 Gbps link).

All disk-servers are interconnected to the storage systems, through the SAN, with 8 Gbps links (4 Gbps in case of the EMC^2 storage) and are equipped with a dual power supply.

Figure 2 - The Oracle/SUN SL8500 tape library

The tape library is an Oracle/SUN SL8500, fully redundant, with 10,000 active slots.

The choice of this library was originally driven by the HSM used in 2007 (CASTOR certified only for some SUN and IBM libraries). Nevertheless this library coupled with the T10k family drives offers the best scalability: the last generation of the drives (T10kc) has a capacity of 5 TB/slot, while with the next one (T10kd), available from Q1 2014, the capacity will increase, with the same cartridge, to 8.5 TB/slot.

In general, in the road map of Oracle, it is foreseen to have a technological jump of the drives every ~ 2 years and of the media every ~ 4 years. In this way, repacking at regular intervals, we are able to cope with the expansion of capacity required from the experiments.

At present we have a mix of 5 TB tapes (served by 10 T10kc drives each with a bandwidth of ~ 200 MB/s) and 1 TB tapes (served by 20 T10kb each with a bandwidth of ~ 100 MB/s). The drives are interconnected to the library and the servers via a dedicated SAN. 13 Tivoli Storage manager HSM nodes access to the shared drives.

In conclusion, the present library has total capacity of 50 PB that will increase to 80 PB in a few months.

4. The SAN model

The storage system at the Tier1 is completely based on a SAN built on several Brocade switches, including two Fabric Director Switches (one 24000 with 128 2GB/s ports and one 48000 with 4GB/s ports extendable to 256 ports).

The SAN has a star topology, centred on the two Director Switches, with several peripheral switches: 16 “blade” Brocade 4424 switches with eight 4 Gb/s ports each and 4 “mid-range class” Brocade 5300 switches with a total of 256 Fibre Channel 8 Gb/s ports. All these switches have been acquired as part of the storage tenders.

A portion of the “central” fabric switches is dedicated to the Tape Area Network (TAN): all the tape drives are directly connected to this central point of our SAN.

Currently we have in production a total of ~200 disk-servers with redundant Qlogic HBA connections to the SAN and 1-10 Gigabit (depending on the generation) connection to the LAN.

The double connection to the SAN from the servers, coupled with the path-failover mechanism, which implements a real load-balancing, allows eliminating several single points of failure (server connections, Fibre Channel switch or controller of the disk storage box).

The flexibility offered by the SAN with the decoupling of disk-servers from the disk boxes and the high efficiency of the infrastructure, allows a robust and performing implementation of clustered file-system like GPFS [6] (no single point of failure).

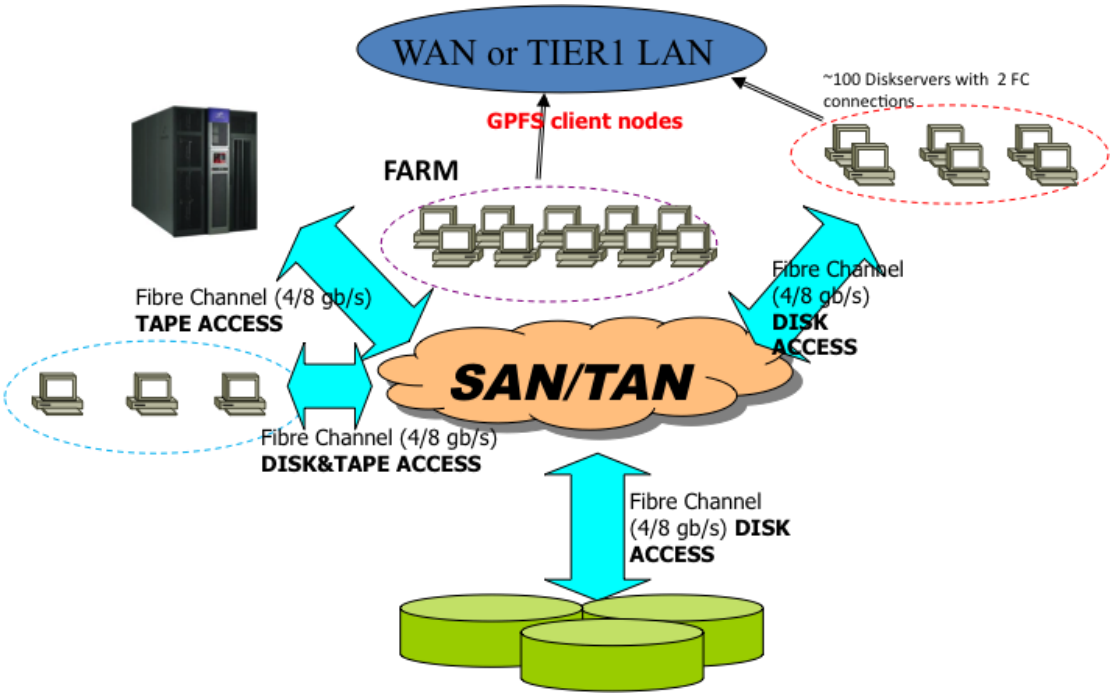


Figure 3. Schematic view of the Tier1 SAN

The SAN in production at CNAF has proven its stability during years of productions; the level of complexity in administration and monitoring of the whole system has been kept low by adopting a uniform infrastructure.

5. GEMSS

The Grid Enabled Mass Storage System (GEMSS) is a full HSM integration of General Parallel File System (GPFS), TSM (Tivoli Storage Manager) [8], both from IBM, and StoRM (developed at INFN); its primary advantages are the high level of reliability and the low management effort needed.

The original idea of GEMSS dates back to 2006, when we were searching for a performing and robust solution as an alternative to the CASTOR [9] system (from CERN) in production at that time. GEMSS underwent a step-by-step evolution:

- Q1 2007: after comparison tests among several cluster file-system GPFS was chosen for disk-based storage;
- Q2 2007: released StoRM (developed at INFN), implementing the SRM 2.2 specifications conforming to the WLCG requirements;
- Q3-Q4 2007: StoRM/GPFS in production for the disk-only storage for LHCb and Atlas (clear benefits for both experiments were reported);
- End 2007: start of realization of a complete grid-enabled HSM solution based on StoRM/GPFS/TSM, where the TSM system manages the tape area;
- Q3 2009: validation of GEMSS;
- Q4 2009: start of migration of experiments to GEMSS.

GPFS is a clustered (fault tolerance and redundant) and parallel file-system: it offers the possibility to have a single file hierarchy (or directory) seen by a set of clients, by aggregating the whole storage resources.

GPFS allows applications to simultaneously access the same files in a concurrent way, ensuring the global coherence, or even different files, from any node mounting a GPFS file system. Therefore GPFS is particularly appropriate in an environment where the aggregate I/O peak exceeds the capability of a single file system server.

The TSM software is dedicated for the tape access layers.

The main elements of the TSM system are the master TSM server and some HSM nodes, running the TSM Storage Agents., that are directly responsible for moving the data to and from the tape back-end. The TSM server is the core component: it relies on a database (replicated over the SAN) for the metadata information and it also provides the space management services to the HSM nodes. A cold stand-by machine is ready for replacing the main server in case of major failure. The TSM Storage Agents enable LAN-free data movements on the HSM nodes, using the dedicated TAN to communicate with the drives, greatly improving the performances avoiding traffic congestion on the LAN. In our present setup a total number of 13 HSM nodes are largely enough for providing all the data movements between disk and tape with optimal performance. Since the HSM nodes are independent from the disk area, it is possible to interrupt the tape access service for maintenance while keeping the disk service online.

The GPFS and TSM interaction is the main component of the GEMSS system: a thin software layer has been developed in order to optimize the migration (disk to tape data flow) and, in particular, the recall (tape to disk data flow) operations. While the vanilla TSM performs recalls file per file, GEMSS collects all the requests in a configurable time lapse and then performs reordering to minimize the number of mount/dismount operations in the tape library and unnecessary tape “seek” operation on tapes. The migrations from disk to tape are driven through configurable GPFS policies.

StoRM implements the SRM interface and it is designed to support guaranteed space reservation and direct access using native POSIX I/O calls to the storage. The main feature of StoRM is the possibility to take advantage of parallel file-systems like GPFS and others.

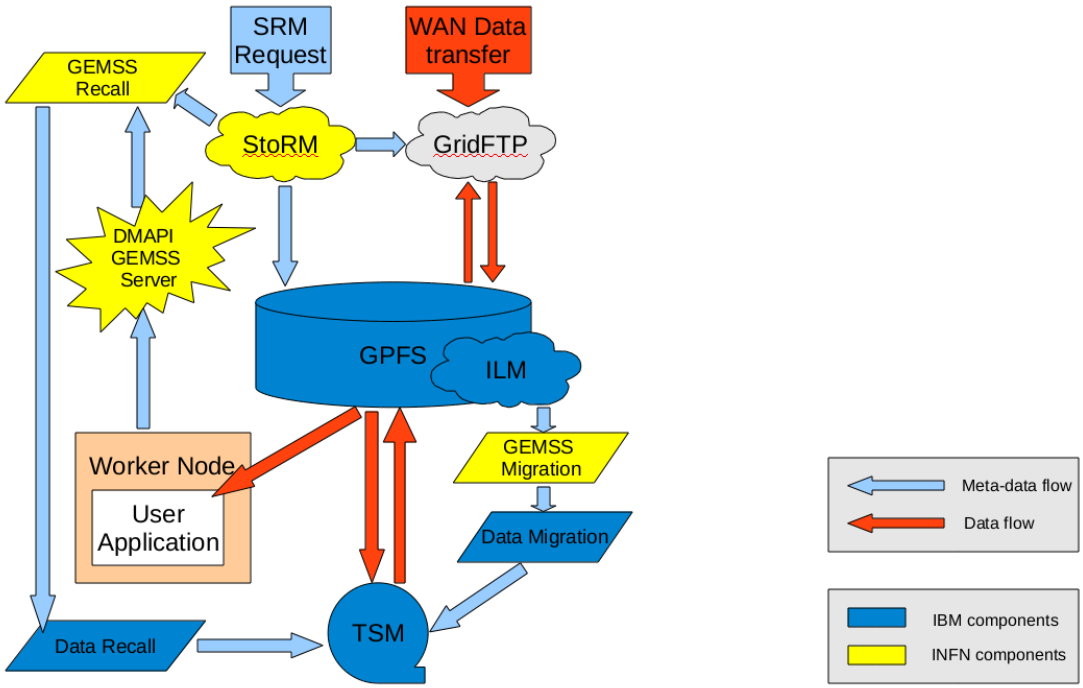


Figure 4 Schema of GEMSS architecture

5.1. GEMSS performances

GEMSS has proved to match nicely the performance required by the experiments, showing good scaling capability both in terms of file system size (currently up to 2.5 PB) and of the storage system data access throughput: currently the SFA 9500 storage system in production has a maximum bandwidth of ~ 11 GB/s which GPFS can nicely reach and sustain as shown in the figure below.

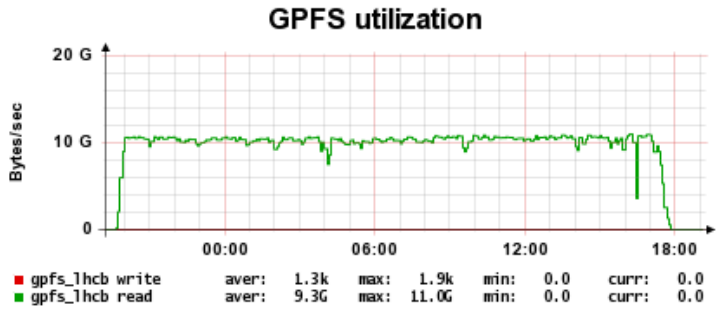


Figure 5 ~800 TB of LHCb data read from the 9500 DDN system

In the figures below the validation stress test of access to the tape system for CMS (September 2009) is shown: 24 TB of data stored in 10.000 files randomly spread over 100 tapes were moved from TSM to GPFS via GEMSS in 19h (no failures) using 6 drives and concurrently data were being migrated to tape (3 drivers used).

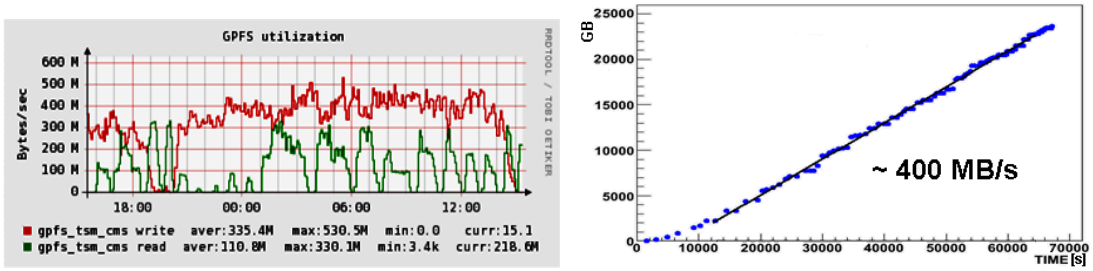


Figure 6 (left) Throughput to (green) and from (red) the tape system; (right) recalled data (GBs versus time)

In the event of exceptionally intense access, GEMSS has also shown excellent performance and robustness, as it was the case when a misconfigured CMS workflow was run at all CMS Tier1s (May 2011).

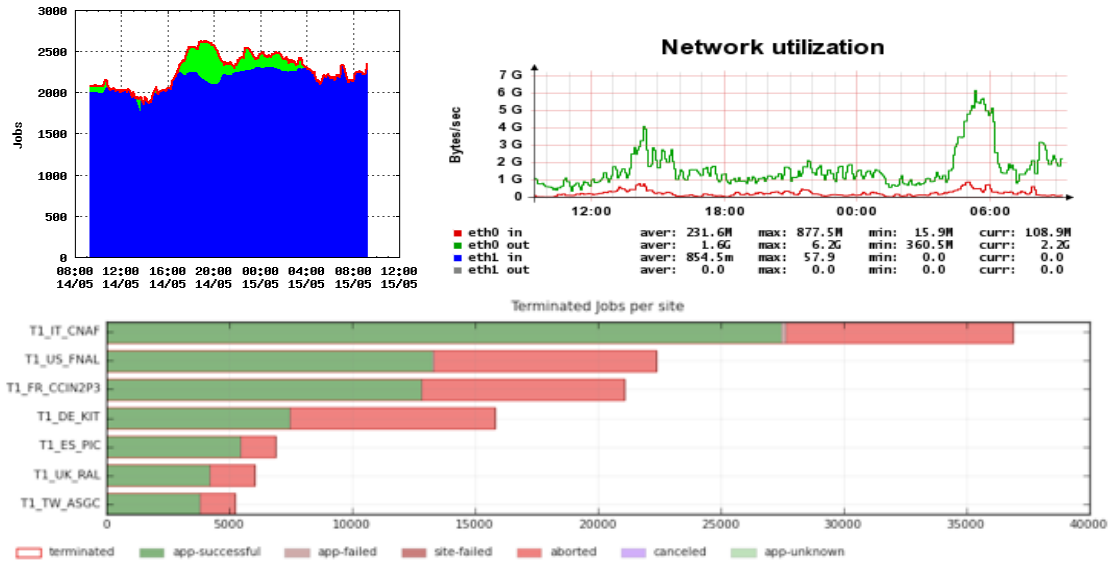


Figure 7 (Left) CMS running jobs (blue), (right) throughput from CMS storage to the farm and (bottom) comparison of the completed jobs in the CMS Tier1 sites

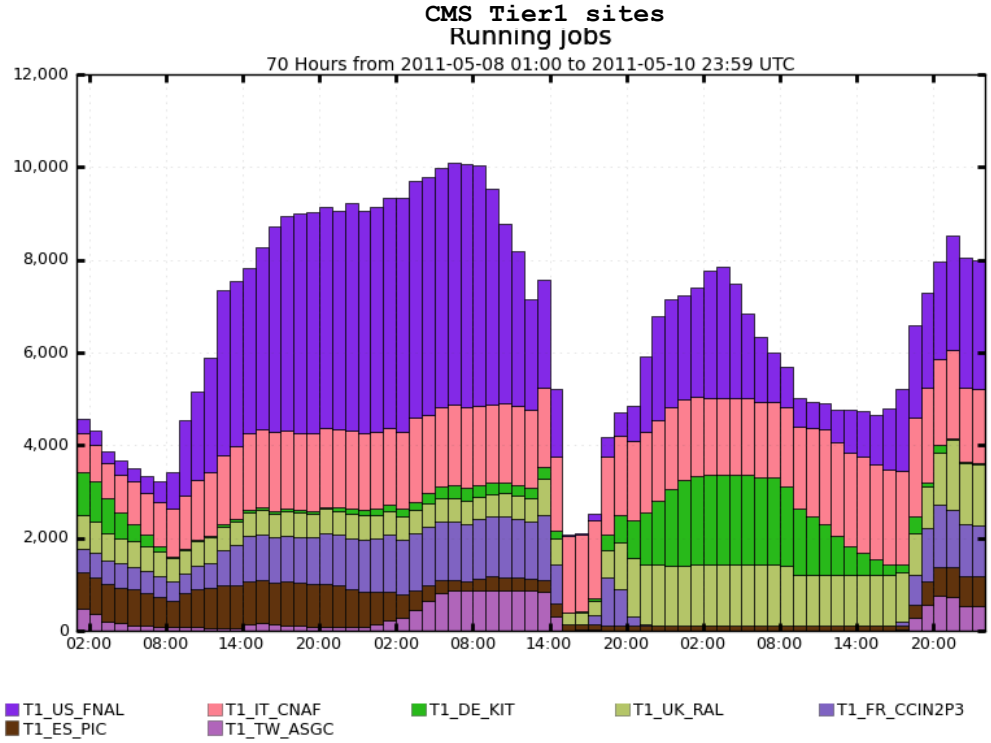


Figure 8 CMS jobs executed at all Tier-1 sites in May 2011. When a particularly IO intensive workflow was submitted by mistake, CNAF was the only site to cope with the load.

In this case, several jobs failed at all the other CMS Tier1s due to frequent crashes of the storage systems due to the very high load: 61% of the total completed jobs were run at CNAF. Also the efficiency was quite high, ranging in the interval 93-96%.

Frequently LHCb users perform analysis at CNAF accessing storage at high rate: in the figure below a typical access pattern is shown.

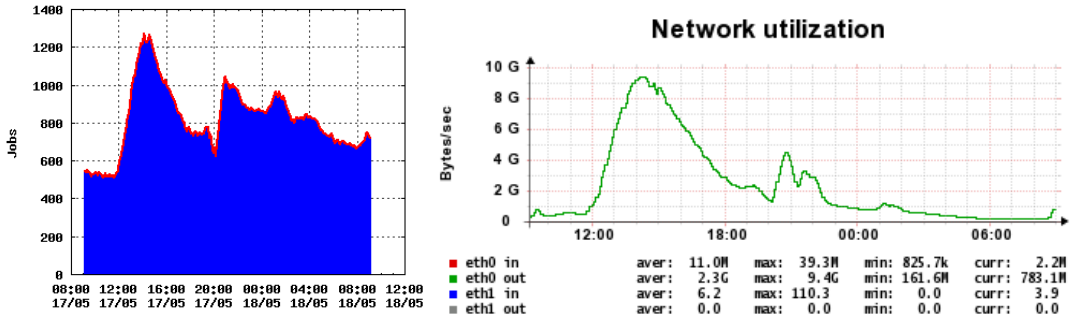


Figure 9 (Left) the LHCb running jobs (blue) and (right) throughput from LHCb storage to the farm (18/5/2013)

To be noticed that the number of TB served by a single disk-server is noticeably higher than other equivalent cases (see figure below) without loss of performances or robustness. This is due to the use of SAN and GPFS.

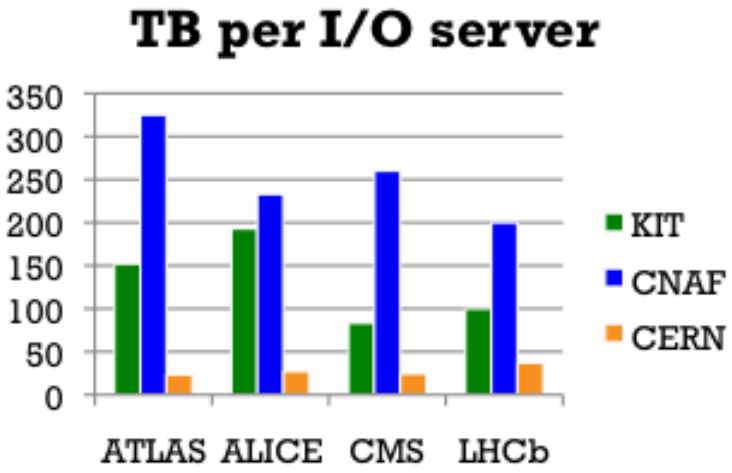


Figure 10 - TB served by a single disk-server at CNAF, KIT and CERN

6. Databases

The main goal of the database service is to provide high available, scalable and reliable Oracle instances. This has been fulfilled through a modular architecture based on two well-known Oracle technologies:

- Oracle Automatic Storage Management (ASM) volume manager, for the storage management implementation of redundancy and striping in an Oracle oriented way.
- Oracle Real Application Cluster (RAC) where the database is shared across several nodes with failover and load balancing capabilities.

In addition the hardware has been particularly selected using dual power redundant server with RAID1 (mirroring) system disks. These database servers use a dual path Fibre Channel layer to the main storage on a dedicated EMC3-80 with 25TB RAW of Fibre Channel technology disks and 10TB RAW space of SATA2 1TB disks. This storage system will be replaced at the end of this year.

The RAC clusters are designed in order to grant service in case of hardware/software failure or patches upgrade on single machines where the primary instance (database) run.

After the phasing-out of several databases (e.g. CASTOR, ATLAS LFC, etc.) the instances in use are mainly for specific service backend such as the FTS database, the Lemon Monitoring system database and the Oracle Grid Console.

7. Long Term Data Preservation

The CDF collaboration has recently started a project aimed at the long-term preservation of the experiment data (Long Term Data Preservation, LTDP) and of the capability to analyse it. The project got a first approval by INFN in September 2012 and we expect the funding will be available from June 2013.

The necessary setup for copying the data from Fermilab and storing it on tape has been setup. The entire operation will be performed in approximately 18 months with 4 PB of data being transferred at a nominal rate of ~150 MB/s.

In addition the LTDP project requires developing a system allowing users in the long-term future to access and use with efficiency the archived data. The design of such system is a complex project and can serve as a prototype for future experiments which are now supported by INFN and which will soon face the problem of long-term data archival. Moreover, INFN has submitted a proposal, together with other scientific institutions, for a wider scope LTDP project: in case of approval, the CDF case will then represent the prototype that will be useful to gain experience.

References

[1] Gridftp – (http://en.wikipedia.org/wiki/GridFTP)
[2] xrootd – (http://xrootd.org)
[3] FTS – File Transfer Service (http://egee-jra1-dm.web.cern.ch/egee-jra1-dm/FTS/)
[4] WLCG – Worldwide LHC Computing Grid (http://wlcg.web.cern.ch/)
[5] Storm - Storage Resource Manager (http://italiangrid.github.io/storm/index.html)
[6] EGI – European Grid Infrastructure (http://www.egi.eu/)
[7] GPFS - General Parallel File System (http://en.wikipedia.org/wiki/IBM_General_Parallel_File_System)
[8] TSM – Tivoli Storage Manager (http://en.wikipedia.org/wiki/IBM_Tivoli_Storage_Manager)
[9] Castor – CERN Advanced Storage Manager (http://castor.web.cern.ch)

The INFN-Tier1: the computing farm

A. Chierici, S. Dal Pra, L. dell'Agnello

E-mail: Andrea.Chierici@cnaf.infn.it

1. Introduction

The farming group is responsible for the management of the computing resources of the centre (including the grid interfaces, CE and site BDII). This implies the deployment of services to install and configure the resources as well as to monitor them and to distribute those resources fairly to the experiments that have agreed to run at CNAF. To perform these activities several software are required.

The installation is performed through QUATTOR, software developed by the European DataGrid [1] collaboration.

2. The computing farm

The computing resources of all the experiments are centrally managed by a unique batch system (typically each experiment has at least a dedicated queue) and dynamically allocated through a fair-share mechanism based on predefined policies, depending on the funding decided by the national scientific committee for each experiment. The centralized resource management system allows the full utilization of the available CPU power for about 95% of the time. In figure 1 the running jobs (blue area) and the waiting jobs (green area) are shown. The saturation of resources (currently ~180 KHS06 for a total of ~16000 job slots) is clearly visible: the step at July 2013 corresponds to the installation of 2013 resources while the holes correspond to scheduled interventions on the farm (middleware rolling updates, security patches and a complete scheduled down of the centre at the beginning of January 2013).

Worker Nodes (WN) have been migrated from Scientific Linux (SL) 5 to SL 6 and from EMI-2 [2] to EMI-3 middleware in the second half of 2013 together with the main Grid Elements (e.g. CEs).

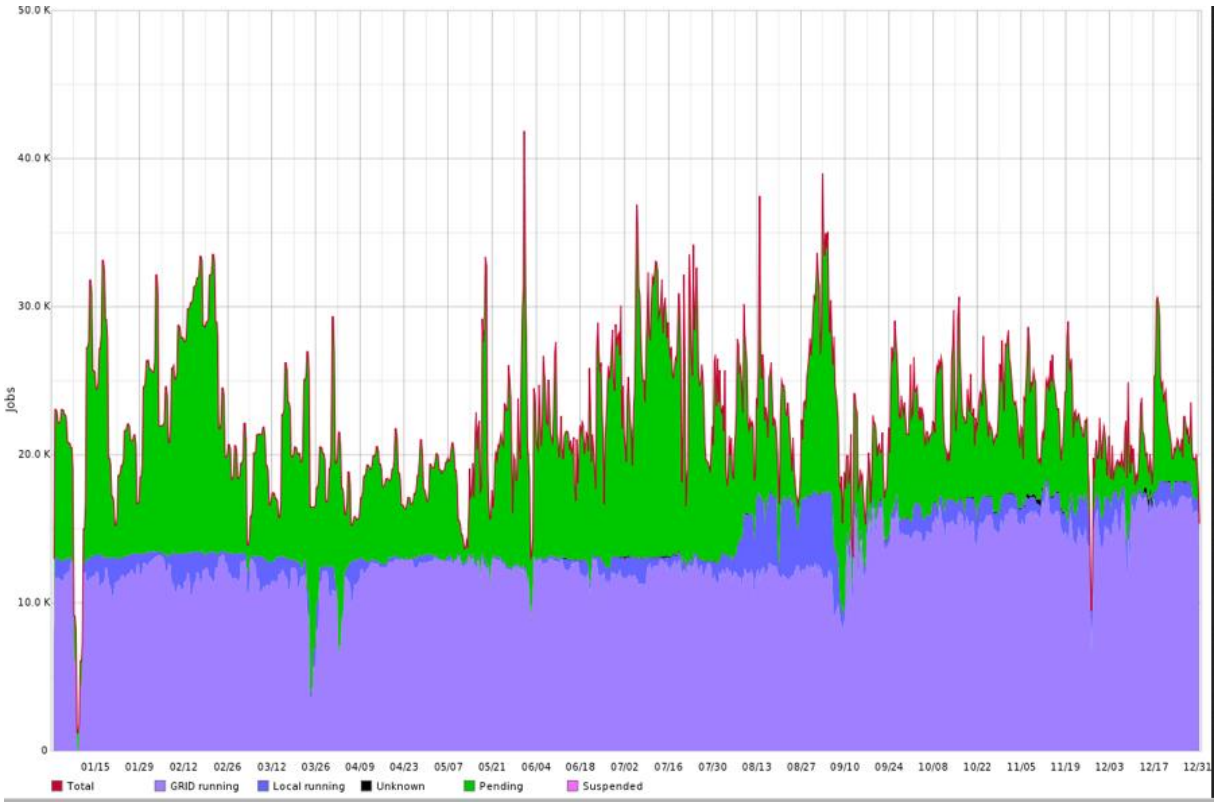


Figure 1 Farm usage trend during 2013

Jobs on the farm directly access disk resources using the file protocol: the storage is divided in several file-systems mounted on the WNs and seen as local disks.

The batch system adopted by CNAF since 2005 is LSF [3] from IBM (formerly Platform). The adoption of this solution was the result of some years of problematic experience with Torque [4], an open source solution derived from PBS [5]. The LSF instance installed at CNAF is fully redundant thanks to the installation of several masters and to a shared area on CNFS (Clustered NFS) [6]: in this way total system reliability is guaranteed.

3. Cloud and Virtualization

On the Tier1 farm we run WnoDeS (Worker Nodes on demand) [7], a system capable to provision virtual WNs on demand via the batch system and to allocate cloud resources via an OCCI [8] interface without the need to partition the farm. As WnoDeS will evolve to converge with Open Stack [9] (apparently Open Stack does not allow having the farm used both via cloud and via grid as WnoDeS does), we will test and deploy this solution.

4. Alarm system and monitoring

The monitoring and alarm system at the Tier1 is based on a hierarchy of Nagios servers [10], (for storage resources, Lemon collects performance metrics).

Nagios advanced features have been used in order to control the status of hosts and services correlated to the status of different hosts or services (the so called cluster control). In particular, several plug-ins have been developed in order to control the SAN and storage clusters (GPFS [11], *gridftp* and Storm [12]) as well as CEs and WNs. In some critical cases Nagios can also execute corrective actions like the restart of a non-responsive service or the modification of the DNS alias of a cluster.

Nagios configuration is populated automatically accessing the central resource database (DOCET [13]) or via a query to the GPFS clusters.

Critical alarms are sent from Nagios to the mobile phone of the on-duty staff and, together with minor critical alarms, sent to the relevant mailing lists: these information are also available on the Tier1 Dashboard, a web-based tool, developed in house. Moreover, a MySQL database keeps track of all monitored services states for historical purposes.

The infrastructure services, on the other hand, are controlled by a proprietary monitoring system (TAC) also able to send email and contact the on-duty staff; infrastructural alarms are collected on the Dashboard too.

The Dashboard web page displays a coloured box for each service, where the colour indicates the condition (green = OK, orange = WARNING, red = CRITICAL, grey = UNKNOWN). Pending errors/warnings are shown inside each box, with a brief description and the time at which they occurred.

The access to the Dashboard is restricted to CNAF staff.

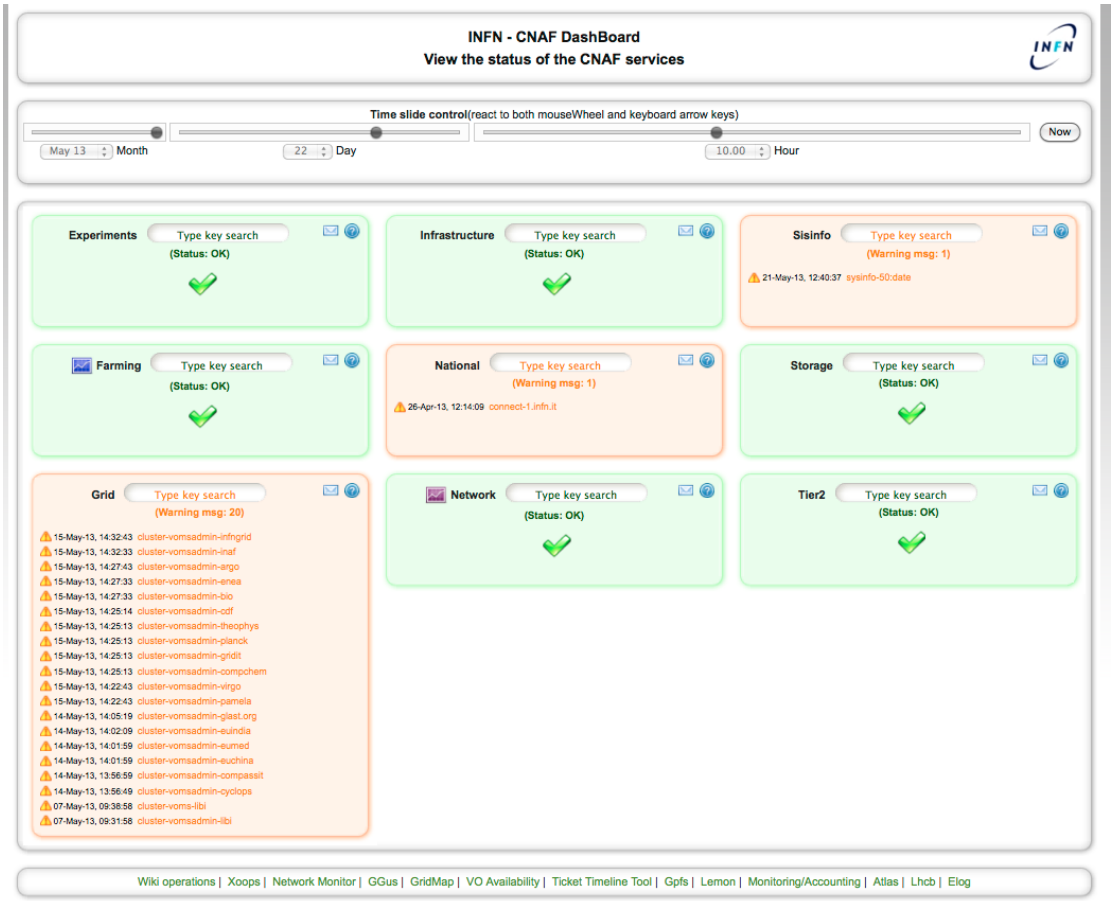


Figure 2 - Screenshot of CNAF dashboard

References

[1] European DataGrid Project (http://eu-datagrid.web.cern.ch/eu-datagrid/)
[2] EMI – European Middleware Initiative (http://www.eu-emi.eu/)
[3] LSF – Load Sharing Facility (http://en.wikipedia.org/wiki/Platform_LSF)
[4] Torque - Terascale Open-source Resource and QUEue Manager (http://en.wikipedia.org/wiki/TORQUE)
[5] PBS – Portable Batch System (http://en.wikipedia.org/wiki/Portable_Batch_System)

[6] CNFS – Clustered Network File System
(http://www.redbooks.ibm.com/abstracts/redp4400.html)
[7] WNoDeS – Worker Nodes on Demand Service
(http://wnodes.github.io/WNoDeS/documentation/articles.html)
[8] OCCI – Open Cloud Computing Interface (http://occi-wg.org)
[9] OpenStack (http://www.openstack.org)
[10] NAGIOS (http://www.nagios.org)
[11] GPFS – General Parallel File System
(http://en.wikipedia.org/wiki/IBM_General_Parallel_File_System)
[12] Storm – Storage Resource Manager (http://italiangrid.github.io/storm/index.html)
[13] DOCET (https://agenda.cnaf.infn.it/conferenceDisplay.py?confId=271)

National ICT infrastructures and services

S Antonelli[1], F Bisi[2], R Giacomelli[2], S Longo[1], S Meneghini[2], M Onofri[1], R Veraldi[1], G Vita Finzi[1] and S Zani[1]

[1] INFN-CNAF, Bologna, Italy
[2] INFN, Bologna, Italy

E-mail: stefano.antonelli@cnaf.infn.it, fabio.bisi@bo.infn.it, roberto.giacomelli@bo.infn.it, stefano.longo@cnaf.infn.it, stefano.meneghini@bo.infn.it, michele.onofri@cnaf.infn.it, riccardo.veraldi@cnaf.infn.it, giulia.vitafinzi@cnaf.infn.it, stefano.zani@cnaf.infn.it

Abstract.
Since the early 90s CNAF has been officially invested with the task to implement, manage, maintain and coordinate services which are critical and fundamental due to their importance and catchment area not only for CNAF itself but also for the whole INFN community. The CNAF department which carries out this activity is called *National ICT Infrastructures and Services* and it is composed of three Full Time persons.

1. Strategic and critical high-priority services

Some of the services managed by this group are critical for the good functioning of the whole INFN network and IT infrastructure:

- DNS for the infn.it TLD: we manage the authoritative DNS for the top level domain infn.it and for all the related subnets reverse name resolution which points to the local *.infn.it DNS servers. This service also acts as a secondary for *.infn.it sites. Over 150 zones are managed including non-infn.it zones for special purpose activities involving INFN like Grid and other projects.
- Mail relay MX backup: we manage the mail service backup MX for all the @*.infn.it email domains with a 15-day retention policy.
- Management of the IT infrastructure at the INFN Headquarters: we provide support for the INFN headquarters located in Rome, either through the use of local hardware or using remote services deployed at CNAF. We manage the local network, fundamental services like wired and WiFi networks, DNS, mailing, as well as implementing specific solutions when necessary to improve and to provide new services.

2. Medium-priority non-critical services

- National Mailing lists: we manage over 1000 lists for the INFN domains with a centralized system based on Sympa.
- Centralized Web Site management: this is a service which allows people to develop a web site for their particular project or experiment, in which INFN may be involved. We chose

a common CMS (Content Management System) based on JOOMLA for everyone. At the moment almost 90 sites are managed.

- NTP service: we manage one of the three Network Time Protocol servers for INFN computers.
- Backup service for the INFN Certification Authority (CA): a daily encrypted backup mirrors data of the INFN CA hosted at the Florence INFN site.
- Eduroam and TRIP management: we coordinate the eduroam and TRIP 802.1x-based WiFi national infrastructure. Actually TRIP is the precursor of eduroam inside INFN, uses the same technology and is widely adopted by many internal INFN users.
- Centralized License distribution: we deployed an infrastructure for license distribution related to several software packages: Ansys, Comsol, Autodesk, NX, Esacomp, Mathematica, Cliosoft, Mathlab. It is based on Linux Flex servers. The management of the infrastructure is done in collaboration with colleagues from other INFN sites: Padova, LNF, Napoli and Pisa.
- Activation of Microsoft Operating Systems and Office: we manage the KMS server for Windows Vista, Windows 7 and Windows 8 Operating Systems and Microsoft Office activation for all the INFN computers.
- National Multimedia Services: the following services are managed:
  - Asterisk server for INFN phone conferences
  - MCU H.323 for INFN video conferences
  - Real Networks and Flash video servers for streaming INFN events
  - SeeVogh reflector
  - vidyo router, part of the CERN vidyo infrastructure
- Centralized INFN Authentication and Authorization (AAI): we host three servers belonging to the INFN-AAI Service.
- AFS and Kerberos: we manage three servers for the INFN national AFS service.
- Design, development and maintenance of central infrastructures for the Cabibbolab: in particular we manage the electronic agenda based on Indico, the document system based on Alfresco and the e-mail server for the domain cabibbolab.it

3. Brand New cutting edge services

- DNS for High-Availability services: this is a new DNS architecture recently deployed to meet strict requirements concerning the DNS resolution for high-availability services. The new DNS architecture is based on a pool of DNS servers distributed across INFN sites and implementing a multi-master ISC bind solution with underlying Dynamic Loadable Zones and GaleraDB technologies.
- INFN document service: we have implemented the official INFN Document Management System, based on Alfresco. A section (a *site* in Alfresco parlance) is available for each INFN site, where documentation and other material related to local services and experiments can be stored. Various scientific collaborations are already using this service for their own documents: BELLE-II, BESIII, !CHAOS, GINGER, LHCb Italia, PRISMA, ReCaS, ReCaS-PRISMA, SPES e SR2S. It is foreseen to extend this service for the INFN official collection of documents.
- Disaster recovery: CNAF coordinates a working group for disaster recovery of important services. The involved sites are CNAF and LNF. At the moment the project is focused and limited to the INFN Information System.

- Synchronization and desktop backup Pandora: the service implements a Dropbox-like solution built at CNAF with standard open-source tools. It includes a synchronization client, multi-platform access via a web interface, data sharing through web URLs. It also offers a WebDAV interface and the possibility to create mini-sites for the distribution of images of operating systems and licensed software that is made available to the entire community. The service has been put into production and is available to all users belonging to the INFN-wide AAI.
- INFN video chat service: this is a Skype-like service for INFN users implemented on top of the XMPP protocol using open-source tools. Any XMMP-compliant client is automatically supported, although we suggest Jitsi, which, among other things, provides built-in support for confidentiality and non-reputation techniques. The service is integrated with the INFN AAI.
- Collaboration tools aimed at supporting proper software engineering practices during software development. These aspects are further described in other contributions to this annual report.

4. HW and SW architecture

The National ICT Infrastructures and Services are mainly implemented using virtualization technology inside cluster architectures (CentOS, VMware, OVirt). Some critical, special-purpose services, notably the DNS, are instead deployed within systems running on a bare hardware machine. We manage over 160 virtual machines to deliver all the described services.

Software Services and Distributed Systems

Software development made easier

S Antonelli[1], C Aiftimiei[2], M Bencivenni[1], C Bisegni[3], L Chiarelli[4], D De Girolamo[1], F Giacomini[1], S Longo[1], M Manzali[1], R Veraldi[1] and S Zani[1]

[1] INFN-CNAF, Bologna, Italy
[2] INFN, Padova, Italy
[3] INFN-LNF, Frascati, Italy
[4] GARR, Roma, Italy

E-mail: stefano.antonelli@cnaf.infn.it, cristina.aiftimiei@pd.infn.it, marco.bencivenni@cnaf.infn.it, claudio.bisegni@lnf.infn.it, lorenzo.chiarelli@garr.it, donato.degirolamo@cnaf.infn.it, francesco.giacomini@cnaf.infn.it, stefano.longo@cnaf.infn.it, matteo.manzali@cnaf.infn.it, riccardo.veraldi@cnaf.infn.it, stefano.zani@cnaf.infn.it

Abstract. This paper describes an infrastructure that is being made available to software developers within INFN to support and facilitate their daily activity. The infrastructure aims at integrating several tools, each providing a well-identified function: project management, version control system, continuous integration, dynamic provisioning of virtual machines, efficiency improvement, knowledge base. When applicable, access to the services is based on the INFN-wide Authentication and Authorization Infrastructure. The infrastructure will be beneficial especially for small- and medium-size collaborations, which often cannot afford the resources, in particular in terms of know-how, needed to set up such services.

1. Introduction

The success of a scientific experiment depends, often significantly, on the ability to collect and later process large amounts of data in an efficient and effective way. Despite the enormous technological progress in areas such as electronics, networking and storage, the cost of the computing factor remains high. Moreover the limits reached by some historical directions of hardware development, such as the saturation of the CPU clock rate with the consequent strong shift towards hardware parallelization, has made the role of software more and more important.

The ISSS (Infrastruttura di Supporto allo Sviluppo Software, Infrastructure in Support of Software Development) project [1] aims at presenting INFN researchers with a variety of tools already configured to support established best practices, so that the quality of the software they produce could be continuously improved at a decreasing cost. The generic term *quality* refers to characteristics such as low presence of defects, runtime performance efficiency, maintainability, easy portability to new platforms. Similarly, the generic term *cost* covers many aspects, such as time devoted to development, test, support and maintenance, money spent in hardware resources and electrical power, low service reliability.

2. Current status of the project

This document summarizes the current status of the project, with a focus on the end user-facing services that have reached a production or pre-production status: project management, version control and continuous integration functionality. Other functionality, notably the virtualization infrastructure that represents the foundation on top of which the abovementioned services rely, is described in another contribution in this report. A more thorough introduction of the whole project is presented elsewhere [2].

It is worth noting that, when applicable, access to the services is based on the INFN-wide Authentication and Authorization Infrastructure (AAI). For web applications the AAI offers a SAML-based Identity Provider (IdP) that allows a user to authenticate via either a personal X.509 certificate or a username/password pair. The AAI is available also to non-INFN users via a simple registration procedure, permitting also to projects not entirely formed by INFN users to access this infrastructure.

2.1. Project management

Any software project that goes beyond a limited-duration single-person exercise would benefit from a project management system. The more complex the project is (more developers, more users, larger code base), the larger the benefit. We have chosen Atlassian *JIRA* [3], a widely-used tool that offers excellent support for large and complex development projects involving many users. It covers many aspects of the software project lifecycle, including: interaction with users for requirements and support; organization of issues, tasks and activities; integration with code repositories; reporting.

2.2. Version control

A *Version Control System* (VCS) allows to store permanently any change applied to a code base (source code, tests, documentation, build and packaging instructions, etc.) along with metadata to keep track of the history of the changes.

Subversion [5] has been chosen as version control system to cover existing needs. The authentication to the system is based on SSH public keys. Beside Subversion, solutions for *git* [6] and *mercurial* [7] repositories hosted by third-parties, such as *GitHub* [8] and *Bitbucket* [9], are encouraged and well supported by the other tools, notably JIRA and Jenkins.

2.3. Continuous integration

In order for the developers to keep their confidence high that changes applied to the code base, by themselves or by their peers, do not cause regressions, it is a recommended practice to verify continuously (e.g. periodically or even at every change) that all changes introduced in the software at least build correctly and pass basic tests. As an additional benefit, the automatic build and test phases are an excellent occasion to run other quality checks. Several static and dynamic analysis tools exist that are able to expose actual or potential defects in the code before they reach production.

Jenkins [4] provides the framework for continuous integration and represents the foundation for any process that aims at producing high-quality software components in a reproducible way, as shown for example in figure 1. A few slaves are attached to the system, covering the most common Linux distributions for 32- and 64-bit platforms.

3. Conclusions

The success of a scientific experiment relies more and more on sound computing practices, notably in the development of scientific software.

The ISSS project aims at providing a one-stop shop for software developers, with a special focus on members of small- and medium-size experiments, where they can find state-of-the-art

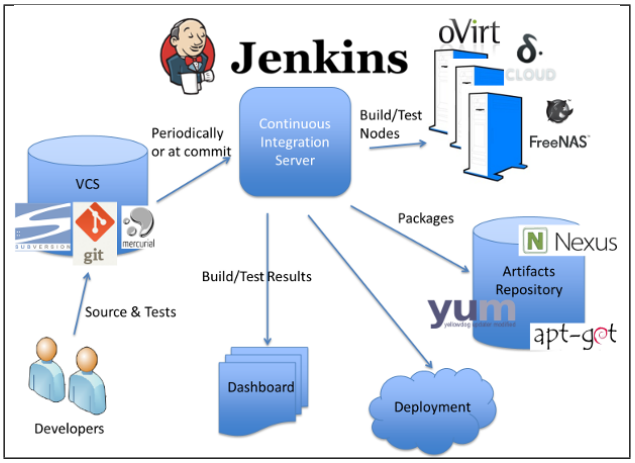


Figure 1. A typical software process using a continuous integration approach.

tools and services that help them deliver software of increasing quality at lower costs and on time, through the adoption of established best practices.

References

[1] The ISSS Project `https://web.infn.it/isss`
[2] Antonelli S et al 2014 An integrated infrastructure in support of software development *J. Phys.: Conf. Ser. Proceedings of CHEP2013* **513** In Press
[3] Atlassian JIRA `https://www.atlassian.com/software/jira`
[4] Jenkins `https://jenkins-ci.org/`
[5] Subversion `https://subversion.apache.org/`
[6] git `http://git-scm.com/`
[7] mercurial `http://mercurial.selenic.com/`
[8] GitHub `https://github.com/`
[9] Bitbucket `https://bitbucket.org/`

The Trigger and Data Acquisition system of the KM3NeT-Italy detector

T Chiarusi[2], F Giacomini[1], M Manzali[1,3] and C Pellegrino[2]

[1] INFN-CNAF, Bologna, Italy
[2] INFN, Bologna, Italy
[2] Università degli Studi di Ferrara, Ferrara, Italy

E-mail: tommaso.chiarusi@bo.infn.it, francesco.giacomini@cnaf.infn.it, matteo.manzali@cnaf.infn.it, carmelo.pellegrino@bo.infn.it

Abstract. KM3NeT-Italy is an INFN project that will develop a submarine cubic-kilometre neutrino telescope in the Ionian Sea (Italy) in front of the south-east coast of Portopalo di Capo Passero, Sicily. It will use thousands of PMTs to measure the Cherenkov light emitted by high-energy muons, whose signal-to-noise ratio is quite disfavoured. This forces the use of an on-line Trigger and Data Acquisition System (TriDAS) in order to reject as much background as possible. In March 2013 a prototype detector, hosting 32 PMTs, has been deployed in the abyssal site of Capo Passero and successfully operated. The existing TriDAS software, used for the prototype, needs a deep revision in order to meet the requirements of the final detector: the adoption of new tools for software development and modern design solutions will bring several improvements and simplifications during this upgrade.

1. Introduction

Neutrinos are the perfect probe to explore the far Universe. They have no electrical charge, are insensitive to magnetic fields and interact only weakly. This means they can travel huge distances from their production sites before reaching a detector on the Earth, transporting direct information on mechanisms acting inside cosmic accelerators. The discovery of astrophysical neutrinos will open a new perspective of astronomy and astrophysics, complementing present gamma ray astronomy and cosmic ray studies [1].

However, because of the same properties that make them unique messengers of high-energy processes, there are severe limitations to detectability of high-energy neutrinos. Very large detection volumes, at least 1 km^3, are required in order to have an unambiguous signal due to astrophysical neutrinos together with an effective shield against the overwhelming background due to atmospheric muons, residuals of the high-energy cosmic ray showers.

KM3NeT-Italy is an INFN project co-financed by the European Community and the Italian Government. Its main objective is the construction of a 1 km^3 high-energy neutrino detector telescope in the Ionian Sea (Italy) that will also host an interdisciplinary observatory for marine sciences [2]. In its final phase, the detector will use about 5000 PMTs to measure the Cherenkov light emitted by high-energy muons created in ultra high-energy astrophysical neutrino interactions, whose signal-to-noise ratio is quite disfavoured. This forces the use of an on-line triggering system in order to reject as much background as possible.

In March 2013, a prototype Detection Unit (DU), following the tower layout and hosting 32 PMTs, has been deployed in the abyssal site of Capo Passero and successfully operated.

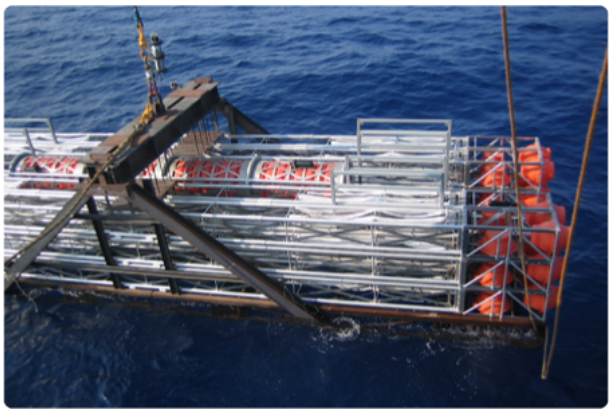


Figure 1. The prototype tower during the deployment.

Since then, the on-line Trigger and Data Acquisition System (TriDAS) has been continuously running at the Portopalo control station [3]. At the beginning of 2015, the detection volume will be expanded with 8 more towers, each hosting 84 PMTs, completing the KM3NeT-Italy Phase 1 detector. KM3NeT-Italy is also the italian node of KM3NeT, a second-generation distributed neutrino telescope [4].

2. The evolution of the TriDAS software

TriDAS, designed for the prototype tower deployed by KM3NeT-Italy and running at the off-shore station of Portopalo, has been stressed with a long period of data acquisition, started more than a year ago. After an important upgrade of the dedicated computing infrastructure and a revision of the existing software, TriDAS will also be used for the data acquisition of the first block of the KM3NeT telescope, that is composed of 8 towers. Each tower follows the prototype layout but with 84 PMTs instead of 32.

In order to meet the new requirements, the revision of TriDAS has made use of the services offered by the ISSS project and has involved the adoption of modern software design solutions and high-level libraries.

ISSS (Infrastruttura per il Supporto allo Sviluppo Software) is an INFN project led by CNAF [5, 6, 7], whose primary goal is the creation of an infrastructure offering a variety of tools and services already configured to support software development for projects inside INFN.

2.1. C++ Boost libraries

Boost [8] is a set of open-source C++ libraries that extend the functionality of the C++ standard library. Many of Boost's founders are themselves in the standards committee and several Boost libraries have been accepted for incorporation into the C++11 standard.

The use of Boost libraries in TriDAS, in addition to increasing the efficiency and maintainability of code, facilitates the transition to future C++ standards.

2.2. ZeroMQ

ZeroMQ [9] (also known as zmq) is both an embeddable networking library and a concurrency framework. It provides sockets that carry atomic messages across various transports like in-process, inter-process, TCP, and multicast. It allows to connect sockets N-to-N with patterns like fan-out, pub-sub, task distribution and request-reply. It has an asynchronous I/O mode,

APIs for the main languages and runs on most operating systems. ZeroMQ has been chosen to implement communication and monitoring functionality in TriDAS.

2.3. CMake

CMake [10] is cross-platform open-source software for managing the build process of software, using a compiler-independent method. It is designed to simplify the compilaton process: distinguished by a modular structure, it provides an easy way to create dedicated Makefiles.

The introduction of CMake in TriDAS has simplified the steps of compilation, testing, packaging and deployment, allowing the execution of unit and integration tests after compilation and the creation of self-installing scripts and archives.

2.4. Jenkins

Jenkins [11] is an open-source continuous integration tool that provides continuous integration services for software development. Builds can be started by various means, including being triggered by a commit in a version control system, scheduling via a cron-like mechanism, building when other builds have completed, or by accessing a specific build URL.

A Jenkins service, together with a pool of nodes with different operating systems and architectures, is available through the ISSS project: this service is used to compile and test TriDAS every night automatically on different platforms.

2.5. git and Bitbucket

git [12] is an open-source distributed version control system, characterized by a strong support for non-linear development, making it ideal for projects that require an extensive collaboration.

For the development of TriDAS git has been chosen instead of SVN, using Bitbucket [13] as a web-based service hosting the source code. Bitbucket offers additional services, notably in support of code review, which have been extensively used during development.

2.6. JIRA

JIRA [14] is a project management system that lets users prioritize, assign, track, report and audit issues, from software bugs and helpdesk tickets to project tasks and change requests. JIRA is also an extensible fully-customizable platform. A TriDAS project is hosted on the INFN JIRA service, which keeps track of the evolution of the code.

References

[1] T. Chiarusi and M. Spurio, 2010, High-Energy Astrophysics with Neutrino Telescope, European Physics Journal C 65, nn. 3-4, 649
[2] A. Margiotta, The KM3NeT project: status and perspectives, Geosci. Instrum. Method. Data Syst., 2, 35-40, doi:10.5194/gi-2-35-2013, 2013
[3] C. Pellegrino et al, The Trigger and Data Acquisition for the NEMO-Phase 2 Tower, Proceedings for the VLVNT 2013, Stockholm, in press
[4] A. Margiotta, The KM3NeT detector, this report
[5] S. Antonelli et al., An integrated infrastructure in support of software development, Proceedings of CHEP 2013, J. Phys. Conf. Ser. 513, 2014
[6] S. Antonelli et al., Software development made easier, this report
[7] S. Antonelli et al., National ICT infrastructures and services, this report
[8] Boost C++ Libraries `http://www.boost.org/`
[9] ZeroMQ `http://zeromq.org/`
[10] CMake `http://www.cmake.org/`
[11] Jenkins `https://jenkins-ci.org/`
[12] git `http://www.git-scm.com/`
[13] Bitbucket `https://bitbucket.org/`
[14] Atlassian JIRA `https://www.atlassian.com/software/jira`

The COKA Project

R Alfieri[1], M Brambilla[1], R De Pietri[1], F Di Renzo[1], A Feo[1], F Giacomini[2], M Manzali[2], G Maron[2], D Salomoni[2], S F Schifano[3] and R Tripiccione[3]

[1] Università di Parma and INFN-Gruppo collegato di Parma, ITALY
[2] CNAF - INFN, Bologna, ITALY
[3] Università di Ferrara and INFN-Ferrara, ITALY

E-mail: alfieri@pr.infn.it, brambilla@pr.infn.it, depietri@pr.infn.it, direnzo@pr.infn.it, feo@pr.infn.it, giacomini@cnaf.infn.it, manzali@cnaf.infn.it, maron@cnaf.infn.it, salomoni@cnaf.infn.it, schifano@fe.infn.it, tripiccione@fe.infn.it,

Abstract.
In this document we describe the activities carried out by the COKA project in 2013 and involving CNAF. In short, we continued to investigate the performance of production-grade MIC-based accelerator boards, released by Intel at the beginning of 2013 under the name Xeon-Phi. We first used Xeon-Phi systems available at external computing centers, but we finally managed, as members of the Intel MIC development program, to have one such board directly from Intel. That board was later integrated in a more general facility to experiment with computing co-processors.

1. Introduction

The *COmputing on Knights Architectures* (COKA) project started in 2012 with the goal of testing the up-coming Intel Many Integrated Core (MIC) architecture for applications relevant for theoretical and experimental physics, assessing its performance and efficiency. In 2013 we started to work with the first production release of a MIC board, called Xeon-Phi.

Our work follows two different but related directions. On one side we try to assess the performance limits of these processors as precisely as possible; this involves careful tuning of the algorithms to the processor architecture, including substantial program re-coding and detailed benchmarking. The goal is to establish the ultimate performance levels that this architecture is able to deliver. On the other side we explore the possibilities of reasonable programming and run-time environment for these processors that does not require substantial handcrafted program re-design, while exploiting at the same time an acceptable fraction of the theoretically available computing power.

A large fraction of our work has been focused on computational physics: we have completed the porting of a production-grade lattice-Boltzmann code for computational fluid-dynamics and we have started to port the Einstein toolkit; work is also in progress with the LGT Chroma and parmalgt packages.

On the experimental physics side we investigated the porting of a set of codes developed by the CMS experiment.

2. Main Results

In the following we highlight some interesting results obtained during the year 2013.

2.1. Experimental Hardware

First of all, CNAF has used funds assigned to the project to procure, install and configure equipment that is now part of an R&D testbed at CNAF. The COKA part of this testbed is currently composed of two servers, with a total of 3 Intel Xeon-Phi boards, 2 NVIDIA K20 boards and some 24 x86-cores processors. These servers are accessible by COKA users through a user interface and a batch system, configured with a number of queues serving different purposes. The CNAF group is currently providing support and customizations for this set-up. Moreover it makes available to the COKA collaboration other state-of-the-art computing devices, such as ARM and GPU boxes, and recently announced and released x86 machines.

2.2. Porting CMS code

CNAF also collaborated with the CMS experiment to evaluate how easy would be to port their reconstruction and analysis software (CMSSW) to the Xeon-Phi platform. There are two practical difficulties when setting up a software environment for the Phi. First, no compilation environment is available on the Phi itself, and cross-compilation must be done on the host. Second, currently only the Intel compiler can be used for the compilation.

The porting turned out to be a non trivial task. The biggest hurdle was the dependency on the Intel compiler, which still lacks proper support for the new C++ standard (C++11) [1], whose features are already in use in CMSSW, notably in its core framework, which is used by most of the rest of the code.

Two releases of the compiler were tested. Version 13.1.3 was almost unusable, due to its very poor support of C++11. Version 14.0.0 was certainly better in this respect but affected by a show-stopper bug, which prevented further progress. At that point 369 out of the 1106 CMSSW packages could be properly built, which allowed at least some preliminary testing [2].

CNAF is also currently working on benchmarks related to efficiency and computing costs of Intel Xeon-Phi boards, in particular for what regards offloading code to the accelerator. These benchmarks are relevant to develop performance models that allow to evaluate the benefits of porting code on Xeon-Phi accelerators.

2.3. Computational Fluido-dynamics

Our activity has been divided in two main parts: the first is focused on benchmarking[3], and the latter on the implementation and optimization of a full production-grade code of a D2Q37 Lattice Boltzmann model for computational fluid-dynamics.

As far as benchmarking is concerned, let us consider memory access and the main computational kernels of the D2Q37 code.

Figure 1 left shows the skeleton of the memory benchmark code used to measure the effective memory bandwidth of the Xeon-Phi. The benchmark copies AVX512 element from array A to array B using three different store instructions:

- `_mm512_store_pd`: this is the conventional store instruction used to update one AVX512 vector in memory; the semantic of this instruction forces a read of the target memory address if not included in cache, wasting memory bandwidth;
- `_mm512_storenr_pd`: this function is intended to speed up memory stores in streaming kernels where we want to avoid wasting memory bandwidth by being forced to read the original content of the entire cache line from memory whose whole contents is going to be completely updated;

```
#define _vload_pd  _mm512_load_pd

#if defined USE_STORENR
  #define _vstore_pd  _mm512_storenr_pd
#elif defined  USE_STORENRNGO
  #define _vstore_pd
_mm512_storenrngo_pd
#elif defined USE_STORE
  #define _vstore_pd _mm512_store_pd
#endif

__m512d A[N], B[N];

#pragma omp parallel private(tid)
{
  tid = omp_get_thread_num();
  th_func((void *) &targv[tid]);
}

// code executed by each thread
void th_func(void * targv) {
  for (i=0; i<L; i++)
    vstore(B+i, vload(A+i));
}
```

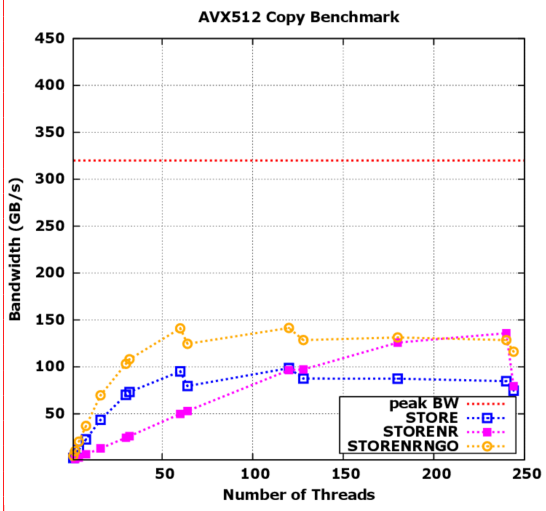


Figure 1. Skeleton of a copy benchmark-code used to measure memory bandwidth.

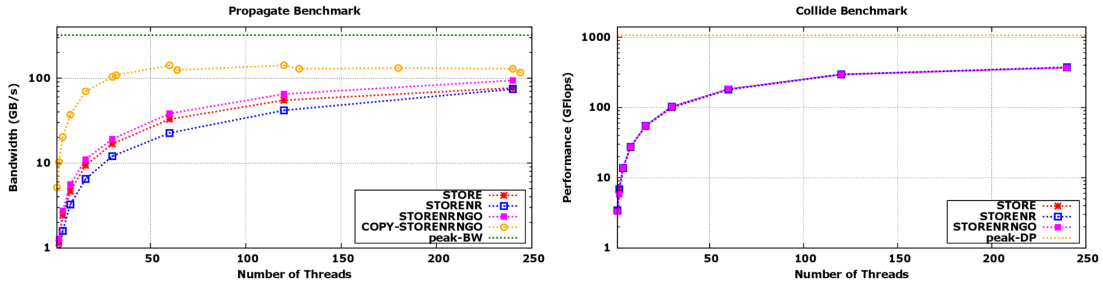


Figure 2. Performance of the propagate and collide kernels as a function of number of threads. Lattice size is 1920×1600.

- `_mm512_storenrngo_pd` this instruction behaves like the previous one, but it adopt a weakly-ordered memory consistency model. Stores performed with this function are not globally ordered, so subsequent stores issued by the same thread may be performed ahead in time.

Figure 1 right showss the corresponding memory bandwidth measured by the benchmark. As we see the benchmark using the `_mm512_storenrngo_pd` instruction performs better and reaches a plateau much faster, delivering an effective bandwidth of $\approx 140 GB/s$, corresponding to $\approx 44\%$ of the available peak.

We have then implemented specific benchmarks to measure performances of the main computing kernels of our LB code, propagate and collide; propagate essentially performs a large set of sparse address memory accesses, while collide is a floating-point intensive kernel. These benchmarks have been coded using AVX vector intrinsic instructions. Figure 2 shows performance and scalability of the propagate and collide kernel as a function of the number of threads. The propagate kernel takes benefit of the storenrngo instruction; using 240 threads, it runs 1.2x faster than the version using the usual store instruction. The maximum bandwidth is $\approx$ 94 GB/s, corresponding to $\approx 67\%$ of the bandwidth measured by the previous memory-benchmark, and to $\approx 30\%$ of the theoretical peak. The collide kernel is strongly compute-bound,

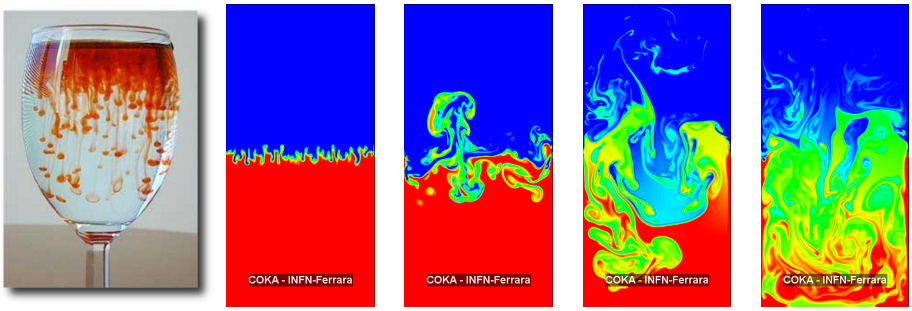


Figure 3. Temperatures maps of a simulation of the Raileigh-Taylor instability at several stages of the time evolution. The simulation runs on a Xeon-Phi board using 60 cores, 240 threads and exploits AVX instructions; the size of the lattice is 512×1024 sites.

so its performances are not affected by the store-instruction used. The maximum performance obtained with 240 thread is ≈ 398 GFlops corresponding to $\approx 37\%$ of the available peak.

Based on the analysis results of the benchmarks, we have then developed a full production code that we plan to use to simulate the Rayleigh-Taylor instability that develops when fluids of different density and temperature are subject to gravity. Our test simulation runs on a lattice of 512×1024 points; this lattice size of is small, and not interesting for real physics simulations, but it is relevant to check the correctness of the porting. The simulation runs in parallel on a variable number of hardware cores, up to 60, and a number of threads ranging from 1 to 4 per core, and exploits AVX vectorization. In figure 3 we show the temperature maps of a at several stages during time evolution.

	C2050	2-WS	2-SB	Xeon-Phi	K20X
propagate GB/s	84	17.5	60	94	160
ϵ	58%	29%	70%	29%	64%
collide GF/s	205.4	88	220	398	506
ϵ	41%	55%	63%	37%	38%
ξ (collide)	–	1.19	1.27	0.76	–

Table 1. Performance comparisons of our D2Q37 lattice Boltzmann code on several platform: C2050 and K20X are two GP-GPUS, while 2-WS and 2-SB are dual-processor systems based respectively on Westmere and Sandybridge architectures.

Finally, Table 1 contains a summary of performances so far measured on different platforms. So far our implementation on Xeon-Phi has a slightly worse performance than that running on the latest generation NVIDIA K20X GP-GPU. The main issue seems to be associated to memory access and it still is under investigation. In an attempt to provide a fair comparison of performances across architectures with widely different number of cores and vector sizes, we have defined the ξ metric

$$\xi = \frac{P}{N_c \times v \times f}$$

where P is the measured performance of a kernel code, N_c is the number of cores on which the kernel has been parallelized, v is the size of vector instruction used, and f is the operating frequency of the processor. The ξ parameter should allow to compare how well a given architecture is able to use all its parallel features, as well as its sheer clock speed, in order to deliver performance to a given application. We see that the more traditional Intel processors

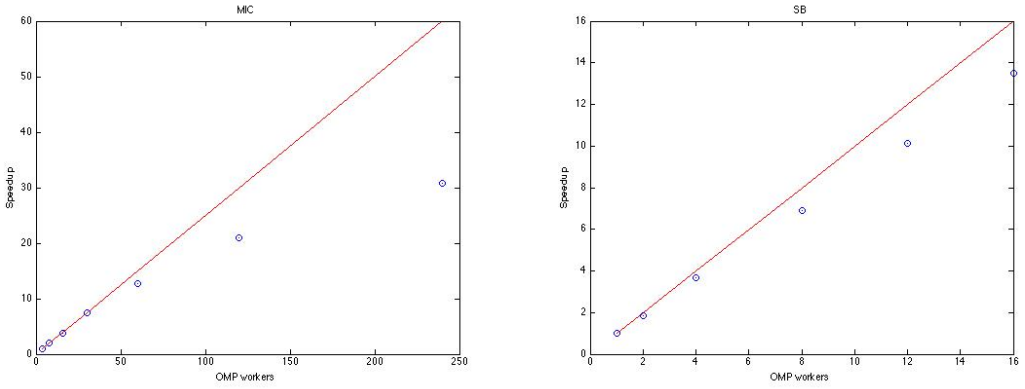


Figure 4. Strong scaling of a parmalgt application: speedup of a sixth order, 32^4 simulation on the Xeon-Phi system (right) and on an Intel SandyBridge dual-processor system (left).

have very similar figures for ξ while the MIC system has a significantly lower value. It will be interesting to see if more clever programming and optimization strategies may improve on these figures; work is in progress in this direction.

2.4. Experimenting with refactoring vs porting codes

One of the goal of the COKA project is to compare the relative merits of a complete refactoring of codes for the Xeon-Phi versus the porting of codes running on different architectures. To understand what to expect from both approaches, a prerequisite was the benchmarking of basic performances such as FP effectiveness in matrix computations or memory operations efficiency. These benchmarks were performed being aware of the main features of the MIC architecture: the introduction of the new set of SIMD ISA extensions (at the moment the widest available in terms of bytes) and the massive number of cores. The Intel compiler environment introduces the *array notation*, intended to clearly point out sections of the code to be mapped to SIMD instructions. We investigated the effectiveness of this syntax in benchmarks and in more realistic context. In particular, we experimented with the joint usage of array notation and thread parallelism in the easy implementation made available by OpenMP and/or Cilk.

All these activities should contribute to the definition of basic guidelines for efficient (more or less native) coding of Lattice QCD applications. The level of performances obtained is at the moment still far from the ones reached on more standard multi-core processors.

At the same time, efforts were dedicated to the porting (as native Xeon PHI application) of some HPC Theoretical Physics applications which are available for other, more traditional architectures. The main goal is to test to which extent good performances can be achieved on the Intel-Xeon PHI architecture without strong changes of the codes. One application which has been considered is the *parmalgt* suite [10], a library intended to enable Numerical Stochastic Perturbation Theory computations for lattice QCD. On this package we have full control, having been developed here in Parma.

Figure 4 displays strong scaling results (the number of threads is varied while keeping the problem size fixed) on the Xeon-Phi and on a dual-processors system based on a standard Intel multi-core processor (SandyBridge). On top of the speedup, one should consider that the percentage of peak-performance obtained on the Phi is roughly one half of that obtained on the SandyBridge system.

On the other side we tested widely used applications that are freely available: the **CHROMA** lattice QCD application [11] and the **Einstein Toolkit** Astrophysical Application suite [12].

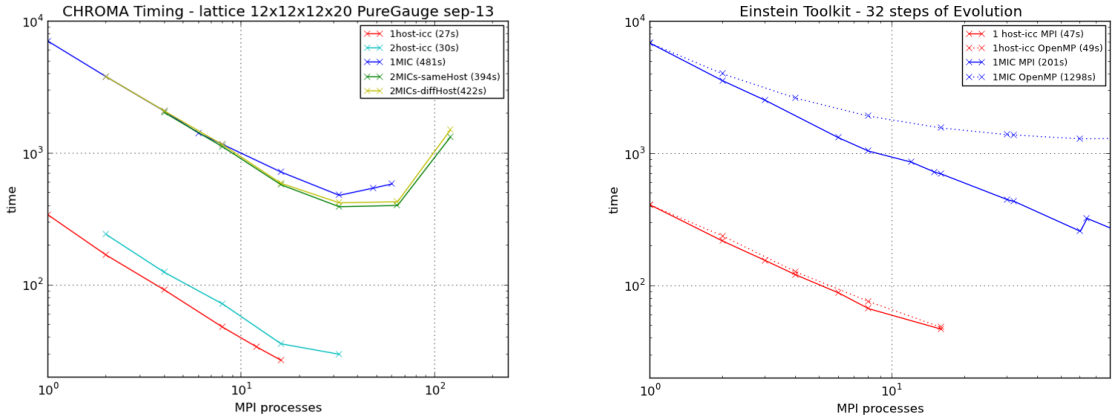


Figure 5. The left hand side panel shows the time needed to perform 200 Montecarlo sweeps on a $12^4 \times 20$ lattice for pure Gauge SU(3) using CHROMA. The execution time is 341 seconds on a host core while it is 7033 seconds on a PHI core. The right hand side panel displays the time needed to compute 32 time evolution steps of a single General Relativistic Star using the EinsteinToolkit on 65^3 grid (0.6 total GBytes allocated memory). On a single host core the requested time is 410 seconds while on a single PHI core the requested time is 6857 seconds.

The main problem we had to deal with was the compilation (for native mode execution) of the auxiliary libraries needed by these applications. The strategy used in testing these applications was to analyze the performance that can be obtained by just recompiling on the MIC environment without any code customizations, that is, using the potential key advantage of the MIC architecture with respect to other accelerators like the General-Purpose Graphical Processing Units (GP-GPUs). The main results obtained using this simple procedure is that one obtains unsatisfactory performance (see Fig. 5) with respect to the execution on the computing host (2x Sandy Bridge, E5-2687W 3.10 GHz, 8 cores) and specific optimizations are needed also in the case of native execution.

References

[1] ISO/IEC 14882:2011, Information technology – Programming languages – C++.

[2] D. Abdurachmanov, et al., *Explorations of the viability of ARM and Xeon Phi for physics processing*, to appear in Proceedings of the 20th International Conference on Computing in High Energy and Nuclear Physics (CHEP), October 14-18, 2013 Amsterdam, Netherlands, arXiv:1311.1001.

[3] G. Crimi, F. Mantovani, M. Pivanti, S.F. Schifano, R. Tripiccione, *Early Experience on Porting and Running a Lattice Boltzmann Code on the Xeon-Phi Co-Processor*, Proceedings of the International Conference on Computational Science, ICCS 2013 Procedia Computer Science, Volume 18, 2013, Pages 551-560, doi:10.1016/j.procs.2013.05.219.

[4] F. Mantovani, M. Pivanti, S.F. Schifano, R. Tripiccione, *Exploiting parallelism in many-core architectures: a test case based on Lattice Boltzmann Models*, Proc. of Conference on Computational Physics October 14-18, 2012 Kobe, Japan, J. Phys. Conf. Ser. 454 Vol. 1, 2013, doi:10.1088/1742-6596/454/1/012015

[5] A. Bertazzo, F. Mantovani, M. Pivanti, F. Pozzati, S.F. Schifano, R. Tripiccione, *Implementation and Optimization of a Thermal Lattice Boltzmann Algorithm on a multi-GPU cluster*, Proceedings of Innovative Parallel Computing (INPAR) 2012, May 13-14, 2012 San Jose, CA (USA), doi:10.1109/InPar.2012.6339603.

[6] F. Mantovani, M. Pivanti, S.F. Schifano, R. Tripiccione, *Performance issues on many-core processors: A D2Q37 Lattice Boltzmann scheme as a test-case*, Proceedings of 24rd International Conference on Parallel Computation Fluid Dynamics (PARCFD), May 21-25, 2012, Atlanta, GE (USA), Computers and Fluids Volume 88, 15 December 2013, Pages 743-752 (2013), doi: 10.1016/j.compfluid.2013.05.014.

[7] L. Biferale, F. Mantovani, M. Pivanti, F. Pozzati, M. Sbragaglia, A. Scagliarini, S. F. Schifano, F. Toschi, R. Tripiccione, *A multi-GPU implementation of a D2Q37 Lattice Boltzmann Code*, 9a International Conference on Parallel Processing and Applied Mathematics (PPAM11), September 11-14, 2011, Torun

(Poland). R. Wyrzykowski et al. (Eds.): PPAM 2011, Part I, LNCS 7203, pp. 640-650. Springer, Heidelberg (2012), doi:10.1007/978-3-642-31464-3_65.

[8] L. Biferale, F. Mantovani, M. Pivanti, F. Pozzati, M. Sbragaglia, A. Scagliarini, S. F. Schifano, F. Toschi, R. Tripiccione, *Optimization of Multi-Phase Compressible Lattice Boltzmann Codes on Massively Parallel Multi-Core Systems*, International Conference on Computational Science (ICCS), June 1-3, 2011, Singapore Procedia Science Vol. 4, pp. 994-1003, 2011, doi:10.1016/j.procs.2011.04.105.

[9] L. Biferale, F. Mantovani, M. Pivanti, F. Pozzati, M. Sbragaglia, A. Scagliarini, S. F. Schifano, F. Toschi, R. Tripiccione, *An Optimized D2Q37 Lattice Boltzmann Code on GP-GPUs* Proceedings of 23rd International Conference on Parallel Computation Fluid Dynamics (PARCFD) May 16-20 Barcelona (Spain), Computers and Fluids Vol. 80 (2013), pp. 55-62, doi:10.1016/j.compfluid.2012.06.003.

[10] M. Brambilla, F. Di Renzo, D. Hesse, *Code development (not only) for NSPT* PoS LATTICE **2013**, 418 (2014).

[11] R. G. Edwards *et al.* [SciDAC and LHPC and UKQCD Collaborations], *The Chroma software system for lattice QCD*, Nucl. Phys. Proc. Suppl. **140** (2005) 832

[12] F. Löffler, J. Faber, E. Bentivegna, T. Bode, P. Diener, R. Haas, I. Hinder, B. C. Mundim, C. D. Ott, E. Schnetter, G. Allen, M. Campanelli, and P. Laguna. *The Einstein Toolkit: A Community Computational Infrastructure for Relativistic Astrophysics*, Classical and Quantum Gravity, 29(11):115001, 2012, doi:10.1088/0264-9381/29/11/115001

Quality in Software for Distributed Computing

M Canaparo and E Ronchieri and D Salomoni

INFN CNAF, Viale Berti Pichat 6/2, 40126, Bologna, Italy

E-mail: marco.canaparo@cnaf.infn.it, elisabetta.ronchieri@cnaf.infn.it, davide.salomoni@cnaf.infn.it

Abstract. In the last decade INFN CNAF has contributed in developing the majority of software solutions for distributed computing. It has been participating in several European projects, such as DataGrid, EGEE, EMI and EGI, which have produced software for achieving the challenges of high energy physics communities in first place and later for supplying other communities' needs. Up to now three main areas belonging to the Grid paradigm ascribe to the software products: storage with StoRM, authorization with VOMS and computing with WMS and WNoDeS. Aware of the experience done at INFN CNAF, we noticed that software researchers, who have been implementing this code, massively forget the quality of their products for different reasons: on the one hand, developers feel pressed for addressing the requests of their users within scheduled budget and short time; on the other hand, they distrust data from existing quality tools since they provide partial analysis of their software. This has led to spend effort maintaining software once released and to develop software without exploiting solutions for managing defects effectively. Notwithstanding that developers perceive quality as an extremely time-consuming task that affects their productivity, in our opinion enhancing quality allows reducing defects, and, as consequence, saving costs and decreasing delivery delays; the software quality models and metrics represent the mainstream to reach high reliability balancing effort and results.

In this report, we are going to describe our solution to the aforementioned issues. Leveraging past and present literature about software quality models, we designed our own mathematical model connecting software best practices with code metrics by defining them in a formal way. To predict the quality at any stage of development, we supplied the model with statistical techniques such as discriminant analysis and linear regression. Input data to this model are some characteristics of packages, while outputs are measures of the defined metrics, which build the data set for the predictive techniques. For the validation process, we used some EMI software products whose defects were already known. Our solution has proved to reasonably reproduce reality.

1. Current State

At the beginning of this study, we marked best practices [1] and metrics [2] suitable for determining software products success and offering the greatest return. However, we planned the validation of our model with a progressive increase in the data set to properly speculate on the variables included in the model.

In [3], we selected a set of best practices (see [4]) referring to software structure, configuration management, construction of the code and deployment. We derived from them some metrics further extended with static code ones [5], such as Lines Of Code and Number of Defects. Once introduced the metrics, we used the mathematical description formalism to express various levels of abstraction from the fundamental concepts of software engineering up to metrics to

design our model. Ultimately, we supplied it with a predictive technique, called risk-threshold Discriminant Analysis (DA) [6], whose starting point is the measure of the foregoing metrics, while its outcomes determine risky software products that may contain defects. Combining the model with DA needed a validation: therefore we compared calculated results with the real data coming from EMI releases [7]. In particular, we selected source code mainly written in `Python` and `sh`, from the WNoDeS (Worker Nodes on Demand Services) [8] and StoRM (STOrage Resource Manager) [9] products released in the EMI 3 Monte Bianco distribution, to highlight similarities and differences among development scenarios. The analysis exploited 1,513 files in 15 software components amounting to a 27,406 total lines of code by using a Matlab-based prototype tool that codes the presented solution. The prototype classified all the components in faulty and non-faulty groups with a correctness of about 83%.

In [10], we validated our model enlarging the data set by increasing the number of metrics and software products. For the former we also considered complexity metrics, while for the latter we used CREAM (Computing Resource Execution And Management) [11], VOMS (Virtual Organization Management System) [12], WMS (Workload Management System) [13] and YAIM (Yet Another Installation Manager) [14] in addition to StoRM and WNoDeS. Furthermore, we used as predictive techniques both risk-threshold DA and linear regression [15] to discriminate the risky software products and defects respectively. As result, DA confirmed the correctness given in [3], while the regression method determined an inaccurate number of defects. However, the outcomes can be improved increasing the data set size and better contextualizing the predictive methods.

2. Future Work

Starting from the current state, there are several improvements that we can conduct in the following periods. In the short-term, we are going to study the correlation among metrics and to express defects as function of various metrics: the former provides us with details about how they influence one another; the latter determines which metric has a greater weight than others. The statistical computing tool, called R, represents the best tool to fulfill these achievements. Our aim is to identify and improve the predictive technique that reproduces reality as much as possible. In the medium-term, we would like to increase the data set by adding new software products and metrics both static and dynamic ones. Our purpose is to improve the % of correctness in the prediction of our model. In the long-term, we will consider and adopt further predictive techniques in addition to DA and regression, such as k-fold cross-validation [16] and support vector method [17], to strengthen the validation of our model. At this point, our solution will be ready for being adopted by developers and integrated in their development environment.

References

[1] CMMI P T 2010 CMMI for Development, Version 1.3 Technical Report CMU/SEI-2010-TR-033 Software Engineering Institute URL `http://resources.sei.cmu.edu/library/asset-view.cfm?assetid=9661`
[2] Kan S H 2002 *Metrics and Models in Software Quality Engineering* (Addison-Wesley Professional)
[3] Ronchieri E and Canaparo M 2013 *The 8th International Conference on Software Engineering and Applications (ICSOFT-EA 2013)*
[4] Perks M 2006 Best practices for software development projects Tech. rep. IBM
[5] Chidamber S R and Kemerer C F 1994 *IEEE Transactions on Software Engineering* **20** 476–493
[6] Guo G and Guo P 2008 *International Conference on Computational Intelligence and Security*
[7] Aiftimiei C, Ceccanti A, Dongiovanni D, Di Meglio A and Giacomini F 2012 *Journal of Physics: Conference Series (JPCS)* **396**
[8] Salomoni D, Italiano A and Ronchieri E 2011 *Journal of Physics: Conference Series (JPCS)* **331**
[9] Zappi R, Ronchieri E, Forti A and Ghiselli A 2011 *An Efficient Grid Data Access with StoRM* (Springer New York) chap VI Grid Middleware and Interoperability, pp 239–250 Data Driven e-Schience. Use Cases and Successful Applications of Distributed Computing Infrastructures (ISGC 2010)

[10] Ciaschini V, Canaparo M, Ronchieri E and Salomoni D 2014 *Journal of Physics: Conference Series (JPCS) (under pub)*
[11] Andreetto P, Bertocco S, Capannini F, Cecchi M, Dorigo A, Frizziero E, Gianelle A, Giacomini F, Mezzadri M, Monforte S, Prelz F, Molinari E, Rebatto D, Sgaravatto M and Zangrando L 2011 *Journal of Physics: Conference Series* **331**
[12] Ceccanti A, Ciaschini V, Dimou M, Garzoglio G, Levshina T, Traylen S and Venturi V 2009 *Journal of Physics: Conference Series* **219**
[13] Cecchi M, Capannini F, Dorigo A, Ghiselli A, Giacomini F, Maraschini A, Marzolla M, Monforte S, Pacini F, Petronzio L and Prelz F 2009 *Advanced in Grid and Pervasive Computing* (*Lecture Notes in Computer Science* vol 5529) (Springer Berlin Heidelberg) pp 256–268
[14] Jayalal M L, Rajeswari S and Murty S A V S 2009 Application of yaim tool in grid computing Tech. rep. Superintendents Advisory Committee on Enrollment and Transfers (SACET)
[15] Fenton N 1990 *Journal of Software Engineering* **5** 65–78
[16] Rodriguez J, Perez A and Lozano J 2010 *IEEE Transactions on Pattern Analysis and Machine Intelligence* **32** IEEE Computer Society
[17] Lo J H 2010 *The 2nd International Conference on Computer Research and Development* (Kuala Lumpur, Malesia) pp 765–769

WNoDeS: a virtualization framework in continuing evolution

V. Ciaschini and S. Dal Pra and G. Dalla Torre and E. Ronchieri and D. Salomoni

INFN CNAF, Viale Berti Pichat 6/2, 40126, Bologna, Italy

Abstract. INFN CNAF hosts the so-called INFN Tier1 that provides computing and storage facilities to high energy physics community and several multi-disciplinary experiments. Optimizing the use of computing resources is therefore essential. This is one of the reasons why WNoDeS is being developed and progressively put in production by INFN Tier1. It is a framework to virtualize computing resources, built on top of batch system such as LSF, allowing full integration with existing scheduling, security policy and monitoring. WNoDeS supports interaction with user requests through traditional batch or Grid jobs.

In this report, we are going to describe the main activities performed on WNoDeS to improve its reliability performing its re-engineering whenever appropriate. In 2013, WNoDeS proved itself able to provision cloud computing resources. We documented the experiences done with various communities, such as bio-informatics and astro particle physics, to highlight its strengths and weaknesses.

1. Current State

In 2013 we performed WNoDeS work on two different lines of development:

- validation of WNoDeS dynamic execution host provisioning for production by several experiments[1]
- service stability and scalability improvements and resource usage optimization

In 2013 we concluded two important collaborations for WNoDeS: the former was with the EMI project [1], where we mainly introduced software engineering practices in the development of WNoDeS; the latter was with the EGI Cloud Federation Task Force [2], where we tested the prototypical feature of WNoDeS to support users' Cloud requests through both a pure command line interface and the IGI Web portal[1]. We presented the Cloud prototype at the EGI Community Forum 2013 [3].

1.1. Service Validation

WNoDeS has been tested for validation by two different research communities: WeNMR and auger.

[1] EMI website, `http://www.eu-emi.eu/`

[2] EGI Cloud Federation Task Force, `https://wiki.egi.eu/wiki/Fedcloud-tf:FederatedCloudsTaskForce`

[3] Egi community forum 2013 website, `cf2013.egi.eu`

WeNMR validated the WNoDeS infrastructure into its own CING network of machine, where newly created WNoDeS virtual machines start off an image containing the full CING software suite, which would have been difficult to install on a new machine, connect to the ToPoS pool framework and pull and execute jobs off of it, killing themselves at the end of computation. In this case WNoDeS provided an easy way to create additional execution hosts without all the overhead of installing and validating a new instance.

Auger instead uses WNoDeS to create execution hosts, which are fully integrated in the Tier1 LSF batch system, to performance considerations. Auger jobs require access to a global MySQL DB. Providing such access on a GPFS file-system leads to poor performances, due to an essentially random access pattern. Instead, executing hosts on VMs and hosting the DB on the hypervisor greatly reduce the stress on the file system, allowing better overall efficiency.

1.2. Stability work

In 2013, development work on WNoDeS mainly focused on increasing the stability and reliability of the code as well as reducing its computing and networking footprint and increasing its debuggability. To achieve this objective, we re-engineered large parts of WNoDeS's internals to provide a richer, more expressive set of logs, as well as a large set of bug fixes applied transversely through all the code, some trivial, some extremely involved. In general, WNoDeS takes much more advantage of LSF's characteristics and strives to be much more lightweight than previous versions were.

2. Future Work

New work on WNoDeS focuses on stability and usability improvements and on virtual machine instantiation.

Usability improvements range from a greater ease of installation and configuration to better reporting tools for discovering the state of the WNoDeS-enabled nodes.

The ability to delegate actual instantiation of virtual machines to external managers providing well-known interfaces, like OpenStack's EC2 interface is also under active development, The objective is to have alternatives to direct handling of VMs via libvirt, which is a little too tied to specific OS versions to be comfortably used.

References

[1] E. Ronchieri, M. Verlato, D. Salomoni, G. Dalla Torre, A. Italiano, V. Ciaschini, D. Andreotti, S. Dal Pra, W. G. Touw, G. Vriend, G. W. Vuister, Accessing Scientific Applications through the WNoDeS Cloud Virtualization Framework, ISGC 2013: International Symposium on Grids and Clouds, 17-22 March, Academia Sinica in Taipei, Taiwan.

[2] E .Ronchieri, A .Italiano, G. Dalla Torre, D. Salomoni, D. Andreotti, M. Caberletti, V. Ciaschini, Distributed open cloud computing, storage and network with WNoDeS: esperienza ed evoluzione, Selected full paper for Workshop GARR, 29-30 November, Rome, Italy Calcolo e Storage Distribuito.

CNAF activities in the MarcheCloud project

E. Fattibene, M. Manzali, D. Salomoni, P. Veronesi

INFN CNAF, Viale Berti Pichat 6/2, 40126, Bologna, Italy

E-mail: enrico.fattibene@cnaf.infn.it, matteo.manzali@cnaf.infn.it, davide.salomoni@cnaf.infn.it, paolo.veronesi@cnaf.infn.it

Abstract. The MarcheCloud project aims to deploy a Cloud infrastructure based on open-source technologies for the Regione Marche Local Public Administration and represents one of the pilot experiences at National level. The experience grown up at CNAF in the operations of computing and storage services and in the deployment of Cloud facilities have been made available to the goal of the project. The MarcheCloud IaaS (Infrastructure as a Service) is based on OpenStack, an open source product that can be executed on open source platforms and has strong support from the industry. On top of this infrastructure, a pilot application, that provides an electronic medical records service, has been deployed. This paper presents the activity carried on by CNAF in the framework of the MarcheCloud project, such as the design study of the architecture, the deployment and the operations of the pilot infrastructure and the training events carried out.

1. Introduction

The MarcheCloud pilot Project started in mid-2012 as a joint collaboration among the Local Public Administration Regione Marche, INFN, University of Camerino and Polytechnic University of Marche. The project starts from the will of Regione Marche to equip a Cloud computing infrastructure able to provide innovative technological services to citizens, public institutions and enterprises, fostering:

- efficiency and innovation, productivity growth;
- new services development;
- business opportunities for the local geographical area;
- fulfilment of important economies of scale using public and private resources;
- circulation of advanced expertise in the ICT sector.

These principles, at the core of Open Government initiatives [1], have found concrete application in the pilot project of a Cloud architecture based on the XaaS paradigm, thought for the deployment of welfare services to citizens. In particular, the Cloud infrastructure has been used to host a specific application that manage data from a network of regional clinical laboratories, in order to provide an electronic medical records service. The service provides citizens with a single point of access to their chemical and physical analysis, regardless of the laboratory used, through a variety of channels (web browsers, smart phones and Android SmartTV), allowing to remotely access results, view, organize and analyze them. The service also allows Regione Marche to optimize the use of computational resources.

2. The MarcheCloud Cloud infrastructure

Within the project, CNAF had the task of designing and deploying the IaaS.

The overall architecture of the MCloud project is depicted in Figure 1. The OpenStack framework [2] was chosen to implement the IaaS infrastructure at the core of the Marche Cloud project. OpenStack, in particular, is an open source product that can be deployed on open source platforms; it has strong backing from the industry, with major ICT players directly supporting it; it enjoys a steady growth in terms of both functionalities and developers; it has an open and extensible architecture, mainly written in Python; it interoperates with other Cloud stacks and APIs; there is significant experience with OpenStack deployment, configuration and extensions within the INFN and in particular at CNAF.

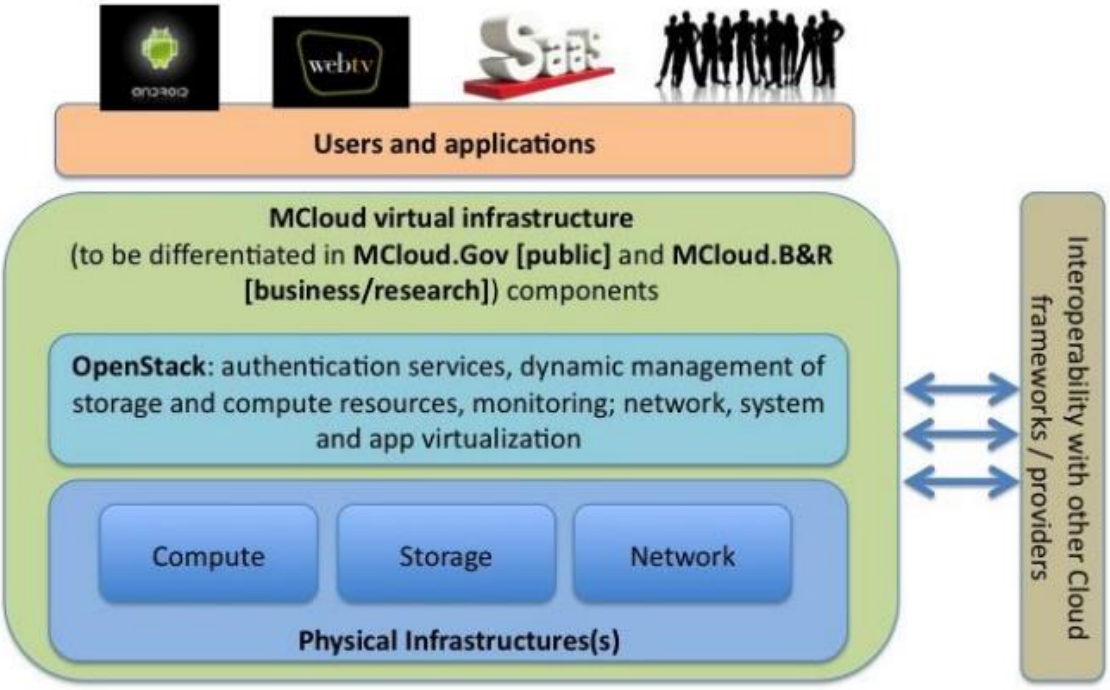


Figure 1: Architecture of the MarcheCloud project

2.1. The first prototype

At the end 2012, a first prototype of the infrastructure, based on the *Folsom* version of the OpenStack software, was deployed. It was decided to focus on some OpenStack components, namely: Dashboard, Compute, Network, Image repository, Authentication. The shared file system chosen for the MarcheCloud infrastructure is GlusterFS [3]. It was configured so that it could be used both to store virtual images in an highly-available configuration, and to define a common storage area across all compute nodes. The GlusterFS deployment was also configured to implement automatic fail-over in case of problems to one of the GlusterFS servers. The disks local to the computing nodes were used to define the IaaS storage volumes.

2.2. The 2013 evolution

During 2013, OpenStack released two new versions, codenamed *Grizzly* and *Havana*. The first evolution of the MarcheCloud IaaS, then, involved the migration from OpenStack Folsom to OpenStack Grizzly. The architecture was reworked to add higher flexibility especially at the network layer, where a mix of per-tenant, private networks, together with external networks (used to

masquerade private VMs and make them accessible to the outside world) and shared networks (used to connect the OpenStack-based MarcheCloud deployment with existing legacy systems like VMware vSphere and Proxmox clusters) was defined. In addition, the MySQL database used by OpenStack was made redundant creating a MySQL cluster; finally, the GlusterFS cluster was connected to volumes derived directly from the existing Storage Area Network of the Regione Marche computing center. Since the first inception of the project, OpenStack made significant advancements not only with the introduction of additional features, but also in terms of manageability and ease of installation. While the initial Folsom release was installed with long and error-prone manual procedures, the Grizzly version benefitted from the automated Puppet-based, OpenStack supported deployment utility called Packstack [4], which drastically reduced provisioning time for the OpenStack IaaS layer from days to hours. It is important to note that, while it is still not possibile to easily migrate from one OpenStack version to another with zero downtime, the simplification of the installation procedures made it easy to test and deploy a Grizzly-based cluster alongside the Folsom one, validate it, and eventually migrate the Folsom state (users, applications, VMs, ad-hoc components) into the Grizzly cluster, which became then the final production cluster. OpenStack-based MarcheCloud deployment with existing legacy systems like VMware vSphere and Proxmox clusters) was defined.

3. Training activity

In the year 2013 CNAF was involved in different training events in the framework of the MarcheCloud project. In April an event comprising both theoretical and hands-on sessions was held in Ancona; in those days, the CNAF experts trained the technical personnel of Regione Marche to the main concepts of Cloud computing, to the functionalities of the OpenStack framework and, thanks to the practical sessions, to the installation and management of an IaaS similar to that running in production.

During autumn 2013 the CNAF personnel involved in the project participated as trainer to two different courses held at the Regione Marche headquarter in Ancona. Both courses have been supplied in two editions and were addressed to people working in enterprises and local public entities in the geographical area of Marche territory. The first event consisted of frontal lessons concerning the main Cloud computing concepts, the OpenStack overall architecture, the MarcheCloud infrastructure and services supplied by these technologies. The second course was more technical, treating in detail the OpenStack project and functionalities and the MarcheCloud solutions; moreover, the hands-on session made the attendees able to perform basic and advanced user actions within an OpenStack-based IaaS.

References

[1] D. Lathrop, L. Ruma (Eds.), “Open Government: Collaboration, Transparency, and Participation in Practice , O’Reilly Media Inc. 2010.
[2] OpenStack, http://www.openstack.org/
[3] GlusterFS, http://www.gluster.org/
[4] Packstack, https://wiki.openstack.org/wiki/Packstack

EMI Testbed Improvements and Lessons Learned from the EMI 3 Release

F Capannini[1], B Hagemeier[2], A Elwell[3], C Bernardt[4], M Kocan[5], F Dvorak[6], D Dongiovanni[1], C Aiftimiei[7], A Ceccanti[1] and A Cristofori[1]

[1] INFN-CNAF, Viale Berti Pichat 6/2, I-40127 Bologna Italy
[2] Forschungszentrum Jülich GmbH, 52428 Jülich, Germany
[3] CERN Geneve, Switzerland
[4] DESY Notkestraße 85 D-22607 Hamburg, Germany
[5] Faculty of Sciences, UPJS Jasenna 5 040 01 KOSICE - Slovakia
[6] CESNET University of West Bohemia, Univerzitni 20, 306 14 Plzen, Czech Republic
[7] INFN Padova, Via Marzolo 8, I-35131 Padova, Italy

E-mail: fabio.capannini@cnaf.infn.it

Abstract. The European Middleware Initiative (EMI) Project has succeeded in merging into a set of releases (EMI 1 Kebnekaise, EMI 2 Matterhorn) more than fifty software products from four major European technology providers (ARC, gLite, UNICORE and dCache). To satisfy end user expectation in terms of functionality and performance, the release process implements several steps of certification and verification. The final phases of certification are aimed at harmonizing the strongly inter-dependent products coming from various development teams through parallel certification paths. This article introduces the new approach in the design and management of the release process envisaged for the EMI 3 release according to the requirement of a more effective integration strategy emerged during the first two EMI releases.

1. Role of central testing facilities as part of EMI release cycle and quality assurance activities

The evolution of EMI software products in order to fix software errors or implement new features follows a defined release cycle, resulting in both monthly release updates (minor releases whilst not breaking backward compatibility) and yearly major releases. To these periodical releases we add revision (fixing defeats without introducing new features) and emergency updates (fixing problems with top priority, generally related to security). The release process periodically cycles over five macro phases:

(i) phase 1-) Requirements analysis phase: inputs collected from EMI user communities representatives are translated into accepted technical requirements;
(ii) phase 2-) Development and test planning phase: technical requirements from phase 1 bring to new development and test plans;
(iii) phase 3-) Development, testing and certification phase: new products versions are developed according to test plans and tested by product teams;
(iv) phase 4-) Release certification and validation phase of new products candidates;
(v) phase 5-) Release and maintenance;

The EMI quality assurance work package is in charge of those activities aimed at harmonizing the parallel work of the developer product teams to obtain as output a single homogeneous EMI release. Therefore, among quality assurance duties we have the definition and monitoring policies, definition and collection of metrics and keys performance indicators (KPIs), quality control verification and reporting, the provision of common tools for products building and the implementation of common and shared infrastructural and operational resources for product inter-component and large scale testing. The present work focuses on this last working area in EMI quality assurance, which is strictly related to the phase 4 and 5 of release cycle. Given the framework described above, we can summarize the role of central testing infrastructure team as a provider of all facilities and certification activities assuring that the sum of components certified in isolation constitutes an EMI release of products deployable from single repository and consistently inter-operating.

2. Overview of EMI Products certification testing

Each EMI component follows an independent path from other components life cycle during the definition of requirements, development/test planning, source coding, build until the component certification in isolation. Then, after certification in isolation has been accomplished, the various component release paths must intersect in order to verify that all components can consistently inter-operate. This means that the logically unitary phase of component certification, aimed at verifying the expected functioning under production environment, is actually split in two separate steps across EMI release cycle phases 3 and 4. Moreover, component certification in isolation during phase 3 is performed on product team resources while inter-component testing performed during release phase 4 occurs on central testbed resources. To give an overview of the types of tests performed in each phase we mention:

(i) Release phase 3: static code analysis, installation tests during repackaging phase in the mock image to verify run time dependencies, deployment tests on product teams local resources, unit tests, functionality and regression tests of the component in isolation;

(ii) Release phase 4: deployment tests on central resources, functionality tests validating EMI components mutual interaction;

(iii) Performance and scalability tests are not mandatory and may occur both in phase 3 and 4 of EMI release cycle.

After a quality check verification step, formally controlling product compliance with agreed release policies and guidelines, the release components candidate for the considered update are deployed on the inter-component testing infrastructure. All component release candidates must pass an inter-component testing certification step to enter the next EMI update. This inter-component testing is performed on EMI central inter-component testing infrastructure instances.

Integration testing is the part of testing and certification process of a software product where the product pieces of functionality and expected behavior are tested against other related grid service components. In EMI decentralized software development model, testing and certification are in charge of different Product Teams, each responsible for one or more software components. Taking place after functional testing of products in insulation has been successfully carried out, integration testing then represents the first centralized point of contact among different products.

3. Improvements in the EMI 3 release testing scenario

EMI 1 and EMI 2 releases have shown that many issues were only discovered during the deployment on the EMI testbed phase and that integration testing for EMI-2 was not implemented successfully due to delays in having the testbed in shape since the products were in many cases not deploy-able. The new strategy conceived for EMI 3 release foresees a two steps deployment scenario. In the first phase the different product teams are responsible of

providing and maintaining an instance of their service where the initial testing validation is carried out. During this phase the SA2 people responsible for the testbed would take care of hosting and configuring a set of "core" services, which lay the foundation of the integration testbed (i.e. voms, top-bdii) that other services will reference during the interoperability and integration testing campaigns. Ideally, the resources are monitored with a nagios instance hosted and maintained by SA2. In a second phase of the testing process the services are deployed on the central testbed and a new set of tests is performed centrally as a second independent check.

4. Impact of the new testing strategy

Releasing software with high production level quality, i.e. satisfying end user expectation in terms of functionality and performance, is the final goal of every software collaborative project. Software continuous and effective testing is then a key step in the software development process to match quality targets. The new testing strategy designed for EMI 3 release should guarantee a more efficient and complete testing activity of the software, eventually resulting in increased product quality and control over the process. The new approach should allow to shorten the delays experienced in the release process thanks to early identification of the problems occurring at the interface between development and testing, thus resulting in a better coupling of the two activities.

Accessing Grid and Cloud Services through a Scientific Web portal

Diego Michelotto, Marco Bencivenni, Andrea Ceccanti, Daniele Cesini, Enrico Fattibene, Giuseppe Misurelli, Elisabetta Ronchieri, Davide Salomoni, Paolo Veronesi, Valerio Venturi, Maria Cristina Vistoli

INFN-CNAF

E-mail: diego.michelotto@cnaf.infn.it

Abstract. Distributed Computing Infrastructures have dedicated mechanisms to provide user communities with computational environments. Concerning the Grid, X.509 certificate is the standard implementation of the authentication component. It grants an appropriate level of security; however, users perceive it as an imposing restriction on the adoption of Grid technology. In addition, the complexity of Grid middleware creates demanding practices to support users' requests. In our opinion, the Web interaction represents the mainstream to overcome these drawbacks and engage new communities.

In this paper, we present a solution to overcome the aforementioned limitations by providing users with several Grid and Cloud services (such as job submission, compute provisioning, and data management) accessible through a community oriented Web portal. Indeed, we have developed the portal within the Italian Grid Infrastructure framework where the major national user representatives influenced its design, the implemented solutions and its validation by testing some specific use cases.

1. Introduction

In the context of the Italian Grid Infrastructure (IGI) [1], we collected requirements from several communities (such as computational chemistry, astronomy, and earth-science) and investigated their particular computing needs. Our aim was to satisfy these users' needs and to hide the X.509 complexity by using a third-party application to generate, store, and manage personal certificates. The paper presents the IGI Web portal, based on the Liferay framework designed and developed to enable scientists to avail themselves of Grid and Cloud resources. We reused third-party components such as MyProxy, WS-PGRADE, Distributed Infrastructure with Remote Control (DIRAC) [2], and Grid User Support Environment (gUSE) [3]. The design of the Web portal is general and suitable for different DCIs, and its implementation refers to the guidelines provided by the Italian and European Grid Infrastructures (IGI and EGI [4] respectively). In this spirit, we also implemented services by using standard modules, called portlets, to easily extend the portal to new functionalities. The portal Grid components are already in production, whereas the Cloud ones are in a prototypical evolving form. Since May 2013 the portal has supported 116 users, distributed in 30 VOs, who have submitted 17.8k jobs via DIRAC and 3.4k jobs via WS-PGRADE for a total amount of about 380k CPU-hours.

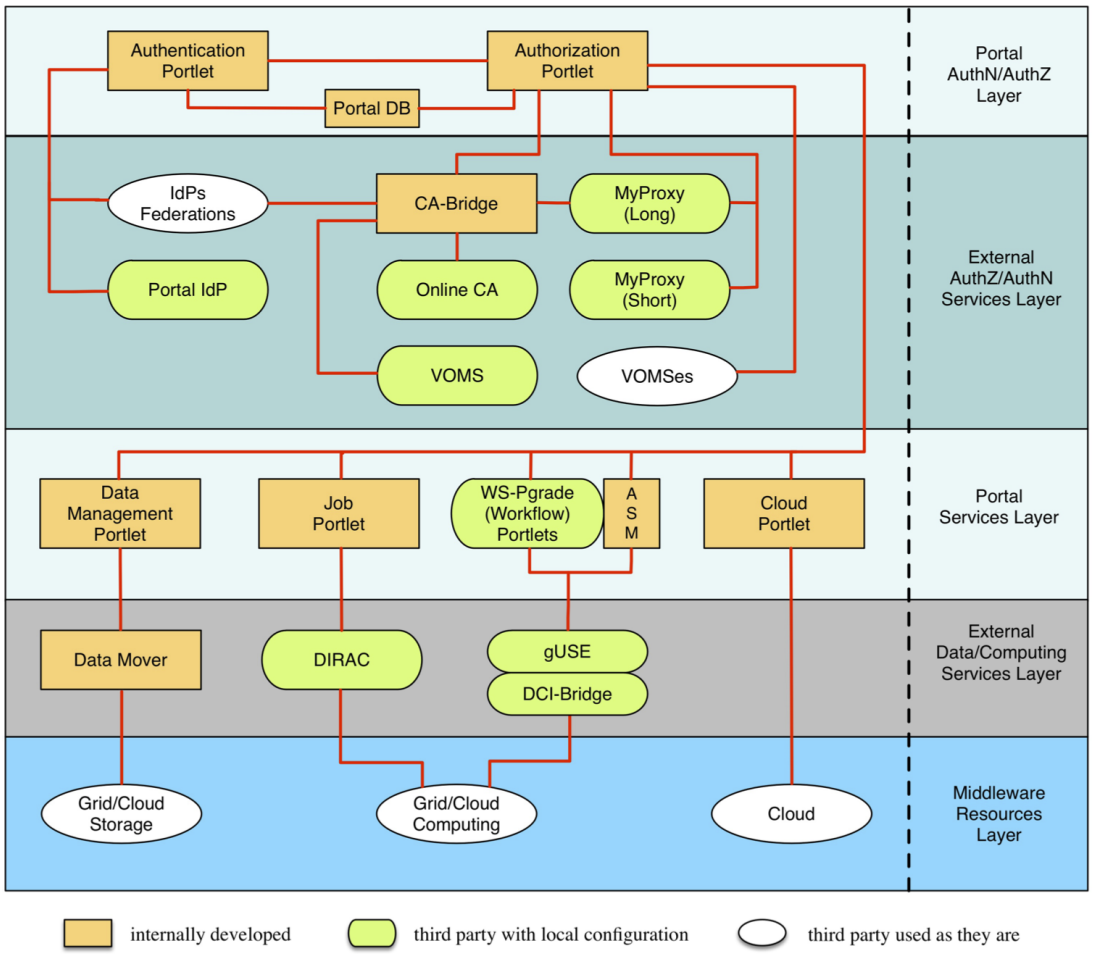


Figure 1. The IGI Web portal architecture: rectangular shapes identify internally developed components, rounded rectangular shapes show third-party components with local configuration, and finally oval shapes represent third-party components used as they are.

2. Architecture

We designed the portal architecture with the aim to reuse existing and maintained open solutions for all the features mentioned in Section 1. In addition, we adopted Liferay - a popular open source portal and collaboration software made of functional units called portlets - to ensure a modular Web portal structure. The portlets are components that build dynamic Web contents and use the JSR168 [5] and JSR286 [6] standards.

The IGI Web portal architecture, presented in Fig. 1, can be conceptually divided into five main layers.

At the highest level, the Portal AuthN (Authentication) and AuthZ (Authorization) Layer verifies all the mandatory credentials provided by the users: X.509 certificate, Virtual Organization (VO) membership and Identity Provider (IdP) trusted by the portal.

At the second level, the External AuthZ and AuthN Services Layer supports the upper level in a set of operations:

- Users' credentials vetting leverages the IdPs federations and VOMSes (VO Membership Service) [7] components invoking single-sign-on or membership-credential requests.

- Providing missing credential exploits the portal IdP, online CA, and VOMS components acting as fallback and supplying an alternative credential chain.
- Storing delegated credentials (proxies) makes use of the MyProxy [8] components that manage short and long proxies.

In addition, the CA-bridge component is paramount for the interconnection among the AuthZ portlet and the other components involved in this layer, performing all the necessary steps to validate users identity, to provide a X509 certificate, to generate the Grid credentials and to archive user data. The CA-bridge, MyProxy server and online CA components implement a certificate provisioning service integrated in the portal framework. The online CA supplies MICS (Member Integrated Credential Services) certificates [9] with a 13-months validity.

At the middle level, the Portal Services Layer contains all the portlets implementing data management, job submission, application and workflow submissions, and cloud provisioning. While a job is a sequence of key pairs (attribute, value) based on the JDL, a workflow is a sequence of connected jobs where the execution of one or more steps depend on the results of the previous ones. WS-PGRADE consists of various portlets (following the Liferay supported standards) providing different functionalities ranging from user registration to data management up to workflow handling; however, we only used the workflow feature. The Application Specific Management (ASM) portlet builds application-specific interfaces and uses the workflows defined in WS-PGRADE. Both WS-PGRADE and ASM interact with the Grid and Cloud user support environment (gUSE). The Job and Cloud portlets allow the portal to interact with the computing Grid and Cloud resources, while the Data Management portlet interfaces with the storage resources.

At the second to last level, the External Data and Computing Layer provides the tools to handle data required by the portlets in the Portal Services Layer. The Data Mover component allows users to transfer data among Grid resources, hiding all the complexity of accessing data so that users can spend no effort to learn Grid data tools. Then, the DIRAC component integrates heterogeneous computing resources and provides solutions for submitting jobs to Grid and Cloud infrastructures and for managing data. Finally, the gUSE and DCI Bridge components assist the workflow submission.

At the lowest level, the Middleware Resources Layer consists of the Grid and Cloud middleware components to provide physical and virtual resources. The interaction with the Grid services is already in production, whereas the Cloud one is in a prototype version and still in pre-production.

3. AuthZ and AuthN Details

The two AuthZ and AuthN Layers (described in Section 2 and shown in Fig. 1) determine the first access to the portal where users through registration must provide the mandatory qualifications to access Grid or Cloud services.

During the authorization phase users must provide the following information: a X.509 certificate issued by a trusted EUGridPMA-member CA, the affiliation to a recognized VO and a trusted IdP. If a user only has a subset of these credentials, the portal could use them to provide the missing ones. Table 1 shows the credentials combinations required to successfully conclude the registration phase: the **x** symbol specifies a missing information; the **o** symbol states an available information.

The request for a personal certificate and its management (such as storing and renewal) often represent tedious operations that many users wish to avoid. We addressed this issue: by interfacing the portal with an online CA, which provides X.509 certificates to users authenticated by a federated identity management system; and by implementing a service to manage these certificates on behalf of the users. Depending on their credentials, users can select whether to upload their certificate or ask for a new one through the portal.

Table 1. The required credentials for the registration phase.

Case	IDP member	X.509 certificate	VO
1	**x**	**o**	**o**
2	**o**	**x**	**x**
3	**o**	**o**	**o**

During the authentication phase, the portal leverages a federated authentication mechanism - based on the Security Assertion Markup Language (SAML) [10] - to offload the portal from managing users' credentials and to exploit a single-sign-on solution. Indeed, the portal trusts all the IdPs belonging to the EDUgain federation [11] that interconnects distributed identity federations around the world; therefore all members of these IdPs can access the portal components by using the credentials contained in their own organization IdP. Since Liferay natively supports the Central Authentication Service (CAS) [12] but not SAML, the portal uses the Casshib [13] software; this enables the CAS server to act as a Shibboleth service provider [14].

4. Data and Computing Services Details

The Portal and External Data/Computing Services Layers (described in Section 2 and shown in Fig. 1) bridge users towards Grid and Cloud resources covering all the requirements described in Table 2. The computing services scenarios comprise the submission of: simple Grid jobs with binaries and input data; workflows with complex use cases; specific applications oriented to custom interfaces; and Cloud provisioning with IaaS (Infrastructure as a Service) allocation. The data service scenario abstracts moving data asynchronously among Grid Storage Elements (SEs) in a drag and drop way.

Table 2. The requirements collected from our user communities.

Services	Requirements
Authentication and Authorization	Single-sign-on based authentication Personal certificate handling Certificate provisioning on demand
Workload Management	Managing workflows Detailing Job Description Language (JDL) customization Getting customized Web interface applications
Data Management	Simplified access to data storage Simplified data moving

4.1. Simple Grid jobs

Users submit jobs by uploading their executables and input files and by adopting JDL to specify the executable parameters. The portal implements this scenario by interfacing the Job portlet with a multi-VOs configured DIRAC server. In this context, we developed a portlet that uses the DIRAC command line interface, handling the communication between the portal and the DIRAC server. In addition, it allows users to: build their JDL by selecting the correct values from a list of attributes and settings; save JDLs as templates for sharing and reusing purposes; show the list of submitted jobs; monitor the job state during its execution; retrieve the output; resubmit an ended job; log files in the end.

4.2. Workflows

The workflow submission is a step-by-step procedure for performing complex computations on different resources optimizing the overall task. Each step represents a specific portion of the entire calculation - identifying what we call a job. It can follow conditional constraints according to the evolution of the computation. The adopted acyclic workflow structure can assume a simple or complex form in relation to the addressed problem. By combining WS-PGRADE and gUSE the portal allows creating, managing and submitting both types of workflows.

4.3. Specific applications

Some communities may need to submit ad-hoc applications exploiting job and workflow submissions according to VO-specific requirements. In this case, we have adopted the following procedure to adapt existing applications to the Web portal: first, we analysed the portability of the application on the Grid; then, we checked hardware and software requirements (such as RAM, number of CPU cores, and specific libraries) as well as required input and output files, and application parameters needed for retrieving log and output files at runtime. In addition, we created a script that takes care of getting users' inputs, parameters and files, executing the application and retrieving the output produced during the entire computation. If the application required a job submission, we defined an ad-hoc JDL template, included in the job portlet and available to the users for the submission of the job; otherwise, we created a custom workflow in the WS-PGRADE framework and we implemented a Web interface, provided by the ASM portlet, which is directly available to the users for the workflow submission. In both cases, we included the mentioned script in the submission operation.

Inspecting produced files and monitoring the application are paramount items for long running applications. The job perusal solution (provided by WMS [15]) supports the first item but works only with small-size output files to avoid network overload. To overtake these limitations and to integrate application monitoring into the IGI portal, we developed a new mechanism, called application progress monitoring, exploiting Grid SEs and Storage Resource Manager (SRM) command line interfaces (CLIs) to make temporary and partial output files available to be inspected at runtime. The adopted SRM CLIs mechanism copies selected files from the computing resource where the job is physically running to a Grid SE that stores the files. Then users can directly access these SE files via the job or ASM portlet.

4.4. Cloud provisioning

Some communities may need to run their applications exploiting the Cloud paradigm. The portal implements this scenario by interfacing a Cloud portlet with a set of services that supply resources according to the IaaS provisioning model. The services, fed with a pre-built configuration file, can interact with various IaaS Cloud providers exploiting the benefits offered by existing Cloud platforms such as WNoDeS [16], OpenStack [17] and Opennebula [18]. The developed portlet exposes a Web interface for each service simplifying users' tasks to create and manage new instances.

To instantiate new virtual machines (VMs), users have to upload their own SSH public key [19], or generate a SSH private and public key pair through the portal. This is necessary to allow logging into a VM with root privileges without requiring any password. The keys generated by the portal can be retrieved by users in a secure way. Users can then create new VM instances choosing from a list of images preloaded in a repository. For each image it is possible to select the size of the cloud environment (defined according to the number of cores, memory and disk size) and how many instances have to be created at a time; each image has a range size that depends on the Cloud platform it belongs to. As soon as users create new instances, a list of their VMs is displayed together with information such as architecture, size and status. By selecting the instance name, users are automatically logged into the VM with

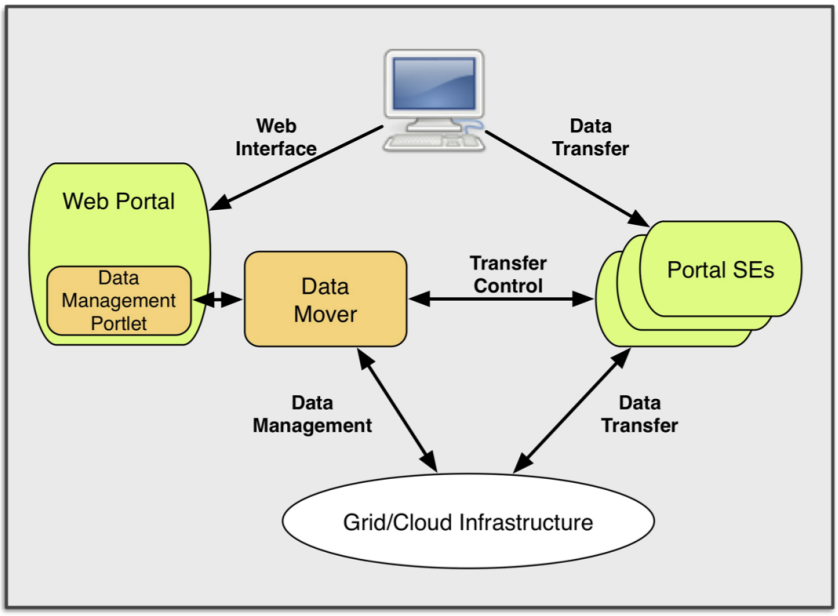


Figure 2. The IGI Portal data management service architecture.

root privileges through a Web terminal [20] that is part of the portal. The portal lets users select VM images offered by different Cloud providers via a repository, which access depends on the resource providers internal policies

4.5. Data management

In a standard Grid environment users may benefit from a set of command line tools to perform data management tasks such as copying files on a Grid SE, registering files in a LCG File Catalogue (LFC) [21], and replicating files on other Grid SEs. To save learning effort, we designed and implemented a data management service [22] sketched in Fig. 2. The service includes several elements intercommunicating in a secure way.

The Data Mover component controls and manages every step of the file transfer operations: uploading and downloading files. It controls data transfers through an external storage service - composed of a set of Storage Resource Manager (StoRM)-based portal SEs [23] acting as a cache memory for the files - until they are transferred to or downloaded from a Grid SE. The Data Mover is a Web P ydio-based data management service that implements and extends the functionalities of the Grid data management command line tools, exposing a Web interface that manages either Grid files or some other types of files. We developed a plug-in for handling data that by the IGI portal allows users to easily browse the content of the VO file catalogue and to perform operations on either the logical data (affecting the catalogue) or the physical files (involving the SE). Table 3 details the possible operations performed on data.

References

[1] Igi website. www.italianGrid.it/about.
[2] A Tsaregorodtsev, M Bargiotti, N Brook, A C. Ramo, G Castellani, P Charpentier, C Cioffi, J Closier, R Graciani, G Kuznetsov, Y Y Li, R Nandakumar, S Paterson, R Santinelli, A C Smith, M S Miguelez, and S Gomez. Dirac: a community grid solution. *Journal of Physics: Conference Series*, 119(062048), 2008.
[3] guse website. guse.hu.
[4] Egi website. www.egi.eu.

Table 3. The possible operations performed on data.

Data Types	Operations
Logical Data	Creating a new folder; Deleting an empty folder; Renaming a folder or file (changing the LFN); Moving a folder or file (changing the LFN); Getting detailed information about a file (LFN, GUID - Global Unique Identifier, list of replicas, owner, ACL); Sharing a file with other portal users.
Physical Data	Replicating files on different storage elements; Downloading files.
Catalogue Data	Deleting files; Uploading files.

[5] Jsr 168: Portlet specification, java community process. http://www.jcp.org/en/jsr/detail?id=168, 2005.

[6] Jsr 286: Portlet specification 2.0, java community process. http://www.jcp.org/en/jsr/detail?id=286, 2008.

[7] R. Alfieri, R. Cecchini, V. Ciaschini, L. dellAgnello, A. Frohner, K. Lorentey, and Spataro F. From gridmapfile to voms: managing authorization in a grid environment. *Future Generation Computer Systems*, 21(4), 2005.

[8] J. Basney, M. Humphrey, and V. Welch. The myproxy online credential repository. *Software: Practice and Experience*, 35(9):801–816, 2005.

[9] Tagpma, profile for member integrated x.509 credential services with secured infrastructure. http://www.eugridpma.org/guidelines/MICS/IGTF-AP-MICS-1.2-clean.pdf.

[10] Thomas Hardjono and Nathan Klingenstein. Saml v2.0 kerberos web browser, sso profile version 1.0. Technical report, OASIS, 2010.

[11] edugain website. www.geant.net/service/eduGAIN/Pages/home.aspx.

[12] Cas website. http://www.jasig.org/cas.

[13] Casshib website. https://code.google.com/p/casshib.

[14] Shibboleth website. http://shibboleth.net/products/service-provider.html.

[15] Marco Cecchi, Capannini Fabio, Alvise Dorigo, Antonia Ghiselli, Francesco Giacomini, Alessandro Maraschini, Moreno Marzolla, Salvatore Monforte, Fabrizio Pacini, Luca Petronzio, and Francesco Prelz. The glite workload management system. In *GPC*, volume 5529 of *Lecture Notes in Computer Science*, pages 256–268. Springer, 2009.

[16] Davide Salomoni, Alessandro Italiano, and Elisabetta Ronchieri. Wnodes, a tool for integrated grid and cloud access and computing farm virtualization. *Journal of Physics: Conference Series*, 331(5: Computing Fabrics and Networking Technologies), 2011.

[17] Openstack website. www.openstack.org.

[18] Opennebula website. opennebula.org.

[19] Daniel J. Barrett, Richard E. Silverman, and Robert G. Byrnes. *SSH, The Secure Shell: The Definitive Guide*. O'Reilly Media, 2005.

[20] Gateone website. http://liftoffsoftware.com/Products/GateOne.

[21] J.P. Baud and S. Lemaitre. The lcg file catalog (lfc). Technical report, CERN, 2005.

[22] M. Bencivenni, R. Brunetti, A. Caltroni, Ceccanti A., D. Cesini, M. Di Benedetto, E. Fattibene, L. Gaido, D. Michelotto, G. Misurelli, V. Venturi, P. Veronesi, and R. Zappi. A web-based utility for grid data management. *PoS(ISGC 2012)004*, 2013.

[23] L. Magnoni, R. Zappi, and A. Ghiselli. Storm: a flexible solution for storage resource manager in grid. In *the IEEE 2008 Nuclear Science Symposium (NSS-MIC 2008)*, Dresden, Germany, 19 – 25 October 2008. IEEE Computer Society.

Grid Operation Service

M. Bencivenni, D. Cesini, A. Cristofori, E. Fattibene, D. Michelotto, G. Misurelli, A. Paolini, P. Veronesi, M. C. Vistoli

INFN-CNAF

grid-operations@lists.cnaf.infn.it

Abstract. The CNAF Grid Operation service provides direct day-by-day operational support to the Italian Grid Infrastructure and it is a reference point for several other INFN experiments, national and international communities.

1. Introduction

Besides HEP the need for a general Grid Infrastructure is common to other research activities. INFN CNAF is not delivering computing and storage capacity in a shared Grid-like manner to LHC experiment only, but it is a reference point for several other INFN experiments, national and international communities.

The Grid Operation Service is started, in its current configuration, during the last two years of EGEE-III, when an EU Project EGI-DS (European Grid Initiative Design Study) was launched for preparing the transition to a sustainable grid infrastructure, based on the National Grid Initiatives (NGIs), in each country, and a central (mostly coordinating) part, called EGI.eu. The NGI's are legal entities with a sustainable structure that mobilize national funding, and ensure the operation of the national e-Infrastructure and its integration in EGI and with EGI.eu.

The implementation of the NGI and of EGI.eu is now taking place with the support of the EGI-InSPIRE Project co-funded by the EC with 25 M Euro for 4 years started in May 2010.

The Italian grid infrastructure supports activities in a vast range of scientific disciplines — e.g. Physics, Astrophysics, Biology, Health, Chemistry, Geophysics, Economy, Finance — and any possible extensions to other sectors such as Civil Protection, e-Learning, dissemination in Universities and secondary schools.

2. Human Resources

The service is provided at National level by several (~15 people) fixed term staff with post-doctoral instruction or equivalent work experience in computing. The staff core (~10 people) and the service coordinator is located at CNAF. The duration of the contracts and their renewal highly depend on National and European founds.

3. Service goals

Through various projects at both national (e.g. Grid.it and INFN-Grid) and international (e.g. DataGrid, EGEE and WLCG) levels, INFN has built one of the largest grid infrastructures in Europe. It is well integrated or interoperates with the most advanced grid infrastructures in the world and it is used by a wide variety of user communities.

The Italian Grid is recognized as a robust, secure and reliable infrastructure. It offers various services from monitoring to operation management, authorization and Virtual Organization management; the operation, control and support of the grid services relies on a robust structure where dedicated teams operate on daily shifts basis.

The Italian Grid, with its 52 sites (71% INFN) providing about 25000 cores (83% INFN) for a total amount of about 248000 HEPSpec for the computing power, 11PB of disk space (90% INFN) and 10 PB of tape space (100% INFN) , provides a consistent contribution to the European e-infrastructure and plays an important role in its overall operation and management.

This Grid Operation service provides the following Core Grid services with direct day-by-day operational support:

- VOMS services for several National and International Virtual Organizations (VOs)
  - VOMS is an EMI product of the security area representing an Attribute Authority (AA) that releases signed security credentials with information beyond pure identities
  - Supported VOs: infngrid, cdf, planck, compchem, enea, theophys, bio, gridit, virgo, inaf, argo, pamela, libi, pacs.infn.it, comput-er.it, ams02.cern.ch, superbvo.org, euindia, cyclops, compassit, ipv6.hepix.org, enmr.eu, tps.infn.it, eumed, euchina, glast.org, icarus-exp.org, net.egi.eu, gerda.mpg.de, igi.italiangrid.it, vo.ingv.it
- TOP-BDII service for the National and International Grid Infrastructure
  - The top-bdii is a mission-critical components in today's production grid infrastructures. It provides detailed information about grid services which is needed for various different tasks.
- Logical File Catalog for National and International Virtual Organizations (VOs)
  - The LFC File Catalog is a general file catalog solution used within the EGI Infrastructure by several VOs.
  - Supported VOs: argo, ams02.cern.ch, babar, bio, cdf, compchem, comput-er.it, superbvo.org, enea, glast.org, gridit, inaf, infngrid, libi, pacs.infn.it, pamela, planck, theophys, tps.infn.it, virgo, euchina, enmr.eu, euindia, cyclops, compassit, gerda.mpg.de, icarus-exp.org, igi.italiangrid.it
- WMS service for National and International VOs
  - WMS is an EMI product of the compute area representing a broker that is able to identify suitable computational resources and the application submissions to them
- General purpose Grid portal, which provides exclusively web graphical user interface access to Grid and Cloud resources for job submission, workflow definition, data management and accounting services.

INFN is also directly responsible for supporting Italian resource Center managers and users in case of problem. The EGI support infrastructure consists of a central part (Global Grid User Support) dealing with global issues and regional and topical subsystems inside various activities and in the NGIs. The central helpdesk also acts as a relay between the different areas of support. At National level, the Italian regional ticketing system is designed to allow the definition of many different support departments, support teams and roles, and features the main functionalities of a typical helpdesk system. The interoperation with GGUS is done through a Web service interface (for both export operations from the regional system to GGUS and import operations from GGUS). The system is a good support tool that makes user/site support tasks easy to manage and trace. The same Web service interface could be potentially deployed to interface the system with other helpdesks at the project or regional level within a country, and with other operational tools such as the regional dashboard.

From the security point of view, the activity of the Italian security working group has been mainly focused on the formal definition of the future IGI-CSIRT, the security group of the Italian NGI. The group is the main contact point between EGI-CSIRT, the Italian grid sites belonging to IGI and the Italian NREN CERT (GARR-CERT). IGI-CSIRT has also the responsibility to coordinate the Incident Response Procedures in the Italian region, provide support to the sites and collaborate to some of EGI-CSIRT's tasks.

4. Current work and achievements

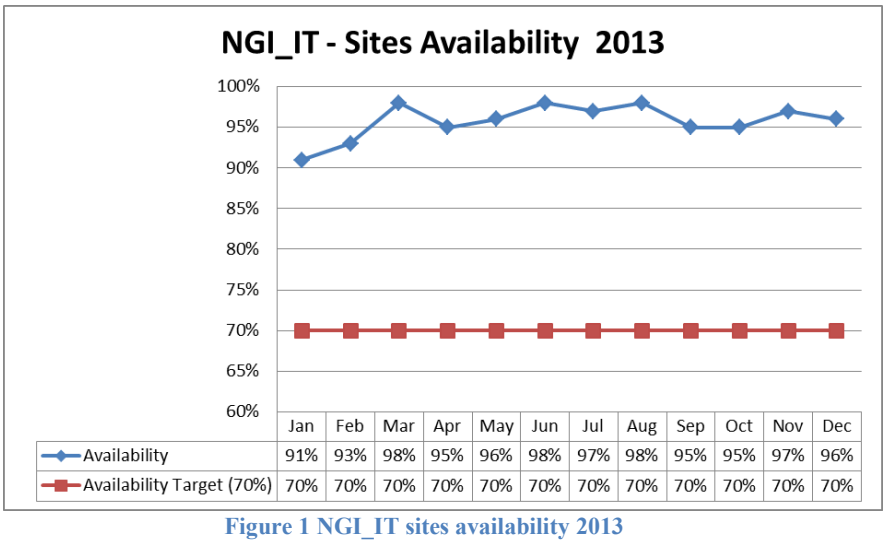
The activities of the Grid Operation service are mostly those agreed at international level with the participation in the EGI-Inspire project. IGI operates the Italian Grid ensuring its full compatibility

with the European Infrastructure, based on the compliance to the agreed shared policies, and taking also care of the support to Italian users.

The support activity is intended in a very broad sense, being comprehensive of the training and assistance to new users, the consultancy on adapting their applications to the distributed environment, as well as the development of instruments facilitating their approach to the grid infrastructure. In particular, those user communities without a strong IT background have usually difficulties approaching the default grid interfaces, consisting in a rather complex UNIX-like command line. In order to overcome this difficulties, high level web interfaces are developed, with the two-folded purpose of simplifying via a graphical interface the beginners' approach to the grid providing at the same time a seamless, location independent access.

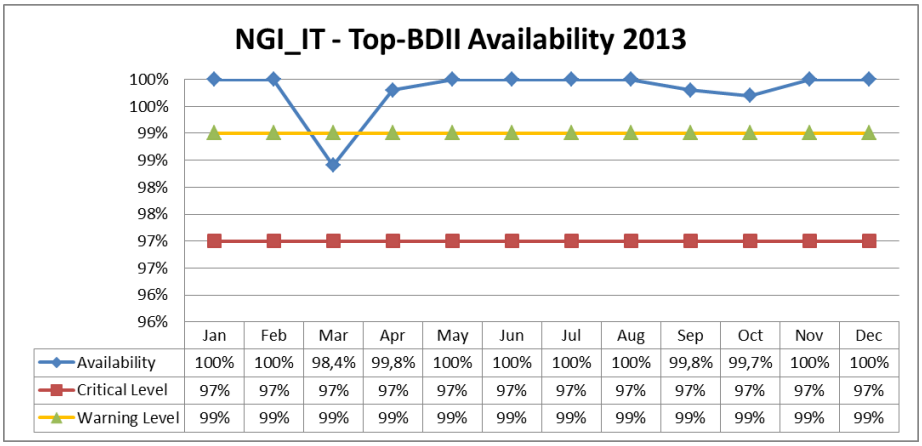
All these support and training activities are done in strong partnership not only with user communities interested in exploiting the grid capabilities, but also in collaboration with other support teams disseminated around Europe. The interaction are mainly mediated by EGI through the so called NGI International Liaisons and Virtual Team frameworks (https://wiki.egi.eu/wiki/Virtual_team) created at the European level in order to improve the efficiency and flexibility of the interaction between the NGIs.

The Italian Service Level Agreement, based on the EGI SLA, requires that production resource centers provide a minimum availability and reliability of 70% and 75% respectively and that resource centers whose availability is less than 75% for three consecutive months be suspended from the production infrastructure. Operational reliability and availability is computed on a monthly basis, site quality performance affects the related global regional statistics depending on the relative weight of the installed compute capacity in a region. The collection of monthly availability and reliability statistics is a service provided to users and operators to assess the level of functionality achieved by the infrastructure.



	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Availability	91%	93%	98%	95%	96%	98%	97%	98%	95%	95%	97%	96%
Availability Target (70%)	70%	70%	70%	70%	70%	70%	70%	70%	70%	70%	70%	70%

Figure 1 NGI_IT sites availability 2013



	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Availability	100%	100%	98,4%	99,8%	100%	100%	100%	100%	99,8%	99,7%	100%	100%
Critical Level	97%	97%	97%	97%	97%	97%	97%	97%	97%	97%	97%	97%
Warning Level	99%	99%	99%	99%	99%	99%	99%	99%	99%	99%	99%	99%

Figure 2 NGI_IT Top-BDII availability 2013

In Italy, availability and reliability [1] have been consolidating, both from a user perspective and from an operational point of view. Figure 1 shows the total NGI_IT availability from January 2013 to December 2013. Figure 2 shows the Top-BDII availability, one of the most important and critical service provided by NGI_IT and used at National and International level, during 2013. It can be noticed that Italian values are very good for both metrics and typically exceed their target.

References

[1] Source: EGI Resource Centres and RP Top-BDII Availability and Reliability https://wiki.egi.eu/wiki/Availability_and_reliability_monthly_statistics#Resource_Centres

Middleware support, maintenance and development

A. Ceccanti, V. Venturi, D. Andreotti, E. Vianello
INFN-CNAF, Bologna, Italy

E-mail: andrea.ceccanti@cnaf.infn.it

Abstract.
INFN-CNAF plays a major role in the support, maintenance and development activities of key middleware components (VOMS, StoRM, Argus PAP) widely used in the WLCG and EGI computing infrastructures. In this report, we discuss the main activities performed in 2013 by the CNAF middleware development team.

1. Introduction

The CNAF middleware development team was composed, in 2013, by four persons dedicated to the support, maintenance and evolution of the following products:

- VOMS [3]: the attribute authority, administration server, APIs and client utilities which form the core of the Grid middleware authorization stack;
- StoRM [7]: the lightweight storage element in production at the CNAF Tier1 and in several other WLCG sites;
- Argus Policy Administration Point (PAP) [6]: the Argus administrative interface and policy repository.

The main activities for the year centered around support and maintenance of the software and on the reorganization of the development and release processes as a consequence of the end of the EMI project [18] in May 2013.

2. The software development and maintenance process

In the beginning of 2013 we started organizing our development and maintenance work following the SCRUM agile process [11]. The objective was to have a faster release cycle centered around 3-weeks sprints and to track all activities consistently in order to know where most of the team effort was spent. We used the INFN internal tracker for the products backlog and to track other support and maintenance activities [12, 13, 14]. The move to our internal tracker was also motivated by the fact that support for the former issue tracker used for the middleware (CERN's Savannah [27]) was being discontinued.

Given that all maintained software was moved, or being moved, to Github [9] we decided to adopt the Gitflow branching model [19], in order to have a flexible and understood way of handling changes and organizing development. At this time we also introduced internal code reviews leveraging git pull requests and the nice collaboration tools provided by Github.

As of today, we can say that adopting the agile methodology and modern development tools boosted our productivity and allowed us to have faster release cycles that resulted in improved stability of the developed products.

2.1. Continuous Integration

During 2013, significant effort has been devoted to the realization of our continuous integration (CI) and testing system, based on the Jenkins CI server [8]. This system was meant to replace the ETICS system [24] that was used for all middleware builds during the EMI project.

The Jenkins server was configured with build nodes for the main supported platforms (Scientific Linux 5 and 6, Debian 6), and configurations were created for all the software packages. The Github Webhooks [25] mechanism provides efficient integration between the code repository and our CI server, so that whenever a change is pushed to one of the managed repositories a new build job is started on our CI infrastructure.

Finally, we added automatic deployment test functionality, in order to check that the latest versions of our products install and run correctly on all supported platforms. These deployment tests are run nightly. The clean environment required for the deployment tests is provided by the integration with a local private cloud infrastructure based on OpenStack [10].

3. VOMS

Work on VOMS focused mainly on:

- daily support and maintenance activities;
- the full refactoring of the Java APIs based on the newly released EMI common authentication library (CANL [20]) and the development of the new Java-based VOMS command-line clients [5, 4];
- the development of a functional and regression testsuite, based on robot framework [21];
- the VOMS Admin transition to a standalone process running on an embedded Jetty container [32].

4. StoRM

In march 2013, an important refactoring activity was launched on the StoRM codebase to align it with the tools and processes already in place for other products (VOMS and Argus PAP) maintained by the group. This meant moving the codebase to Github, adopting modern build tools (maven [15]) for the backend server, rationalizing the packaging and including StoRM in the CI infrastructure.

During this transition, two test suites were developed to assess the service stability and scalability:

- a functional and regression testsuite, based on robot framework [16, 22]
- a load testsuite, based on the Grinder tool [17, 23]

In the final months of the year, the team focused on improving the performance of the StoRM GridHTTPS server in view of the ATLAS DQ2 to RUCIO file renaming campaign [29], and fixing a set of high priority vulnerabilities [28] reported by the EGI Software Vulnerability Team.

5. Argus

In 2013, work on the Argus PAP focused on daily support and maintenance activities and the preparation of the EMI-3 release, for which the main change was the adoption of the EMI CANL library [20].

6. Web presence

The VOMS and StoRM web sites [3, 7] have been migrated to Github pages [26], a free hosting service provided by Github. There are several advantages in using Github pages for the web presence of our products:

- no hosting costs;
- versioned content and automatic publishing: all content is hosted on a Github repository, whenever a change is pushed to the repository the new content is published on the web site;
- effective workflow: the same workflow used for reviewing changes in the code can be used for the website (i.e., forked preview repositories, pull requests);
- simple but powerful static content-management engine based on Jekyll [30], which allows for easy management and update of the content of the web site by editing markdown pages [31].

7. Support

The Global Grid User Support (GGUS) system is the main support access point for EGI and WLCG. In the EGI support model, requests are first routed to the Deployed Middleware Support Unit (DMSU), managed by EGI, which provides expert support by Grid service operators. If an incident cannot be resolved by the DMSU, it is escalated to specialized, third-level support, typically provided by middleware developers.

In order to have a clearer view on support requests assigned to our support units (VOMS, StoRM and Argus), we developed a simple web tool, termed ggus-monitor, that groups support requests per priority and status, so that it is always evident on which requests the support team should focus at any given time.

8. Future work

Besides ordinary support and maintenance, in the future we will focus on the following activities:

- Refactoring of the StoRM services, to reduce code-base size and maintenance costs, to provide horizontal scalability to all services and to decouple the WebDAV service from the SRM backend service;
- Evolution of the VOMS attribute authority for better integration with SAML federations;
- Continuous integration and delivery, by leveraging lightweight virtualization environments (e.g., Docker) for integration testing and simplified deployment in production.

Acknowledgements

This work is dedicated to the memory of our friend and guide Valerio Venturi, suddenly deceased on December, 25th 2013. His kind presence, phenomenal sense of humour and sharp technical vision are now missing ingredients in our daily working life.

References

[1] European grid Infrastructure `http://www.egi.eu`
[2] The Worldwide LHC computing Grid `http://wlcg.web.cern.ch`
[3] The VOMS website `http://italiangrid.github.io/voms`
[4] VOMS clients code repository `https://github.com/italiangrid/voms-clients`
[5] The VOMS Java APIs `https://github.com/italiangrid/voms-api-java`
[6] Argus authorization service website `http://argus-authz.github.io`
[7] StoRM website `http://italiangrid.github.io/storm`
[8] Jenkins `https://jenkins-ci.org/`
[9] GitHub `https://github.com/`
[10] Openstack `http://www.openstack.org`
[11] SCRUM `http://bit.ly/scrum-wikipedia`

[12] INFN issue tracker https://issues.infn.it
[13] StoRM on INFN JIRA https://issues.infn.it/jira/browse/STOR
[14] VOMS on INFN JIRA https://issues.infn.it/jira/browse/VOMS
[15] Apache Maven http://maven.apache.org
[16] Robot framework http://robotframework.org/
[17] The Grinder http://grinder.sourceforge.net
[18] The European Middleware Initiative http://www.eu-emi.eu
[19] The Git flow branching model http://nvie.com/posts/a-successful-git-branching-model/
[20] The EMI Common Authentication Library https://github.com/eu-emi/canl-java
[21] The VOMS clients testsuite https://github.com/italiangrid/voms-testsuite
[22] The StoRM testsuite https://github.com/italiangrid/storm-testsuite
[23] The StoRM load testsuite https://github.com/italiangrid/grinder-load-testsuite
[24] The ETICS system http://etics-archive.web.cern.ch/etics-archive/
[25] Github webhooks https://help.github.com/articles/about-webhooks
[26] Github pages https://pages.github.com
[27] CERN's Savannah issue tracker https://savannah.cern.ch
[28] StoRM Vulnerability Advisory http://j.mp/storm_vuln
[29] Serfon, C., et al. ATLAS DQ2 to Rucio renaming infrastructure. 2014 J. Phys.: Conf. Ser. 513 042008
[30] The Jekyll static site generator http://jekyllrb.com
[31] Markdown https://en.wikipedia.org/wiki/Markdown
[32] Embedding Jetty https://wiki.eclipse.org/Jetty/Tutorial/Embedding_Jetty

Additional Information

Organization

Director

Mauro Morandin	*(till September 30^{th})*
Gaetano Maron	*(since October 1^{st})*

Scientific Advisory Panel

Chairperson (till September 30^{th})	Gaetano Maron	*INFN – Laboratori Nazionali di Legnaro, Italy*
Chairperson (since October 1^{st})	Michael Ernst	*Brookhaven National Laboratory, USA*
	Gian Paolo Carlino	*INFN – Sezione di Napoli, Italy*
	Patrick Fuhrmann	*Deutsches Elektronen-Synchrotron, Germany*
	Josè Hernandez	*Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas, Spain*
	Donatella Lucchesi	*Università di Padova, Italy*
	Vincenzo Vagnoni	*INFN – Sezione di Bologna, Italy*

User Support

Head: C. Grandi

M. Tenti	L. Morganti	S. A. Tupputi	S. Taneja
S. Virgilio	S. Perazzini	P. Franchini	

Tier 1

Head: L. Dell'Agnello

Farming	**Storage**	**Networking**	**Infrastructure**
A. Chierici	P. Ricci	S. Zani	G. Bortolotti
S. Dal Pra	A. Cavalli	A. Barilli	A. Ferraro
M. Donatelli	M. Favaro	L. Chiarelli	A. Mazza
F. Rosso	D. Gregori	D. De Girolamo	M. Onofri
A. Simonetto	B. Martelli	G. Giotta	
	M. Pezzi		
	A. Prosperini		
	V. Sapunenko		
	G. Zizzi		

R&D Service

Head: D. Salomoni

M. Caberletti	V. Ciaschini	F. Giacomini	M. Manzali
M. Perlini	E. Ronchieri		

National ICT Services

Head: R. Veraldi

S. Antonelli	S. Longo

Grid Services

Head: M. C. Vistoli

D. Andreotti	M. Bencivenni	F. Capannini	A. Ceccanti
M. Cecchi	D. Cesini	A. Cristofori	G. Dalla Torre
M. Di Benedetto	E. Fattibene	T. Ferrari	D. Michelotto
G. Misurelli	A. Paolini	V. Venturi	P. Veronesi
E. Vianello			

Hardware and Software Support

Head: G. Vita Finzi

Information System

Head: G. Guizzunti

S. Bovina	S. Cattabriga	E. Capannini	M. Canaparo
C. Galli	C. Simoni		

Administration

Head: M. Pischedda

A. Aiello	G. Grandi	A. Marchesi

Seminars

Feb. 1^{st}, 2013	Francesco Giacomini **The New C++ Standard (The Library)**
Feb. 13^{th}, 2013	Pier Paolo Deminicis **La costruzione di un business plan e le problematiche sulla creazione d'impresa**
Feb. 14^{th}, 2013	Fabrizio Furano **Storage Federations In Ambito HEP**
Feb. 22^{nd}, 2013	Andrea Negri **Hierarchical Data Format 5 (HDF5): Why and How to Use It**
May 10^{th}, 2013	Davide Salomoni **Hype in the Cloud, Stacks in the Ground - Andata e Ritorno per il Cloud Computing**
May 17^{th}, 2013	Stefano Spataro **Computing Design Choices for the PANDA Experiment**
May 22^{nd}, 2013	Charles Loomis **Scientific Cloud Computing - Present and Future**
Jun. 7^{th}, 2013	Michele Pezzi, Paolo Veronesi **Installazione e Configurazione di un Centro di Calcolo con Puppet e Clobber**
Jun. 26^{th}, 2013	Giuseppe Misurelli **Sursum Log: Introduzione a OSSEC e Splunk**
Jul. 17^{th}, 2013	Elisabetta Ronchieri **A Software Quality Predictive Model**
Oct. 1^{st}, 2013	Salvatore Alessandro Tupputi **Automating Usability of ATLAS Distributed Computing Resources**
Oct. 23^{rd}, 2013	Tim Mattson **Predicting the Future in a Rapidly Changing Many-core World**

Oct. 28^{th}, 2013	**Report da CHEP 2013**
Nov. 20^{th}, 2013	Davide Salomoni, Paolo Veronesi **Report da OpenStack Summit 2013**
Nov. 21^{st}, 2013	Rosa Brancaccio **Tomografia computerizzata e software parallelo di ricostruzione**
Dec. 3^{rd} – 4^{th}, 2013	**CUDA/OpenACC Workshop**
Dec. 18^{th}, 2013	Andrea Petrucci, CERN **XDAQ: A Software Framework for Distributed Trigger and Data Acquisition Systems in HEP Experiments**


50°
INFN
CNAF

Authors of the texts

Antonia Ghiselli
Pietro Matteuzzi
Mauro Morandin
Cristina Vistoli

With contributions by

Lorenzo Chiarelli
Francesco Giacomini
Luca Dell'Agnello
Davide Salomoni
Stefano Zani
Umberto Zanotti

50^{th} anniversary event

agenda.cnaf.infn.it/event/CNAF50

Bubble chambers and the *Flying Spot Digitizer*

The acronym CNAF, which stands for Centro Nazionale Analisi Foto- grammi (National Center for Photographic Analysis), may sound quite strange to those who know the Center for its current research activity and who may wonder what photography has to do with networks or distributed computing. The truth is that CNAF had to pass through some radical transformations in order to keep up with the evolution of nuclear and particle physics.

The origin of the name and the reasons why the Center was founded can be understood only if we take a look at what was happening in nuclear physics during the fifties. The invention of bubble chambers by Donald A. Glaser in 1952 literally revolutionized the way nuclear and particle physics research was performed. Compared to nuclear emulsion, the most precise and compact tracking detector for charged particles used in those days, bubble chambers allowed to create 3D spatial images of elementary particle interactions that were impressed on photographic films at a relatively higher rate. As a consequence, the production of data increased remarkably.

Bubble chambers started to be built in many nuclear physics labs and soon grew in number and size. In Italy the interest for such a promising technology became immediately apparent: it was only 1953 when the founding members of the future Bubble Chamber Group, led by Marcello Conversi, started their experimental activity in Pisa. Soon, relatively large quantities of films were produced and then analyzed in a few sites of the National Institute of Nuclear Physics (INFN) where suitable measuring projectors were used to display the magnified images. With the help of these devices, trained technicians were able to locate many points on the tracks. This information used to be converted into digital form and then collected and transferred to obtain the geometrical track parameters and the kinematic reconstruction of the events.

In the late fifties, the urgent need for more precise and faster measurements was finally addressed by an international cooperation effort (including CERN, DESY, Saclay, Brookhaven, Berkeley and several other universities) whose goal was to build an automatic measurement system of the bubble chamber pictures with precision of one micron.

Giampietro Puppi, director of the INFN unit based in Bologna, saw in the local university environment the ideal conditions to set up a new center with the specific mission of providing support for automatic film measurements and for data processing, paving the way to the creation of CNAF by INFN in 1962. The cooperation with the Bologna unit of CNEN (National Committee for Nuclear Energy) was crucial both for providing access to one of the most advanced computers of those times (the IBM 7094), and for hosting the activities of the new center.

Following an idea originally proposed by P.V.C. Hough and B.W. Powell, the newly formed CNAF team designed and built a Flying Spot Digitizer (FSD). The FSD scanned with a luminous spot and explored sequentially every single film coming from the bubble chambers. This small spot of light was projected by a lamp and shifted mechanically backwards and forwards across each image in search for particle tracks. The tracks modulated the light signal which passed through the film before it got detected by photomultipliers. This procedure generated information on track positions and made them available in electronic form (digital coordinates) to a computer.

Since all films contained many tracks but just a few of them were associated with significant

particle interactions, it was necessary to filter the data so that the noise and all irrelevant information could be removed. This operation was implemented through "pre-digitization", a manual procedure which was carried out with film projectors. The completion of the project owed most of its success to the hard work and passion of the first Technical Director Massimo Masetti and his group of about twelve collaborators. The project started in 1962, whereas the entire system FSD-7094 was operational in the early months of 1966 and inaugurated in the presence of the University Minister Luigi Gui the following year.

Many were the original electronic, mechanical and optical components that had to be developed in cooperation with specialized companies. A ferromagnetic thin film memory was built in collaboration with the company Olivetti to achieve read/write access time performance suitable for the data acquisition peaks of the FSD. Thanks to this innovation, it was possible to overcome the limitations of the IBM 7094 standard magnetic memory.

In 1968 CNAF bought an IBM 360/44, a new fast and efficient computer for the on-line acquisition of the data produced by the FSD. CNAF built an original advanced electronic interface to interconnect the new IBM 360/44 to multiple external devices like a new FSD and some new analyzer devices. Some INFN groups based in Bologna, Padua and Trieste started to bring their films to CNAF, and the Center began to play a major coordinating role for precise measurements and for central computing. The FSD was made available to all Italian INFN sites, and CNAF was in charge of coordinating its use by all different groups.

To face the increasing computing demand from INFN sites, a new FSD was made operational in 1971, and it reached the remarkable amount of 300.000 measured events per year. The bubble chamber pictures were becoming more and more complex, and that implied an increase of the noise produced; it was necessary to reduce the amount of data. A new special purpose processor was designed and implemented to run online, and it was able to transform a cloud of points in segments identified by coordinates and slope. This processor, built with the most advanced technologies of the time, was programmed in firmware.

In the meantime (ca. 1968) Proteus, another analyzing system device whose features were similar to the FSD, was built in about one year. The FSD opto-mechanical core device was replaced by a high-resolution Ferranti cathode ray tube, and the scanning spot was obtained by focusing the light produced by an electron beam which hit a fluorescent screen.

Compared to the FSD, Proteus was less precise and had a lower resolution, but it was more versatile because the electron beam could be deflected as required by applying an electromagnetic field. Proteus allowed to select the area to scan and so it decreased the computational effort necessary for the noise reduction. This device was particularly efficient in the analysis of spark chamber events whose topology was simpler than bubble chambers.

The measuring process was sped up to such an extent that it was possible to analyze up to 240 pictures per hour, i.e 100.000 events in 45 days. As a clear indication of both the versatility of the device and of the openness of the Center to cooperate with other research institutes, it is worth mentioning a paper produced in partnership with the Department of Human Anatomy of the University of Modena. Proteus was used for scanning photographic images of human chromosomes in order to study their banding patterns.

The last development in the study of bubble chamber events was ERA- SME, a CRT device far more complex and performing than Proteus. The project started in 1974, and was intended for the analysis of films produced by CERN with the big bubble chamber called BEBC. Based on the measurement system built in Geneva, ERASME followed the steps of pre-digitization and of automatic measurement. This effort was made possible by a new advanced graphic and interactive system developed at CNAF which allowed to overlap analog and digital information. The expertise and know-how gathered by CNAF physicists and engineers in inventing and managing digitizing devices was quite peculiar. Beside the obvious competences in mechanics and optics, great efforts were devoted to programming and optimizing data carrying connections.

These are the words used by Albert Werbrouk, first Scientific Director of the Center, to recollect those achievements:

> "Some programs bore names like Mist, Fog or Haze, names which summon the climate of the San Francisco Bay. The Hough-Powell device was initially developed at Berkeley, and we kept those names as a tribute. Other programs whose names were evocative of their functions, such as Thresh and Grind, came from CERN. We had to write some original code to let different tools interact, as in the case of BIND, which converted data from the card-reader to a magnetic tape"

The names used for those programs witness the heterogeneity of their origins. This heterogeneity forced CNAF computing professionals to get a deep and well-grounded knowledge of different software, ranging from programs developed by researchers to the managing systems of commercial machines such as IBM computers. People at CNAF had to learn to work efficiently and to adapt quickly to different platforms. At the same time, it was necessary to cautiously manage the flux of data from FSDs to computers and from computers to recorders or printers. A few years later, the sound expertise in data transmission and storage acquired by CNAF technicians in those years would be of great help when bubble chambers had to be replaced by more effective electronic instruments and the Center was to reconsider its mission and role.

The birth of INFNet

Bubble chambers gave origin to a trend which shaped the future of the experimental research in nuclear and particle physics because they produced massive quantities of data which could not be processed manually, and proved that automatic computing machines were essential tools for physicists.

Gradually, electronic computers were introduced at every level of the experimental procedure. The first commercially successful minicomputer manufactured by DEC appeared on the market in 1965, and this new line of machines became soon popular in HEP labs where it was exploited for relatively simple and specific tasks such as instrumentation control and data acquisition. On the other hand, bigger and more powerful machines were used to perform more demanding computational jobs like geometrical and kinematic reconstruction of events, data storage, electronic simulation of phenomena, and the statistical analysis of the experimental results that had been collected.

As a consequence, in the late sixties CNAF and some INFN units began to acquire computers of different size and the equipment necessary to access mainframes. The commitment of INFN in using computational devices for research purposes also promoted the creation of important regional academic computing centers like CINECA, CILEA, CSATA, etc.

The earliest data communication facilities in INFN sites were terminal clusters for remote job entry (RJE) which were connected to mainframes such as IBM/370, CDC/6600 and UNIVAC. Due to the lack of interconnections, INFN sites were compelled to set up independent connections to any computing center they were interested in. The first attempt to overcome this situation was made in 1979 when an experimental dial-up connection was established between some PDP-11 minicomputers located in different INFN sites (Milan, Pavia, Rome, Frascati).

In the late seventies, DEC presented its first release of a data communication package (DECnet) meant to interconnect its minicomputers (PDP) and, at about the same time, it introduced the novel concept of RJE station software emulation, i.e. the possibility to use the minicomputers for remote job submission to mainframes. The original idea of INFN was to integrate the different emulation packages with DECnet and to develop dedicated gateways which were directly connected to each mainframe. CNAF, that at that time was facing the slow phase out of bubble chambers in favor of more modern elementary particle detectors equipped with electronics read-out, quickly became the reference site within INFN for management, experimentation and planning of the new network developments.

The initial results were so good that INFN set up a digital network based on a leased lines star topology built around CNAF and on the gateways located in different sites close to the mainframes: the setting up of this simple infrastructure marked the dawn of the INFN network, INFnet. CNAF played an important role by implementing the gateway to the CINECAs CDC mainframe with batch and interactive functions. The advantages of such an innovation were considerable: INFN physicists were able not only to submit jobs to the mainframes, but also to communicate remotely by e-mail, to exchange files, and to access remote computers and data. Users were hence enabled to work easily in an heterogeneous environment and through a simple and unique user interface.

In those years physicists also started to set up international connections. In 1983 the first leased line CNAF-CERN was installed and connected on an INFN computer, named CERNGW. It was a gateway node because the CERN machines were connected through a CERN-developed LAN protocol named CERNet. In the meanwhile, new requirements about data communication were coming from the Large Electron-Positron Collider (LEP) which was under construction at CERN; in this respect, CNAF developed, in partnership with INFN-Bologna, a gateway to access the CERNs IBM in interactive mode from the DECnet environment. 1985 saw the installation of the first leased line CNAF-Fermilab, and the protocol adopted was DECnet on both sides of the Atlantic. Later on, also the DESY laboratory in Hamburg was connected to CNAF. In the early months of 1989 the INFNet infrastructure was based on 64Kbit/sec leased lines. The HEP community realized the importance of supporting a worldwide dedicated network infrastructure (HEPnet) for High Energy Physics, and CNAF participated actively in the coordination effort carried out within HEPCCC, the HEP Computing Coordination Committee.

Connecting Italian research networks: GARR

In the eighties, other Italian research Institutions like CNR and ENEA had also started to organize and connect their computational resources. CINECA, which ran the biggest computing center for the Italian research community, was working on a project for a national network enabling its users to take advantage of its powerful computing facility. Other academic computing centers, such as CILEA in Lombardy and CSATA in Apulia, decided to set up some networks on a regional scale too. All these networks needed to interact and to communicate, and so the main Italian research institutions decided to found a purpose-specific study group, i.e. GARR (Gruppo per lArmonizzazione delle Reti della Ricerca — Group for Research Networks Harmonization).

In 1988 GARR started to plan the creation of a national 2Mbps backbone dedicated to research institutes and universities and funded by the University and Research Ministry. The network scenario in Italy was very heterogeneous, and several protocols were in use. The most reliable were well-tested proprietary protocols (such as SNA for IBM and DECnet for DEC) and protocols based on open international standard models (like the OSI X25 protocol adopted by the Italian Postal, Telegraph and Telephone services - PTT). The Internet TCP/IP protocol, albeit considered not sufficiently mature, was spreading rapidly, and 1989 saw the birth of the organization that allocated IP addresses in Europe (RIPE - Réseaux IP Européens). In general, users were reluctant to abandon the dependable protocols they knew so well, and therefore the new 2Mbps backbone (the first implementation of the GARR network) was based on Time Division Multiplexing (TDM), a technology employed to share a communication channel among several protocols by allocating a short time access to each of them. CNAF, on the basis of its experience, was the coordinating partner in the definition of the plan and the implementation of the project. The first 2Mbps line was set up between CNAF and CERN in 1989, and that was a remarkable achievement, because this connection was the very first 2Mbps link which CERN had established with other research institutions. In a short time, the 2Mbps backbone was extended with links to CINECA, and to the Rome and Milan physics departments.

Mailing was, and still is, one of the most important network services, and in a short time it became necessary to find a way to communicate by e-mail with users of different networks. In 1989 CNAF, in collaboration with INFN-Trieste, developed the mail gateway called GIVEME (General Interface on VMS for Electronic Mail Exchange) that represented an original and quite innovative development. GIVEME was able to interwork between DECnet, EARN/BITNET, X25, and the Internet. In 1993 a second gateway based on a distributed architecture was implemented.

The availability of 2Mbps links dramatically improved the GARR network potential, and users started to shift to innovative and demanding services, such as videoconferences; in particular, an experimental backbone for IP multicast traffic across the Internet known as Mbone became quite popular. In this respect, TDM technology, with its static allocation of bandwidths, was no longer an attractive option for efficient network links sharing. The need to change transmission technologies and the topology of infrastructures became therefore compelling. Those years saw the start of an experimental phase intended to test new 'transfer

modes' technologies which allowed more flexibility and efficient bandwidth usage. The two new transfer mode technologies in the limelight were Frame Relay and Cell Relay.

On the one hand, Frame Relay was data oriented and targeted at speeds up to 2Mbps, and it was offered by Telecom Italia through a service called C-LAN. The most interesting feature of C-LAN was the possibility to define several virtual circuits for building meshed topologies, which allowed to increase the aggregated throughput of the backbone. On the other hand, Cell Relay was targeted at higher speeds, and therefore designed for transmitting data, voice and video. CNAF, in collaboration with INFN-Bologna, tested a new device called IPX, which used a proprietary form of cell switching called FastPacket. On the basis of their good results, these devices were used to set up the first backbone based on cell switching.

Both Frame Relay and Cell Relay proved to be performing adequately, and in 1995 they were used as a keystone in the realization of the GARR-2 backbone, the first upgrade of the original GARR infrastructure. Shortly afterwards, Telecom Italia made available other fast links at 34 Mbps, and, after long and careful testing on high speed LAN and WAN, contributed to the evolution of the Cell Relay technology with the Asynchronous Transfer Mode (ATM).

Tests disclosed both qualities and limitations of the project, and a lot of tuning was necessary to reach a satisfactory efficiency. Final results were however appreciable, and the new pilot product named GARR-B saw the light based on ATM 34Mbps links. Due to the positive outcomes of this and similar experiences elsewhere, the same technology was also employed in the new European network TEN-34. Such a successful experimentation had been possible thanks to the wide cooperation that had taken place among some INFN sites coordinated by CNAF. In the meanwhile, the data transmission protocols scenario was changing. Although until that moment the development of data communication networks had been conducted through the implementation of proprietary protocols, the new tendency was to pursue a more selective approach; i.e. public data networks such as DATAPAC in Canada, TRANSPAC in France, ITAPAC in Italy, and Coloured Book protocols for academic network in the United Kingdom were all based on a single standard protocol X25. In 1996 the GARR group had to take a crucial decision. DECnet, up to that moment the most popular architecture, was running out of addresses, and therefore DEC was about to launch the new version, DECnet phase V, which was based on OSI international standards and boasted a new 160bit addressing system. Meanwhile, the alternative TCP/IP protocol was gaining worldwide consent as an open architecture supported by the growing diffusion of Unix and Linux operating systems. CNAF studied and tested DECnet phase V within the HEPnet communities, but the consolidation of this new version took far too long, and therefore GARR-B, the new GARR network released in 1998 and based on 34Mbit backbone links, was configured to support only the TCP/IP protocol. GARR-B was the last network infrastructure produced by CNAF, since GARR evolved later into an independent entity located in Rome and took over the management of the Italian research network. Since 1997, CNAF was getting interested in the rapid evolution of the HEP computing models, and, after the delivery of GARR-B, it gradually reoriented its research activity towards this field. In addition, the Center had acquired throughout the years a relevant role as a reference center for the development and the coordination of INFN activities in other ICT sectors. In many cases, the initial interest to explore emerging technologies eventually resulted in the organization of new national services being deployed for the benefit of all INFN sites. This happened for instance in the videoconference sector, where several test campaigns were carried out using various network protocols, or in the case of the early adoption of web technologies for document management, or in the development of a national software and hardware maintenance and license management system.

The computing infrastructure of LHC

In 1998, while CNAF started shifting its research mission, INFN was forced to rethink its policy regarding the management of its computing resources. Until that moment, the bulk of computing resources owned by INFN had been limited to small computational farms for local use, while mainframes owned by High Energy Physics Laboratories worldwide and other Italian centers such as CINECA, CILEA, CASPUR had been employed by the Institute for bulk computation. However, the decision of participating in the Large Hadron Collider (LHC) experiments in Geneva was going to change this situation completely. It was soon realized that, with its expected 15 PB of data generated each year, LHC would have represented a completely different challenge in terms of data analysis management.

As Laura Perini, a member of the ATLAS experiment, recalls:

> "We realized that maintaining all the computational facilities in Geneva as had occurred for LEP was no longer possible. Contributors definitely preferred having direct control on the infrastructures they paid for, and they would have liked them to be in their own country. However, thanks to data transmission connections, the situation was different for computing centers. Our idea was to allow every single country involved in the LHC project to develop its own human and computing resources in loco, in order to spread skills and expertise all over Europe."

A research project named Monarc (Models of Networked Analysis at Regional Centers) was launched. CNAF was involved in the project from the beginning and gave a significant contribution to it. The resulting proposal was a hierarchical architecture with computing resources to be organized in three levels (TIERs). The data produced by the LHC accelerator from CERN (TIER-0) would have been transferred to the main centers (TIER-1), located in some European countries, in the US and in Asia, and responsible for data management and processing at a regional scale. Secondary centers (TIER-2) would have referred to TIER-1 centers in order to get the data they were interested in, with no direct connection to TIER-0.

This simple and rigid structure was justified by a precautionary approach, since at that time networks were considered a possible limiting factor for the expected huge amount of data to be managed on a worldwide scale.

The introduction of a new, geographically distributed approach was ac companied by other important changes: the object-oriented programming paradigm was gaining consensus within the big HEP collaborations, and the combination of commodity servers employing the Personal Computer architecture and the open Linux operating system gradually became the platform of choice for building up large scientific clusters.

INFN therefore started a big campaign with the aim of training physicists and computer scientists to deal with new computing technologies. CNAF, which took part in the newly established INFN CNTC committee (Comitato Nuove Tecnologie di Calcolo — New Computing Technologies Committee), decided to organize courses to teach programming languages like C++, Unix-based operating systems and the TCP-IP protocol suite. Meanwhile, the INFN computing infrastructure for LHC started to take form, and CNAF was soon identified as the

hosting site for the Italian TIER-1 because of its expertise in running networks infrastructures, in providing computing national services and for the experience gained in managing the INFNet and GARR projects. In 2001 the project for the INFN TIER-1 at CNAF entered its operational stage with the design of a prototype installation. Unlike other sites across Europe, CNAF could not count on pre-existing hardware, software and technical infrastructures suitable for a TIER-1 center.

Circumstances were challenging because a large and performant computing center had to be to set up from scratch in a relatively short time. The location for the facility was sought within the premises of the Physics Department of the Bologna University, which was already hosting the CNAF center. This choice allowed a constant and close interaction between the two communities, but it inevitably posed a few logistical issues, since no suitable place was immediately available. In the end, a 1000 sqm garage floor in the basement of the main building was identified for the purpose, and the design of the computing rooms had to face several environmental constraints. At the end of 2005, the TIER-1 prototype was fully functioning, and its computing facilities consisted of about 1000 bi-processor servers mounted on 35 racks, 450 TB of disk storage and a magnetic tape library of 500 TB, managed by fully automated software procedures for installation and configuration. The local area network consisted of 35 switches to aggregate the servers, and two core switches connecting the center to the GARR network at 2Gbps.

The experience obtained with the prototype was very useful for training the personnel and for identifying the most critical aspects to be taken into account for delivering the services with the very high (99,9%) availability and reliability levels required by the LHC experiments. It was realized that in order to achieve such goals, complete redundancy of the electrical and HVAC systems was necessary to avoid single points of failures and perform maintenance operations with no impact on the running systems. The storage system was identified as the most critical computing service: the first solution adopted at CNAF could hardly comply with the LHC requirements. In this respect, several campaigns of tests were launched to investigate alternative solutions. The results of these tests led to the adoption of a solution based on industrial standards: a parallel file-system meant to ensure high availability and performances, and an infrastructure based on Storage Area Network to allow flexibility and easiness of management. The batch system responsible for managing the delivery of computing jobs to the servers was also replaced by a more mature and scalable solution, and, finally, the unification of all computers into a single cluster made it possible to avoid any static partitioning.

In 2005, a substantial infrastructure upgrade including building refurbishment, new racks arrangement for the hot aisle containment and a complete electrical and HVAC systems redesign was planned, and a series of European tenders was announced. Improvement activities started in 2008 and were completed after only nine months, as in March 2009 the Center attained its expected computing, storage and technical configuration just in time for the start of data acquisition at LHC. With 3.8 MW of available electrical power (3.5 MW in absolute continuity), and six independent cooling plants capable of extracting up to 1.5 MW of released heat, the computing room upgrade was designed to exceed the foreseeable needs of the LHC experiments. At the beginning of 2013 the Italian Tier1 was hosting more than 1300 servers with 13,000 computing cores, 13 PByte of disk storage and a magnetic tape library with 14 PByte of data. Network connections to the external world were provided by GARR through three separate 10 Gbps links. About 20 international scientific experiments were supported at the Center; on average, 90,000 batch jobs were being executed every day, with computing resources optimally used 24x7 by all experiments thanks to sophisticated sharing policies applied to a single large computing cluster capable of supporting local, Grid and Cloud resource requests on both real and virtual environments.

The computing infrastructure of Grid

In the late nineties, the idea of taking advantage of clusters of personal computers and workstations to gather the computational power needed by large scientific endeavors was gaining wider acceptance, and the availability of high-performance connections encouraged different groups all over the world to embrace this approach. Until that moment, these clusters were employed only on a local scale, and the new idea that CNAF contributed to investigate was to extend this kind of computing infrastructures on a geographical scale.

CNAF started exploring this possibility, and in 1999 it began an in-depth evaluation of two systems that had been created in the United States: Globus, developed at the Argonne National Laboratory in Chicago, and Condor, developed by the University of Winsconsin Madison. The Globus toolkit provided some basic tools to export diverse computing and data resources through a common interface. Combined with a directory service and a strong security model, the tools enabled the creation of a distributed computing system on a geographical scale that became known later as the "Grid". Originally Condor was used only for relatively small clusters of computers residing within the University district, but CNAF, in collaboration with several INFN sites, extended its use on a national scale. Thanks to this effort, researchers acquired full awareness of the great potential of the recently produced tools, but also realized that those tools were not yet mature enough for the production mode deployment on a geographical scale.

More development was needed, and in 2000 INFN launched the INFN-GRID project, which included several computing facilities distributed in seven INFN sites. The goal of the project was to implement a dependable and consistent Grid model (with uniform interfaces to a wide variety of resources) which could enable a pervasive access to all available resources. In 2001, under the coordination of CERN and in partnership with other institutes from most European countries, INFN-GRID promoted the first European Grid Project: DataGrid. DataGrid gave INFN-GRID an international dimension and, after three years of tumultuous development, succeeded in delivering a European e-infrastructure for science.

From that moment on, Grid projects proliferated, with CNAF heavily involved in many of them. A contemporary project to DataGrid was DataTAG, whose purpose was to address the issues raised by the interaction with the Grid infrastructures deployed in the US. In particular, this project focused its activities on two fundamental aspects of the interoperability of different Grid systems, i.e. the necessity of a common authentication and authorization model, and the set-up of a common schema to describe Grid resources.

In 2002 the LCG (LHC Computing Grid) project was launched in order to manage the growing LHC computing infrastructure, and in 2006 it was renamed WLCG (Worldwide LHC Computing Grid) to stress the worldwide distribution of the collaboration. The DataGrid project was followed by the "Enabling Grids for E-sciencE" initiative (EGEE): a sequence of three projects funded by the European Commission which took over both the operation of the infrastructure and the evolution of many middleware products originally created by DataGrid, as required by the whole European Research community.

EGEE was built on the EU research network GÉANT, and exploited the Grid expertise brought forth by many national and international projects. After the end of EGEE, the

management of the deployed e-infrastructure and the maintenance of the middleware took different routes, with the former being embedded in the European Grid Infrastructure (EGI) project and the latter absorbed in the European Middleware Initiative (EMI) project. CNAF has played a prominent role in managing and developing several components of the Grid services within the above mentioned European projects, the most notable ones being:

- The Workload Management System (WMS), aimed at distributing the computational workload on the Grid in an effective and efficient way based on the current state of computing resources and on the data availability.
- The Virtual Organization Membership Service (VOMS), a Grid attribute authority which serves as a central repository for user authorization in formation, providing support for sorting users into group hierarchies and keeping track of their roles and other attributes within a user community.
- The Storage Resource Manager (StoRM), and the Grid Enabled Mass Storage System (GEMMS), that integrates disk and tape storage in a seamless way and exposes the resources through a Grid interface.
- The definition of the Resource schema, i.e. a conceptual model of the Grid resources to be used in the Grid Information Service for discovery and monitoring purposes.
- The Worker Nodes on-Demand Service (WNoDeS), that provides virtualized and customized computing resources on demand, building upon a tight integration with the batch system managing a computing cluster.

Besides these products, CNAF developed a variety of tools for controlling, supervising and monitoring the Grid computing infrastructure, and for managing the Grid software release distribution.