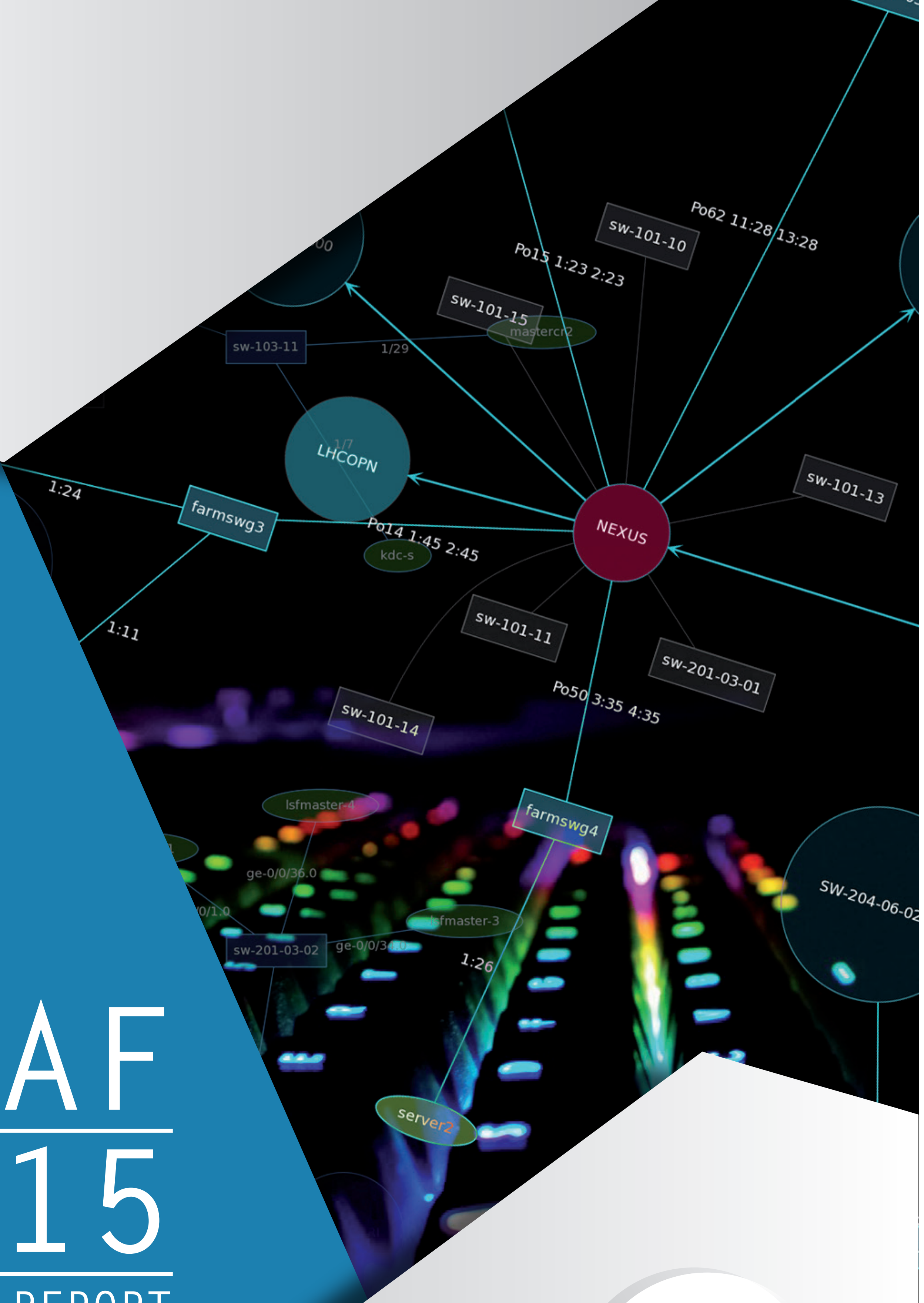


CNAF 2015

ANNUAL REPORT

INFN-CNAF Annual Report 2015

www.cnaf.infn.it/annual-report
ISSN 2283-5490 (online)

Editors

Luca dell'Agnello
Francesco Giacomini
Matteo Tenti

Cover Design

Francesca Cuicchio

Address

INFN CNAF
Viale Berti Pichat, 6/2
I-40127 Bologna
Tel. +39 051 209 5475, Fax +39 051 209 5477
www.cnaf.infn.it

Contents

Scientific Exploitation of CNAF ICT Resources

The INFN-Tier1 Center and National ICT Services

Software Services and Distributed Systems

Additional Information

Introduction

2015 has been a year full of changes and developments for our center. Among the most notable: the approval of the new organizational structure of the center, which, together with the support to scientific computing, reinforces the activity on research, development and technology transfer; the introduction of the elastic expansion of the data center, through the use of remote computing resources, be they offered on an opportunistic basis, provided by a data center with temporary excess capacity or rented from a cloud provider; the start of the first important H2020 projects; the accreditation of our center by Regione Emilia-Romagna as an industrial research lab, which gives us the opportunity to interact more efficiently with the local productive system.

The Tier1 Data Center ran smoothly during the year, serving appropriately all the four LHC experiments engaged in the second LHC physics Run, at a centre-of-mass energy of 13 TeV, that started in late Spring 2015 and ended in December, including 3 weeks dedicated to heavy-ion collisions. Beside this activity, aimed to support the World Wide LHC Computing Grid (WLCG) e-infrastructure, CNAF is committed to provide computing resources to astroparticle experiments in which INFN is involved. Currently more than 30 Astroparticle experiments use the computing resources hosted at the Tier1 Data Center and recently new experiments, such as PADME, DAMPE, FAMU, LSPE and CUPID, have joined the group. Research groups active in the Digital Cultural Heritage domains represented by the INFN CHNet network and researchers belonging to the Computational Chemistry and Biomedical domains also access the CNAF facility through Grid or Cloud services.

All experimental groups working at CNAF are assisted by the User Support team, which is skilled both in proposing computing models for experiments and in the management of daily operations by offering to the users a well defined interface to the various groups at CNAF, from the data center operations to the groups managing the Grid middleware.

In 2015 the computing power of the Tier1 reached 188000 HS06, while the fast storage exceeded the capacity of 18 PB and the long-term storage the capacity of 20 PB. During the year significant improvements to the overall service infrastructure have been made, providing a well established configuration and management tool, strengthening the virtualization of all the main services, updating the monitoring and alarming infrastructure and finally upgrading the data center batch system. Investigations and tests continued for low-power processors. The need to serve peaks of requests of computing power has lead to the successful development of an elastic data center extension towards resources located at external providers, both public and private.

The Software Development and Distributed Systems (SDDS) group has been heavily engaged on the startup of the INDIGO – DataCloud (EINFRA-1-2014) project, an H2020 project on the development of a data/computing platform targeting scientific communities and provisioned over hybrid (private and public) e-infrastructures. CNAF-INFN is leading the consortium coordination and the project management of this 30-month project that sees the collaboration of 26 European partners in 11 countries, including developers of distributed software, industrial partners, research institutes, universities, e-infrastructures. Work is in progress towards the first software release foreseen for mid-2016. The Software Development group, apart from the contribution to the INDIGO project, has focused on the development of a large-scale prototype for the event building component of the LHCb experiment and of the data acquisition system of the KM3NeT experiment. The group is also involved in the management of the Grid-WLCG infrastructure and on the maintenance of some Grid services, including VOMS, StoRM/SRM and Argus.

Within the European H2020 framework program, CNAF has participated to the 2014 and 2015 calls related to the areas of development more suited for our center, especially on distributed systems (Grid and Cloud), on infrastructure for our communities, but also on the development of new IT technology. In this context we won five projects, including the above mentioned INDIGO, and all of them started the activities along the year. All these projects provide significant resources, in particular human resources, to CNAF, thereby strengthening its development activities for the coming years.

Gaetano Maron
CNAF Director

Scientific Exploitation of CNAF ICT Resources

The User Support unit at CNAF

E-mail: exp-support-cnaf@lists.infn.it

Abstract. Many different research groups, typically organized in Virtual Organizations (VOs), exploit the Tier-1 facilities for computing and/or data storage and management. The User Support unit provides them with a direct operational support, and promotes common technologies and best-practices to access the ICT resources, in order to facilitate the usage of the center and maximize its efficiency.

1. Current status

Born in April 2012, the User Support team is presently composed by one coordinator and six fellows (Assegno di Ricerca) with post-doctoral education or equivalent work experience in scientific research or computing.

The main activities of the team include:

- providing a prompt feedback to VO-specific tickets on the VOs ticketing system, or via mailing lists or personal emails from users;
- forwarding to the appropriate Tier-1 units those requests which cannot be autonomously satisfied, and taking care of answers and fixes, e.g. via the tracker JIRA, until a solution is delivered to the experiments;
- supporting the experiments in the definition and debugging of computing models in distributed and Cloud environments;
- helping the supported experiments by developing code or monitoring frameworks;
- porting applications to new parallel architectures (e.g. GPUs);
- providing the Tier-1 Run Coordinator, who represents CNAF at the Daily WLCG calls, and reports about resource usage and problems at the monthly meeting of the Tier-1 management body (Comitato di Gestione del Tier-1).

People belonging to the User Support team represent CNAF Tier-1 inside the VOs. In some cases, they are directly integrated in the supported experiments. In all cases, they can play the role of a member of any VO for debugging purposes.

The User Support staff is also involved in different CNAF internal projects, in particular during 2015, collaborated to the migration of the Tier-1 resource provisioning system to a new system based on Puppet and Foreman. It also stated the collaboration with other CNAF groups to renew the data center monitoring system including a complete refactoring of the Tier-1 experiments dashboard; this work will be completed in 2016.

Members of the User Support group also participated to the activities of the Computing on SoC architectures (COSA) project (www.cosa-project.it). The project is dedicated to the technology tracking and benchmarking of the modern low power architectures for computing applications.

2. Supported experiments

Besides the four LHC experiments (ALICE, ATLAS, CMS, LHCb), for which CNAF acts as a Tier-1 site, the User Support also takes care (or has taken care) of the direct day-by-day operational support of the following experiments from the Astrophysics, Astroparticle physics and High Energy Physics domains: Agata, AMS-02, Argo-YBJ, Auger, BaBar, Belle II, Borexino, CDF, CTA, Cuore, DarkSide-50, Gerda, Glast, Icarus, LHAASO, Juno, Kloe, KM3NET, LHCf, Magic, NA62, Opera, Pamela, Panda, Virgo, and Xenon100.

Recently, the tier1 has started supporting new experiments such as PADME, DAMPE, FAMU, LSPE, CUPID and research groups active in the Digital Cultural Heritage domains represented by the INFN CHNet network.

Researchers belonging to the Computational Chemistry and Biomedical domains also access the center through Grid services.

ALICE at the INFN CNAF Tier-1

Domenico Elia[1]
[1]INFN Bari, Bari, Italy
E-mail: `domenico.elia@cern.ch`

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 configuration of the experimental apparatus has been upgraded for Run2 by installing a second arm complementing the EMCAL at the opposite azimuth and thus enhancing the jet and di-jet physics. This extension, named DCAL for "Dijet Calorimeter" has been installed during the Long Shutdown 1 (LS1) period of LHC from the April 2013 to the end of 2014. Other detectors were also upgraded or completed: in particular the last few modules of TRD and PHOS were also installed while the TPC was refilled with a different gas mixture and equipped with a new redesigned readout electronics. Also the DAQ and HLT computing farms were upgraded to match the increased data rate foreseen in Run2 from the TPC and the TRD. 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 heavier 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. Data taking, physics results and plans for the upgrade

During the year 2015 the re-start of the physics program at the LHC has taken place, after the upgrade of the machine and the consolidation of the experiments during LS1. The ALICE detector restarted data taking operations in January 2015 with a full cosmic data taking campaign. During the ensuing run with protons at 13 TeV ALICE operated very smoothly, adjusting the choice of triggers to the evolving running conditions. In the following intensity ramp up phase with 50 and 25 ns bunch spacing, ALICE has been operating at instantaneous luminosities up to 5 Hz/µb collecting 620 M minimum bias events and integrating 4.35 pb^{-1} di-muon triggers and 1.81 pb^{-1} high-multiplicity triggers in proton-proton collisions. The ALICE luminosity targets for 2015 have been reached. The heavy-ion (HI) run for 2015 was characterized by a very packed schedule as it included one week (3 days of setup and 4 of data taking) of p-p reference running at 2.51 TeV/beam, one lead source refill, one machine development shift for crystal collimation, the ALICE polarity flip which requires loss maps revalidation, VdM scans (also for the reference run) and two quench tests (luminosity and collimation induced, both successful).

ALICE operated successfully during the proton-proton reference run, collecting 130 M minimum bias events which amounts to 13% of the request formulated for the entire Run2 (1000M). During the 3 weeks of HI running the LHC has delivered high-luminosity Pb-Pb collisions and ALICE has achieved the historical milestone of running at its design luminosity of 1000 Hz/b for an average of 2.4 h at the beginning of each high-intensity fill. The overall delivered luminosity was 433 μb^{-1}. The target for the number of minimum bias triggers and rare triggers (muon, ultraperipheral and calorimeter) have been matched or are very close of the desired statistics. The average running efficiency during the 4 weeks (proton reference and lead run) was around 90%. A further improvement for the 2016 run will be obtained deploying tools to minimize the time between the start of consecutive runs. Many new physics results have been obtained from pp, p-Pb and Pb-Pb collisions. In particular no cold nuclear matter effects have been measured in p-Pb collisions, while several signals of collective effects have been unexpectedly observed in collisions of smaller systems (pp and p-Pb) triggering a considerable interest in the theorists [4, 5]. There have been 43 publications and more than 500 conference presentations.

For the upgrade program of Run3, the five TDRs, namely for the new ITS, the TPC GEM-based readout chambers, the Muon Forward Tracker, the Trigger and Readout system, and the Online/Offline were fully approved by the CERN Research Board. A transition from the R&D phase to the construction of prototypes of the final detector elements is progressively taking place.

3. Computing model and R&D activity in Italy

The ALICE computing model is still 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 [6]. 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 event reconstruction tasks 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.

The Computing Model has not changed significantly for Run2, except for scavenging of some extra computing power by opportunistically use the HLT farm when not needed for data taking. The much higher data rate forseen for Run3, tough, will require a major rethinking of it in all its components, from the software framework to the alorithms to the distributed computing infrastructure. The design of the new computing framework for Run3, started in 2013 and mainly based on the concepts of Online-Offline integration (“O2 Projec”), has been finalized with the corresponding Technical Design Report [7] approved by the LHCC in September 2015. An addendum to the MoU on the “Ugrade of the Online-Offline Computing System (O2)” has been circulated to the funding agencies for signatures and final commitment of the participating institutes. Development and implementation phases as well as performance tests are currently ongoing.

The Italian share to the ALICE distributed computing effort (currently about 15%) includes resources both form the Tier-1 at CNAF and from the Tier-2s in Bari, Catania, Torino and Padova-LNL, plus some extra resources in Cagliari and Trieste. The contribution from the Italian community to the ALICE computing in 2015 has been mainly spread over the usual items, such as the development and maintenance of the (AliRoot) software framework, the management of the computing infrastructure (Tier-1 and Tier-2 sites) and the participation in the Grid operations of the experiment. In addition, the R&D activities connected with the development of the Virtual Analysis Facility (VAF) in the framework of the STOA-LHC national project (PRIN 2013) have been finalized. Starting from the experience with the Torino VAF which is active already since few years, similar infrastructures have been deployed in Bari, Cagliari, Padova-LNL and Trieste. They have been also provided with an XRootD-based storage Data Federation (DF), with a national redirector in Bari and local redirectors in each of the involved centers: a complete performance study campain has been carried out to compare the data access from the DF with those from the local storage and from the central ALICE catalogue (Alien) at CERN. Results on the VAF/DF developments have been presented at CHEP 2015 [8, 9].

Still on the R&D side in Italy, the design and development of a site dashboard project already started in 2014 has been continued and enforced during 2015. In its original idea, the project aimed at building a monitoring system able to gather information from all the available sources to improve the management of a Tier-2 datacenter. A centralized site dashboard based on specific tools selected to meet tight technical requirements, like the capability to manage a huge amount of data in a fast way and through an interactive and customizable Graphical User Interface, has been developed. Its current version, running in the Bari Tier-2 site since more than one year, relies on an open source time-series database (InfluxDB), a dashboard builder for visualizing time-series metrics (Grafana) and dedicated code written to implement the gathering phase. The Bari dashboard will be exported in all the other sites in the first months of 2016, in order to allow a next step where a unique centralized dashboard for the ALICE computing in Italy will be implemented. The project prospects also include the design of a more general monitoring system for distributed datacenters able to provide active support to site administrators in detecting critical events as well as to improve problem solving and debugging procedures. A contribution on the Italian dashboard has been recently presented at CHEP 2016 [10].

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 cloud-like computing model, it has no special assigned task or reference community, but provides computing and storage resources to the whole collaboration, along with offering valuable support staff for the experiments computing activities. It provides reliable XRootD access both to its disk storage and to the tape infrastructure, through a TSM plugin that was developed by CNAF staff specifically for ALICE use.

Running at CNAF in 2015 has been remarkably stable: for example, both the disk and tape storage availabilities have been better than 99%, ranking CNAF in the top 5 most reliable sites for ALICE. The computing resources provided for ALICE at the CNAF Tier-1 centre were fully used along the year, matching and often exceeding the pledged amounts due to access to resources unused by other collaborations. Overall, about 71% of the ALICE computing activity was Montecarlo simulation, 9% raw data processing (which takes place at the Tier-0 and Tier-1 centres only) and 20% analysis activities.

In order to optimize the use of resources and enhance the “CPU efficiency” (the ratio of CPU to Wall Clock times), an effort was started in 2011 to move the analysis tasks from user-submitted “chaotic” jobs to organized, centrally managed “analysis trains”. The effort went on in the years 2013-2015 with relative year-on-year increases of the number of train jobs of about 50-60%. This leads up to a split of analysis activities, in terms of CPU hours, between about 30% individual jobs and 70% organized trains (6% and 14% of the total ALICE computing activity, respectively).

In 2015, CNAF provided about 6% of the total CPU hours used by ALICE, thus ranking second of the ALICE Tier-1 sites, following only FZK in Karlsruhe. This amounts to about 33% of the total INFN contribution: it successfully completed nearly 6.5 million jobs, for a total of more than 20 millions CPU hours. Figures 1 and 2 show the running job profile at CNAF in 2015 and the cumulated fraction of CPU hours along the whole year for each of the ALICE Tier-1 sites, respectively.

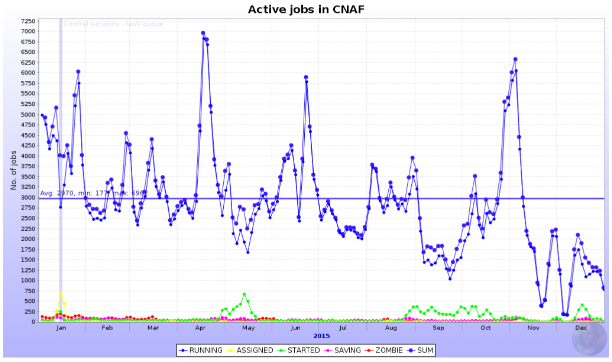


Figure 1. Running jobs profile at CNAF in 2015.

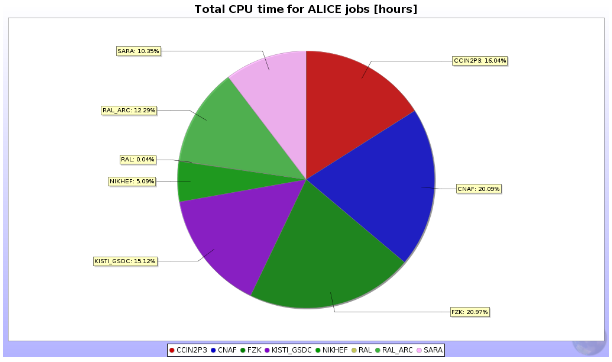


Figure 2. Ranking of CNAF among ALICE Tier-1 centres in 2015.

At the end of the last year ALICE was keeping on disk at CNAF about 2 PB of data in nearly 40 million files, plus about 4 PB 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; the 40Gb/s of the WAN links makes CNAF one of the better-connected sites in the ALICE Computing Grid.

References

[1] B. Abelev et al. (ALICE Collaboration), Int. J. Mod. Phys. A 29 1430044 (2014).

[2] B. Abelev et al. (ALICE Collaboration), Eur. Phys. J. C 74 3054 (2014).

[3] B. Abelev et al. (ALICE Collaboration), Physics Letters B 728 25-38 (2014).

[4] B. Abelev et al. (ALICE Collaboration), Physics Letters B 719 29-41 (2013).

[5] J. Adam et al. (ALICE Collaboration), Physics Letters B 758 389-401 (2016).

[6] P. Cortese et al. (ALICE Collaboration), CERN-LHCC-2005-018 (2005).

[7] J. Adam et al. (ALICE Collaboration), CERN-LHCC-2015-006 (2015).

[8] S. Piano et al., Journal of Physics: Conference Series 664 (2015) 022033

[9] D. Elia et al., Journal of Physics: Conference Series 664 (2015) 042013

[10] G. Vino et al., “A Dashboard for the Italian Computing in ALICE, contribution to the 22nd International Conference on Computing in High Energy and Nuclear Physics (CHEP2016), San Francisco (California, US), October 10-14, 2016.

AMS-02 data processing and analysis at CNAF

B Bertucci^{1,2}, **M Duranti**^{1,2}, **D D'Urso**^{1,3}, *

^{1} Università di Perugia, I-06100 Perugia, Italy
^{2} INFN, Sezione Perugia, I-06100 Perugia, Italy
^{3} ASDC, I-00133 Roma, Italy
AMS experiment http://ams.cern.ch, http://www.AMS-02.org,
http://www.pg.infn.it/ams/

E-mail: * domenico.durso@pg.infn.it

Abstract. AMS [1] is a large acceptance instrument conceived to search for anti-particles (positrons, anti-protons, anti-deutons) coming from dark matter annihilation, primordial anti-matter (anti-He or light anti nuclei) and to perform accurate measurements in space of the cosmic radiation in the GeV-TeV energy range. Installed on the International Space Station (ISS) in mid-May 2011, it is operating continuously since then, with a collected statistics of ~ 75 billion events up to the end of 2015. CNAF is one of the repositories of the full AMS data set and contributes to the data production and Monte Carlo simulation in the international collaboration. It represents the central computing resource for the data analysis performed by Italian collaboration. In the following, the AMS computing framework, data transfer to/from CNAF will be shortly discussed and use of the CNAF resources in 2015 will be given.

1. Introduction

AMS is a large acceptance instrument conceived to search for anti-particles (positrons, anti-protons, anti-deutons) coming from dark matter annihilation, primordial anti-matter (anti-He or light anti nuclei) and to perform accurate measurements in space of the cosmic radiation in the GeV-TeV energy range.

The layout of the AMS-02 detector is shown in Fig. 1. A large spectrometer is the core of the instrument: a magnetic field of 0.14 T generated by a permanent magnet deflects in opposite directions positive and negative particles whose trajectories are accurately measured up to TeV energies by means of 9 layers of double side silicon micro-strip detectors - the Tracker - with a spatial resolution of $\sim 10\mu m$ in the single point measurement along the track. Redundant measurements of the particle's characteristics, as velocity, absolute charge magnitude (Z), rigidity and energy are performed by a Time of Flight system, the tracker, a RICH detector and a 3D imaging calorimeter with a 17 X_0 depth. A transition radiation detector provides an independent e/p separation with a rejection power of $\sim 10^3$ around 100 GeV.

AMS has been installed on the International Space Station (ISS) in mid-May 2011 and it is operating continuously since then, with a collected statistics of ~ 75 billion events up to the end of 2015. The signals from the ~ 300.000 electronic channels of the detector and its monitoring system (thermal and pressure sensors) are reduced on board to match the average bandwidth of ~10 Mbit/s for the data transmission from space to ground, for a ~ 100 GB/day of raw data produced by the experiment.

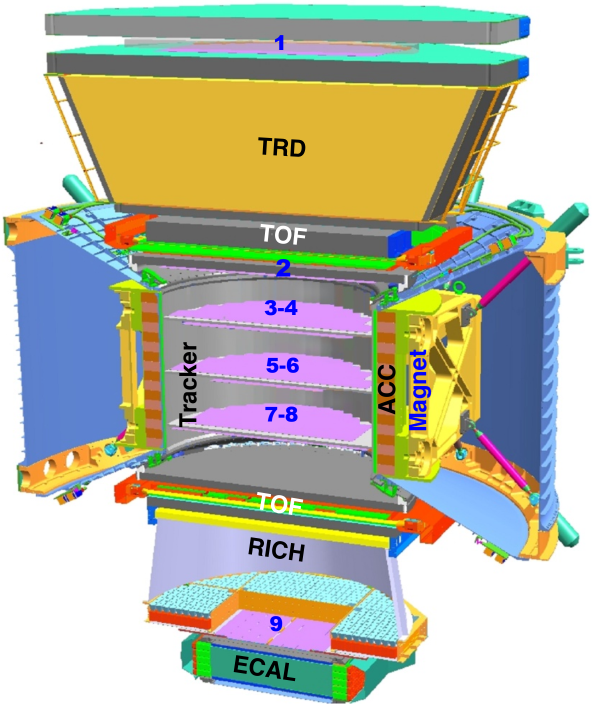


Figure 1. AMS-02 detector consists of nine planes of precision silicon tracker, a transition radiation detector (TRD), four planes of time of flight counters (TOF), a permanent magnet, an array of anticoincidence counters (ACC), surrounding the inner tracker, a ring imaging Cherenkov detector (RICH), and an electromagnetic calorimeter (ECAL).

Due to the rapidly changing environmental conditions along the ~ 90 minutes orbit of the ISS at 390 Km of altitude, continuous monitoring and adjustments of the data taking conditions are performed in the Payload and Operation Control Center (POCC) located at CERN and a careful calibration of the detector response is needed to process the raw data and reconstruct physics quantities for data analysis.

CNAF is one of the repositories of the full AMS data set, both raw and processed data are stored at CNAF which represents the central computing resource for the data analysis performed by Italian collaboration and contributes as well at the data production and Monte Carlo simulation in the international collaboration.

2. AMS-02 Computing Model and Computing Facilities

As a payload on the ISS, AMS has to be compliant to all of the standard communication protocols used by NASA to communicate with ISS, and its data have to be transmitted through the NASA communication network. On the ground, data are finally stored at the AMS Payload Operation Control Center (POCC) at CERN. Data are continuosly collected, 24 hours per day, 365 days per year. Data reconstruction pipeline is mainly composed by two logical step:

1) the **First Production** runs continuously over incoming data doing an initial validation and indexing. It produces the so called "standard" (STD) reconstructed data stream, ready within two hours after data are received at CERN, that is used to calibrate different sub-detectors as well as to monitor off-line the detector performances. In this stage Data Summary Files are produced for fast event selections.

2) Data from the First Production are reprocessed applying all of sub-detector calibrations, alignments, ancillary data from ISS and slow control data to produce reconstructed data for the physics analysis. This **Second production** step is usually applied in an incremental way to the std data sample, every 3-6 months - the time needed to produce and certify the calibrations. A full reprocessing of all AMS data is carried out periodically in case of major software major updates, providing the so called ”pass” production. Up to 2015 there were 6 full data reproductions done, published measurements were based on the pass6 data set.

The First Production is processed at CERN on a dedicated farm of about 200 cores, whereas Monte Carlo productions, ISS data reprocessing and user data analysis are supported by a network of computing centers (see fig. 2).

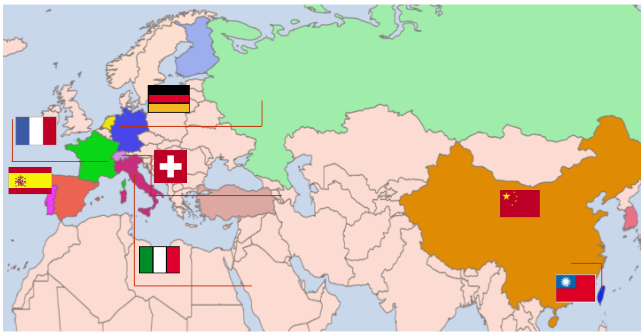


Figure 2. AMS-02 Major Contributors to Computing Resources.

China and Taiwan centers are mostly devoted to Monte Carlo production, while CERN, CNAF and FZJ Julich are the main centers for data reprocessing. A light-weight production platform has been realized to run on different computing centers, using different platforms. Based on perl, python and sqlite3, it is easily deployable and allows to have a fully automated production cycle, from job submission to monitoring, validation, transferring.

3. CNAF contribution

CNAF is the main computing resource for data analysis of the AMS Italian collaboration with 7904 HS06 and 1543 TB of storage allocated. A full copy of the AMS raw data is preserved on tape, the latest production and part of the Monte Carlo sample are available on disk. More then 50 users are routinely performing the bulk of their analysis at CNAF, transferring to local sites just reduced data sets or histograms.

An integrated solution, which enables transparent and efficient access to on-line and near-line data through high latency networks, has been implemented, between the CNAF (Bologna) and the ASI Science Data Center (ASDC) [2] in Rome to allow an efficient use of the local computing resources (384 cores and $\sim$ 100 TB). The solution is based on the use of the General Parallel File System (GPFS) and of the Tivoli Storage Manager (TSM). In the next months, that kind of solution will be implemented also between CNAF and the local computing resources in INFN-Perugia (150 cores and 90 TB).

4. Activities in 2015

AMS activities at CNAF in 2015 have been related to data reprocessing, Monte Carlo production and data analysis. Those activities have produced two publications reporting the measurement of proton [3] and helium flux [4] perfomed by AMS.

Two local queues are available for the AMS users: the default running is the *ams* queue, with a CPU limit of 26400 minutes and a maximum of 600 job running simultaneously, where as for test runs the *ams_short* queue, with high priority but a CPU limit of 360 minutes and a max 100 jobs running limit. For data reprocessing or MC production the AMS production queue *ams_prod*, with a CPU limit of 5760 minutes and 2000 jobs limit, is available and accessible only to data production team of the international collaboration and few experts users of the Italian team. In fact, the *ams_prod* queue is used within the data analysis process to produce data streams of pre-selected events and lightweight data files with a custom format [5] on the full AMS data statistics. In such a way, the final analysis can easily process the reduced data set avoiding the access to the large AMS data sample. The data-stream and custom data files productions are usually repeated few times a year.

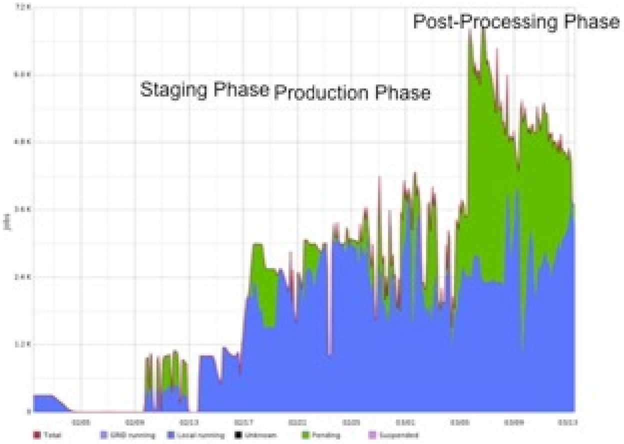


Figure 3. Submitted production jobs as a function of time from February to the March 13th .

Data reprocessing
From the February to the end of March, CNAF supported with additional 10k HS06 the reprocessing of AMS-02 data, to produce the AMS-02 pass6. During the first week, a massive staging of raw data from tape have been manged by CNAF staff while the production pipeline at CNAF has been widely tested. This test phase has been followed by the Production phase, where almost 10% of AMS data from May 2011 to October 2014 have been reprocessed and validated at CNAF and finally copied at CERN. After the Production phase, there was a post-production phase, dedicated to the copy to CNAF of all the available pass6 data and to process them to produce streams of events for user data analysis. To speed-up the production of event streams, data not yet available at CNAF have been processed on the CNAF farm directly from EOS at CERN, via xrootd [6] protocol. As shown in Figure 4, the number of ams running jobs (blue curve in the top plot) is strongly correlated with the amount of traffic rate of the Tier1, due to the processing of AMS data via xrootd, wiht a traffic peak of 20 Gb/s, the bandwidth limit of the CNAF/CERN connection. These activities has been the verification of xrootd as protocol for data access, it has been possible to have almost 3000 running jobs processing AMS-02 data. The effective bottle neck has been just the bandwidth limit.

Monte Carlo production
As part of the network AMS computing centers, CNAF has been involved in the Monte Carlo campaign devoted to the study of protons, helium nuclei and heavier ions for AMS publications.

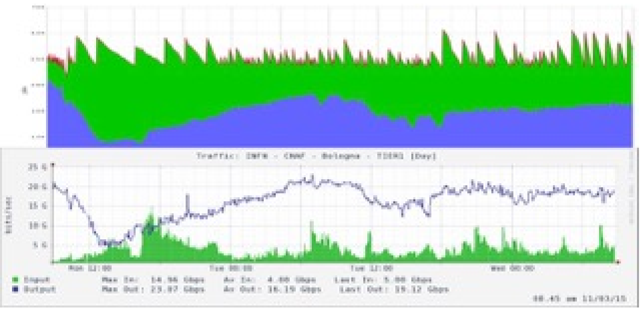


Figure 4. Number of pending (green) and running (blue) AMS jobs, in the top plot, and input (green)/output(blue) network traffic rate, on the lower plot, as a function of time
.

To support Monte Carlo campaign, special LSF profile has been implemented to allow AMS users to submit multi-thread simulation jobs. In the last months of 2015, CNAF supported the effort of the Monte Carlo campaign with an extra pledged of 20000 HS06. At CNAF more then 80 billions Monte Carlo events have been produced, equivalent to the 5% of the total amount of AMS simulations.

Data analysis
Different analysis are carried on by the Italian collaboration. In 2015, most of the CNAF resources for user analysis have been devoted to the electron/positron analysis, both in terms of flux measurement and anisotropies, the evaluation of the geomagnetic cutoff for all the on-going analyses in the collaboration, the study of time dependence of electron/positron and proton fluxes, the antiproton to proton ratio and ion analysis.

The disk resources pledged in 2015, ∼ 1.5 PB, were mostly devoted to the PASS4/PASS6 data sample (∼ 800 TB), MC data sample (∼ 400 TB), selected data streams (∼ 100 TB of pre-selected data used for common electron/positron, antiproton to proton and ion analysis) and scratch area for users.

References

[1] M.Aguilar *et al.*, AMS-02 Collaboration, Phys.Rev. Lett,110 (2013) ,141102.1-10
[2] http://www.asdc.asi.it
[3] M.Aguilar *et al.*, AMS-02 Collaboration, Phys.Rev. Lett,114 (2015) ,171103.1-9
[4] M.Aguilar *et al.*, AMS-02 Collaboration, Phys.Rev. Lett,115 (2015) ,211101.1-9
[5] D. D'Urso & M. Duranti, Journal of Physics: Conference Series, 664 (2015), 072016
[6] http://xrootd.org

ATLAS activities at the INFN CNAF Tier1

A De Salvo[1]

[1] INFN Roma 1, Roma, Italy

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 are briefly discussed, together with the impact of the Italian community.

1. Introduction

ATLAS is one of two general-purpose detectors at the Large Hadron Collider (LHC), as shown in Figure 1. 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. During the so-called Run-2 phase ATLAS collected and registered in 2015 at the Tier0 about 3.9 fb^{-1} of integrated luminosity at center of mass energies of 13 TeV.

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.

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.

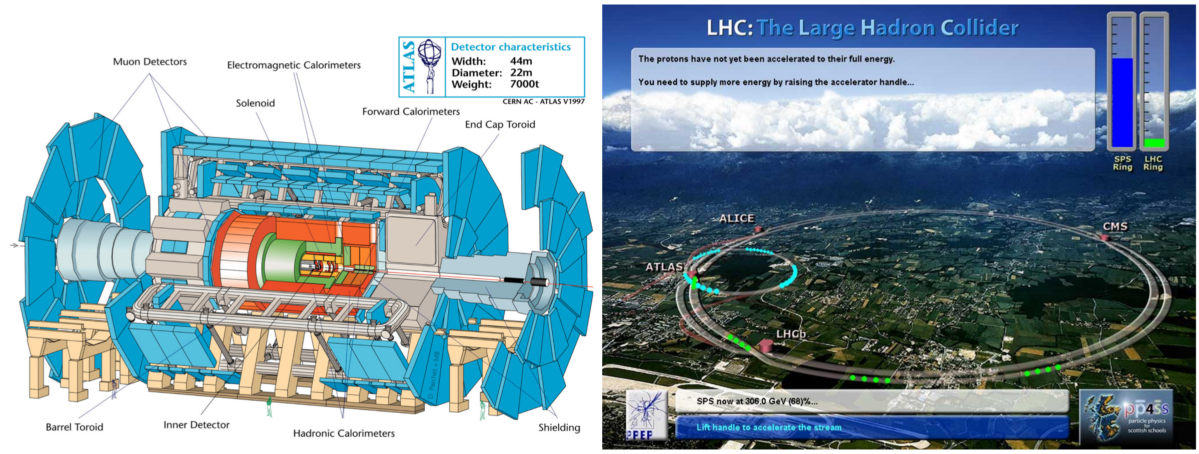

Muon Detectors
Electromagnetic Calorimeters
ATLAS
Detector characteristics
Width: 44m
Diameter: 22m
Weight: 7000t
CERN AC - ATLAS V1997
Solenoid
Forward Calorimeters
End Cap Toroid
Barrel Toroid
Inner Detector
Hadronic Calorimeters
Shielding
LHC: The Large Hadron Collider
The protons have not yet been accelerated to their full energy.
You need to supply more energy by raising the accelerator handle...
SPS Ring
LHC Ring
ALICE
CMS
ATLAS
LHCb
SPS now at 306.0 GeV (68)%...
Lift handle to accelerate the stream
pp4SS
particle physics for scottish schools

Figure 1. The ATLAS experiment at LHC.

2. 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 and 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.

2.1. 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 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: EGI, 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 and 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 and Computing Project and the national resources and their Funding Agencies.

3. 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. All 4 Tier-2 sites are 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. 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 9% 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 2015 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, reprocessing 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 fine measurement of the Higgs [3] properties in 2015.

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 systems, the storage federations and the monitoring tools.

The T1 at CNAF has been very important for the ATLAS community in 2015, for some specific activities:

(i) 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;

(ii) improvements on the WebDAV/HTTPS access for StoRM, in order to be used as main renaming method for the ATLAS files in StoRM and for http federation purposes;

(iii) improvements of the dynamic model of the multi-core resources operated via the LSF resource management system;

(iv) network throubleshooting via the Perfsonar-PS network monitoring system, used for the LHCONE overlay network, together with the other T1 and T2 sites;

(v) planning, readiness testing and implementation of StoRM for the future infrastructure of WLCG

(vi) prototyping of new accesses to resources, including the Cloud Computing Infrastructures.

4. Main achievements of ATLAS Computing centers in Italy

The computing activities of the ATLAS collaboration have been constantly carried out over the whole 2015, in order to finalize the analysis of the data of the Run-1, produce the Monte Carlo data needed for the 2015 run and analyse the data of the first part of Run-2.

The LHC data taking started in June 2015 and, until the end of the operation in December 2015, the CNAF Tier1 and the four Tier2s, have been involved in all the computing operations of the collaboration: data reconstruction, Monte Carlo simulation, user and group analysis and data transfer among all the sites.

Besides these activities, the Italian centers have contributed to the upgrade of the Computing Model both from the testing side and the development of specific working groups. Several improvements in the Computing Model has been achieved in 2014 and the first part of 2015, more precisely in the software domain and the infrastructure.

ATLAS collected and registered at the Tier0 about 3.9 fb^{-1} and about 8 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 16 PB.

The data has been replicated with an efficiency of 100% and an average throughput of the order of about 7 GB/s during the data taking period, with monthly average peaks above 8 GB/s. For just Italy, the average throughput was of the order of 400 MB/s with monthly average peaks around 500 MB/s. The average number of simultaneous jobs running on the grid has been of about 100k for production (simulation and reconstruction) and data analysis, with peaks up to 130k in Sep/Nov, with an average CPU efficiency up to more than 90%.

The use of the grid for analysis has been stable on about 30k simultaneous jobs, with peaks around the conference periods to over 48k, 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, ATLAS introduced the so-called Federation of Xrootd storage systems (FAX), on top of the existing infrastructure. Using FAX, the users have the possibility to access remote files via the XRootd protocol in a transparent way, using a global namespace and a hirerarchy 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 FAX federation is now in production mode and used as failover in many analysis tasks.

The contribution of the Italian sites to the computing activities in terms of processed jobs and data recorded has been of about 9%, corresponding to the order of the resource pledged to the collaboration, with very good performance in term of availability, reliability and efficiency. All the sites are always in the top positions in the ranking of the collaboration sites.

Besides the Tier1 and Tier2s, in the 2015 also the Tier3s gave a significant contribution to the Italian physicists community for the data analysis. The Tier3s are local farms dedicated to the interactive data analysis, the last step of the analysis workflow, and to the grid analysis over small data sample. Many italian groups set up a farm for such a purpose in their universities and, after a testing and validation process performed by the distributed computing team of the collaboration, all have been recognized as official Tier3s of the collaboration.

5. 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_2015_12.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.

1. Introduction

The Pierre Auger Observatory, built near the town of Malargüe in Argentina, has been gathering data since January 2004 [1]. Measurements of the Auger Observatory have dramatically advanced our understanding of ultra-high energy cosmic rays (UHECRs). Particularly exciting is the observed behavior of the depth of shower maximum with energy, which changes in an unexpected, non-trivial way. Around 3×10^{18}eV it shows a distinct change of slope with energy, and the shower-to-shower variance decreases. Interpreted with the leading LHC-tuned shower models, this implies a gradual shift to a heavier composition. A number of fundamentally different astrophysical model scenarios have been developed to describe this evolution. The high degree of isotropy observed in numerous tests of the small-scale angular distribution of UHECRs above 4×10^{19}eV is remarkable, challenging original expectations that assumed only a few cosmic ray sources with a light composition at the highest energies. Interestingly, the largest departures from isotropy are observed for cosmic rays with E above 5×10^{19} eV in $\sim$20° sky windows . Due to a duty cycle of $\sim$ 15% of the fluorescence telescopes, the data on the depth of shower maximum extend only up to the flux suppression region, i.e. 4×10^{19}eV. Obtaining more information on the composition of cosmic rays at higher energies is of central importance for making further progress in understanding UHECRs.

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 spectrum 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 specifically 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 is calculated using a sample of simulated events that reproduce the exact conditions of the experiment. This simulation is

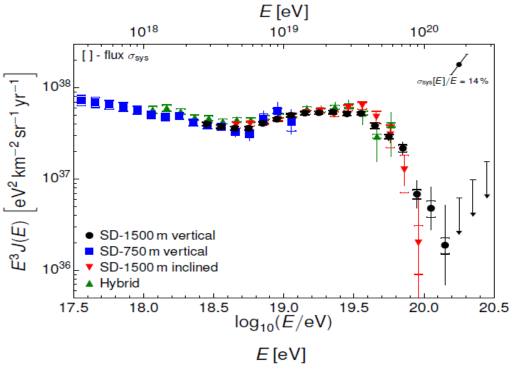


Figure 1. Energy spectra derived from SD and hybrid data recorded at the Pierre Auger Observatory. The error bars represent statistical uncertainties. The upper limits correspond to the 84% C.L.

performed at CNAF. The measurement of the energy spectrum is derived from vertical SD data sets recorded by both the 750m and 1500m arrays up to 31 Dec 2014, and hybrid data up to 31 Dec 2013, togheter with the spectrum derived from inclined events recorded by the 1500m array up to 31 Dec 2013 [4].

The four independent measurements of the energy spectrum of cosmic rays are shown in Figure 1. The comparison shows that all spectra are in agreement within uncertainties. The four independent measurements of the energy spectrum of cosmic rays are then combined using a method that takes into account the systematic uncertainties of the individual measurements. The combined spectrum shows a flattening above the ankle energy 4 $\times 10^{18}$ eV, up to the onset of the flux suppression. This suppression is clearly established with a significance of more than 20 σ.

5. The upgrade program of the experiment

The Auger upgrade operation is now planned from 2018 until 2024, event statistics will more than double compared with the existing Auger data set, with the critical added advantage that every event will now have mass information. This will allow us to address some of the most pressing questions in UHECR physics, including that of the origin of the flux suppression, the prospects of light particle astronomy and secondary particle fluxes, and the possibility of new particle physics at extreme energies. Obtaining additional composition-sensitive information will not only help to better reconstruct the properties of the primary particles at the highest energies, but also improve the measurements in the important energy range just above the ankle. Furthermore, measurements with the new detectors will help to reduce systematic uncertainties related to modeling hadronic showers and to limitations of reconstruction algorithms. This improved knowledge of air-shower physics will likely then also allow a re-analysis of existing data for improved energy assignments, for mass composition studies, and for photon and neutrino searches. Several massive productions in the Simulation-Reconstruction framework have been performed in order to evaluate and optimize the impact of the upgrade.

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] The Pierre Auger Collaboration, accepted by JCAP (2014) arXiv:1503.07786.

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 second physics Run, at centre of mass energy of 13TeV, started in late Spring 2015, and ended in November 2015; more than 4 fb-1 of collisions were collected. 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 for Run II are on one side on the completion of the Standard Model, with a precision study of the Higgs sector, on the other side on the discovery of 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 450 physics papers were produced from Run I data, including the now renowned paper on the Observation of a 126 GeV Higgs Boson, which sets the final cornerstone to the Standard Model. Publications on Run II data are being prepared, with 6 already published at the end of March 2015.

2. The CMS Computing Model

CMS trigger rates increased to 1.5 kHz for Run II (standard + parking); this, combined with large event sizes and computational needs, has 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 Runs I and II 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 amount to the 13% of the total Tier1 resources, the fraction being equal to the fraction of the Italian component in CMS; they amount (2015 numbers) to

- 39 kHS06 computational power;

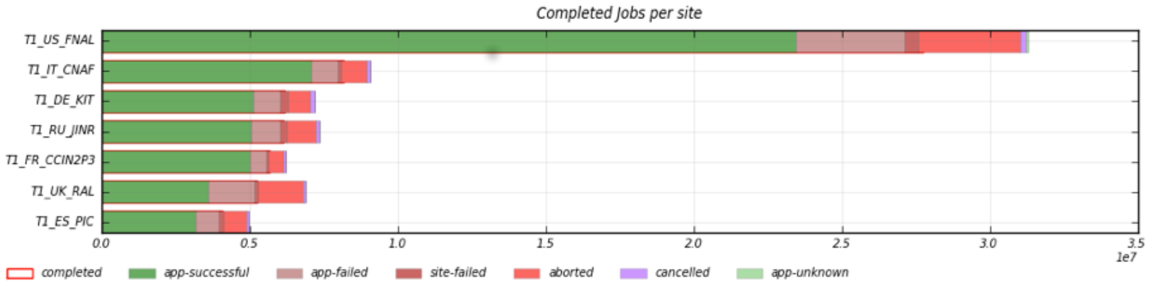


Figure 1. Number of jobs processed by each CMS Tier1 during 2015.

- 9620 TB of tape;
- 338 TB of disk.

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 tape resource has been used at levels exceeding 90%, resulting in CNAF as the Tier1 holding more custodial data, again after the US Tier1. The disk resource has been used to more than 95%, as controlled by CMS Dynamic Disk Management system.

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/GPFS solution is being investigated or implemented at a number of CMS Tier2 sites.

Starting from the end of Run I, the CMS storage setup has evolved at CNAF. Access to the files has been granted from remote locations via the Xrootd access protocol, and later 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, protecting the tape area from chaotic accesses; indeed, the files on tape are accessible only after an explicit movement to the disk area. The new setup for the CPU + the disk area reduced significantly the differences between a Tier1 and a Tier2; indeed, during 2017 CNAF has opened the batch queues also for the standard analysis jobs, ramped from virtually zero at the end of 2013, to O(30%) level, limited just by the higher priority of production.

During 2015, CNAF has again been the second Tier1 in CMS as number of processed jobs, as already since 2012 (see Fig.1).

3. New developments

A complex system like a CMS Tier1 is under constant change, in order to keep components up to date, and to introduce new features. During 2015, CMS has used CNAF as a testbed for elastic expansion of resources beyond the pledged ones. In particular, two efforts should be noted:

- the utilization of remote resources at ReCaS/Bari, as part of the standard pool of computing resources. This has been implemented in a transparent way by integrating those resources into the LSF batch system;
- the utilization of remote commercial cloud systems. The first test (now in production), has been in collaboration with Aruba, a major italian Cloud provider. The setup has used VPN

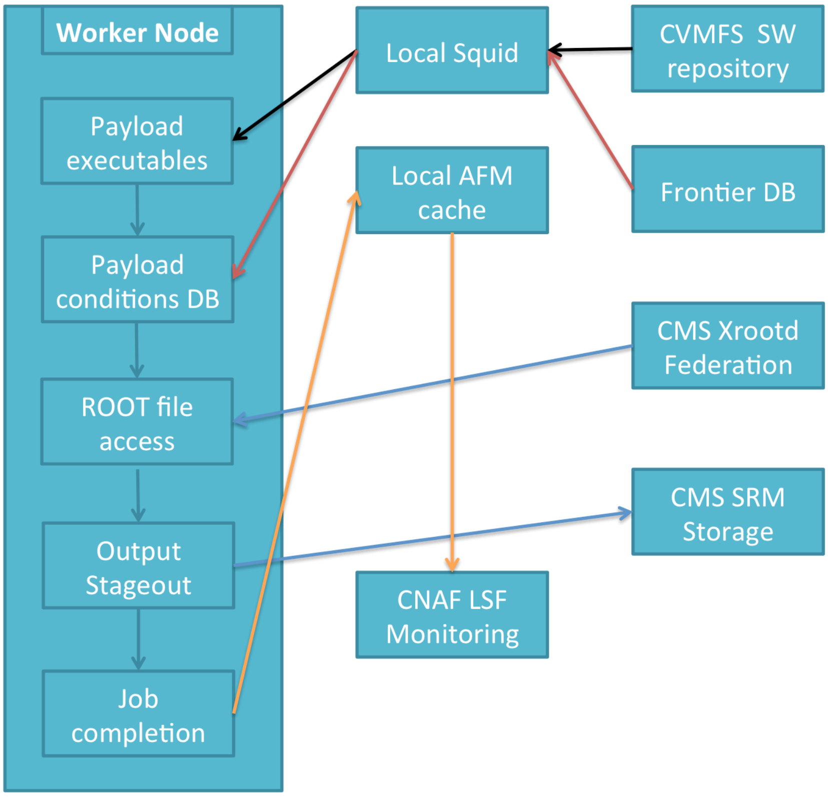


Figure 2. Interactions between CMS processes running on a Aruba Worker Node and CMS central services.

connections (using an ad-hoc CNAF system) between machines in Arezzo and CNAF, in order to merge the remote resources with CNAF's LSF batch system. The schema of the Aruba setup is shown in Fig2.

In both cases, a major problem of the setup has been the absence of storage local to the new resources. The solution has been to fallback to Xrootd usage for tasks running not at CNAF; this is automatic for LHC Experiments' workflows, but not a slution for other VOs. At Bari, an attempt using GPFS AFM caching has been deployed, as a general solution, and is currenty under tuning.

4. Expected resource growth

The LHC collider is at the moment (April 2016) about to restart for the second data taking year in Run II. CMS expects to carry on an extensive study of the Higgs boson properties during Run II, while performing searches for new physics at the newly available energy. In 2016, the instanstaneous LHC luminosity is expected to exceed $15^{34}\mathrm{cm}^2\mathrm{s}^{-1}$, a factor 2 with respect to 2015. This requires additional computing resources at CNAF, and indeed 2016 pledges are set to 48 kHS06 for CPU ; disk pledges are at the level of 3960 TB, and tape pledges to 12 PB. Expectations for 2017 prospect another large increase, due to anothe jump in LHC performance, with CPU reaching 66 kHS06, disk 5.8 PB and tape 16 PB, and are currently under scrutiny by the RRB-CRSG scrutiny group.

The Cherenkov Telescope Array

Ciro Bigongiari
INAF - Osservatorio Astrofisico di Torino,
Strada Osservatorio 20 - 10025 Pino Torinese, Torino, IT
E-mail: bigongiari@oato.inaf.it

Abstract. The Cherenkov Telescope Array (CTA) is an ongoing international project to build a new generation ground-based gamma-ray detector composed by tens of Cherenkov telescopes of different sizes. Imaging Cherenkov telescopes have already discovered more than 170 VHE gamma-ray emitters providing plentiful of valuable data and clearly demonstrating the power of this technique. CTA aims to increase the sensitivity by an order of magnitude compared to current facilities, to extend the accessible gamma-ray energies from a few tens of GeV to hundreds of TeV, and to improve on other parameters like angular and energy resolution. CTA will cover the full sky by featuring an array of imaging atmospheric Cherenkov telescopes in both hemispheres and will be operated as an open observatory. CTA project combines guaranteed scientific return, in the form of high precision astrophysics, with considerable potential for major discoveries in astrophysics and fundamental physics.

1. Introduction

CTA will provide a deep insight into the non-thermal processes which are responsible of the high energy emission by many astrophysical sources, like Supernova Remnants, Pulsar Wind Nebulae, Micro-quasars, Active Galactic Nuclei and Gamma Ray Bursts. Very High Energy gamma-rays can be produced in the collision of highly relativistic particles with surrounding gas clouds or in their interaction with low energy photons or magnetic fields. Possible sources of such energetic particles include jets emerging from active galactic nuclei, remnants of supernova explosions, and the environment of rapidly spinning neutron stars. High-energy gamma-rays can also be produced in top-down scenarios by the decay of heavy particles such as hypothetical dark matter candidates or cosmic strings. The CTA observations will be used for detailed studies of above-mentioned astrophysical sources as well as for fundamental physics measurements, such as the indirect search of dark matter, searches for high energy violation of Lorentz invariance and searches for axion-like particles. High-energy gamma-rays can be used moreover to trace the populations of high-energy particles, thus providing insightful information about the sources of cosmic rays. Close cooperation with observatories of other wavelength ranges of the electromagnetic spectrum, and those using cosmic rays, neutrinos and gravitational waves are foreseen. To ensure a full coverage of the sky the CTA detector will be composed actually by two arrays, one placed in the Southern hemisphere and one in the Northern one. Negotiations with ESO (European Southern Observatory) to place the Southern array at a site close to Cerro Paranal, Chile, are ongoing, as well as negotiations with Spain to build the Northern one at the Observatorio del Roque de Los Muchachos (ORM) on the La Palma island. It has already been decided that the prototype of the large size telescope will be built at ORM and the foundation laying is foreseen by July 2016. CTA Observatory is expected to become fully operational in

2020. A detailed description of the project and its expected performance can be found in a dedicated volume of the Astroparticle Physics journal [1].

2. Computing needs

The CTA project is presently in the pre-production phase when many innovative technologies are being tested and developed for the construction of the various classes of telescopes. Meanwhile detailed Monte Carlo simulation of the entire array are ongoing to estimate its overall performance and to optimize many parameters like the telescope layout and the trigger strategy. A huge effort has been dedicated so far to the evaluation of the expected performance at many different site candidates to provide the information needed for a fair comparison to the site selection committee. Each site requires a full simulation because the performance of a Cherenkov telescope depends on many site-dependent parameters like its altitude, atmospheric conditions, geomagnetic field and night-sky brightness. Due to the very effective hadronic background rejection achieved with the imaging air Cherenkov technique a huge amount of simulated background events is needed to achieve reliable estimates of the array performance. About 10^{10} cosmic ray induced atmospheric showers for each site are needed to properly estimate the array sensitivity, energy and angular resolution requiring extensive computing needs in term of both disk space and CPU power. About 1.9 million of GRID jobs have been executed in 2015 for such task corresponding to about 88.6 millions of HS06 hours of CPU power and 1840 TB of disk space. CNAF contributed to this effort with about 12.4 millions of HS06 hours and 172 TB of disk space, corresponding to 14% of the overall CPU power used and the 9% of the disk space, resulting the second contributor in terms of CPU time and the third in terms of disk space. For 2016 even larger needs of CPU power and disk space are foreseen due to the ongoing simulations for the array layout optimization of the selected sites. CNAF will surely keep its role among the main contributors to the CTA virtual organization thanks to the greatly increased dedicated resources.

References

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

The Borexino experiment at the INFN CNAF Tier1

Alessandra Carlotta Re[a] **and Alessio Caminata**[b]
on behalf of the BOREXINO collaboration

[a]Università degli Studi e INFN di Milano, via Celoria 16, 20133 Milano, Italy
[b]Università degli Studi e INFN di Genova, via Dodecaneso 33, 16146 Genova, Italy

E-mail: alessandra.re@mi.infn.it, alessio.caminata@ge.infn.it

Abstract. Borexino is a large-volume liquid scintillator experiment designed for low energy neutrino detection, installed at the National Laboratory of Gran Sasso (LNGS) and operating since May 2007. The exceptional levels of radiopurity Borexino has reached through the years, have made it possible to accomplish not only its primary goal but also to produce many other interesting results both within and beyond the Standard Model of particle physics.

1. Introduction

Borexino is an experiment originally designed for real-time detection of low energy solar neutrinos. It is installed at the INFN underground National Laboratory of Gran Sasso (Assergi, Italy) where the average rock cover is about 1,400 m with resulting in a shielding capacity against cosmic rays of 3,800 meter water equivalent (m.w.e.): at the LNGS, the muon flux is reduced of a factor 10^6 respect to the surface.

In Borexino, neutrinos are detected via elastic scattering of the liquid scintillator electrons. The active target consists of 278 tons of pseudocumene (1,2,4-trimethylbenzene) doped with 1.5 g/L of a fluorescent dye (PPO, 2,5-diphenyloxazolo) and it converts the energy deposited by neutrino interactions into light. The detector is instrumented with photomultiplier tubes that can measure the intensity and the arrival time of this light, allowing the reconstruction of the energy, position and time of the events. The Borexino detector was designed exploiting the principle of graded shielding: an onion-like structure allows to protect the inner part from external radiation and from radiation produced in the external shielding layers. The requirements on material radiopurity increase when moving to the innermost region of the detector[1].

2. The Borexino recent result and future perspectives

Borexino started taking data in 2007 and, since then, it has been producing a considerable amount of results including the first direct measurement of proton-proton solar neutrino interaction rate, the precision measurement of the ^{7}Be solar neutrino rate (with a total error of less than 5%), the first direct measurement of the so-called pep solar neutrinos and the measurement of the ^{8}B solar neutrino rate with an unprecedented low energy threshold. Borexino has also published significant results on non-solar neutrino physics, such as the first observation of anti-neutrinos from the Earth (the geoneutrinos) and several limits on rare or forbidden processes.

Among the most important scientific results obtained during 2015 we recall the test of electric charge conservation[2] and the spectroscopy of geoneutrinos from 2056 days of Borexino data[3].

Besides its application in the solar physics and geophysics fields, the Borexino detector offers a unique opportunity to perform a short-baseline neutrino oscillation study. This is the idea of SOX (Short distance neutrino Oscillations with boreXino). The SOX experiment[4] aims at the complete confirmation or at a clear disproof of the so-called neutrino anomalies, a set of circumstantial evidences of electron neutrino disappearance observed at LSND, MiniBoone, with nuclear reactors and with solar neutrino Gallium detectors. If successful, SOX will demonstrate the existence of sterile neutrino components and will open a brand new era in fundamental particle physics and cosmology. A solid signal would mean the discovery of the first particles beyond the Standard Electroweak Model and would have profound implications in our understanding of the Universe and of fundamental particle physics. In case of a negative result, SOX would be able to close a long-standing debate about the reality of the neutrino anomalies, would probe the existence of new physics in low energy neutrino interactions, would provide a measurement of the neutrino magnetic moment, and would yield a superb energy calibration for Borexino which will be very beneficial for future high-precision solar neutrino measurements. The SOX experiment will use a powerful and innovative antineutrino generator made of ^{144}Ce. This generator will be located at a short distance from the Borexino detector and will yield tens of thousands of clean antineutrino interactions in the internal volume of the Borexino detector. The SOX experiment is expected to start in spring 2017 and will take data for about two years.

3. Borexino computing at CNAF

At present, the whole Borexino data statistics and the user areas for physics studies are hosted at CNAF. The Borexino data are classified into three types: raw data, root files and DSTs. Raw data are compressed binary files with a typical size of about 600 Mb corresponding to a data taking time of ∼6h. Root files are reconstructed events files each organized in a number of `ROOT TTree`: their typical dimension is ∼1Gb. A DST file contains only selected events for high level analyses. Borexino standard data taking requires a disk space increase of about 10 Tb/year while a complete Monte Carlo simulation of both neutrino signals and backgrounds requires about 7 Tb/DAQ year.

CNAF front-end machine (`ui-borexino.cr.cnaf.infn.it`) and pledged CPU resources (about 100 cnodes) are currently used for root files production, Monte Carlo simulations, interactive and batch analysis jobs. For few weeks a year, an extraordinary *peak usage* (up to 500 cnodes at least) is needed in order to perform a full reprocessing on the whole data statistics with an updated version of the reconstruction code.

4. Conclusions

During next years, the amount of CNAF resources needed and used by the Borexino experiment is expected to increase. In fact, Borexino will not only continue in its rich solar neutrino program with the ambitious target of CNO neutrino flux measurement but will also be devoted to the SOX project, a short baseline experiment, aiming at a clear proof or disproof of the sterile-neutrino hypothesis.

References

[1] Alimonti G. et al. 2009 *Nucl. Instrum. Methods A* **600** 568.

[2] Agostini M. et al. 2015 *Phys. Rev. Lett.*, **115** 231802.

[3] Agostini M. et al. 2015 *Phys. Rev. D* **92** 031101.

[4] Bellini G. et al. 2013 *JHEP* **8** 038.

CUORE experiment

CUORE collaboration

E-mail: cuore-spokesperson@lngs.infn.it

Abstract. CUORE is a ton scale bolometric experiment for the search of neutrinoless double beta decay in ^{130}Te. The detector is in the advanced commissioning phase at the Laboratori Nazionali del Gran Sasso of INFN, in Italy. It is composed by an array of 988 TeO_2 bolometers, for a total mass of 0.75 ton. The projected CUORE sensitivity for the neutrinoless double beta decay half life of ^{130}Te is of 10^{26} y after five years of live time. The configuration of the CUORE data processing environment on the CNAF computing cluster is almost complete, and a more intense use of resources is expected in 2016.

1. The experiment

The main goal of the CUORE experiment [1] is to search for neutrinoless double beta decay (0νDBD) of the isotope ^{130}Te. In this spontaneous decay a nucleus changes its atomic number by two units, and two electrons are emitted. Its observation would imply the Majorana nature of the neutrino mass, and could give information on the neutrino mass hierarchy and absolute scale. To date there is no experimental evidence for this decay, and the half life limits lie in the range $10^{22} \div 10^{25}$ y, depending on the isotope that is being considered. In a calorimetric detector, the sum energy of the two electrons emitted in 0νDBD produces a sharp peak in the spectrum, centered at the Q-value of the decay. Typical 0νDBD Q-values are in the few MeV range, therefore the tiny signal is submerged by background from natural radioactive decays. The CUORE detector is an array of 988 $^{nat}TeO_2$ bolometers, with a total mass of 741 kg (206 kg of ^{130}Te). The bolometers are arranged in 19 towers, each tower is composed by 13 floors of 4 bolometers each. A single bolometer is a cubic TeO_2 crystal with 5 cm side and a mass of 0.75 kg. The bolometer array is enclosed in a dilution refrigerator whose mixing chamber is cooled to ~10 mK and thermally coupled to the copper support structure holding the detectors. The CUORE bolometers act at the same time as source and detectors for the sought signal. When a particle interacts in a CUORE crystal, it produces a sizable temperature rise, $\Delta T = E/C$, that can be read by a NTD thermal sensor. The CUORE collaboration aims at reaching a background of 10^{-2} counts/(keV·kg·y) in the region of the energy spectrum where the 0νDBD signal is expected ($Q_{\beta\beta} \simeq 2528$ keV), and a FWHM energy resolution of 5 keV. With these parameters, the experiment will reach a ^{130}Te half-life sensitivity of about 10^{26} y in five years of live time.

2. Status of CUORE and CUORE-0

The CUORE experiment is currently in the advanced commissioning phase at the Laboratori Nazionali del Gran Sasso of the INFN, Italy. All the 19 bolometer towers were successfully assembled and are now stored underground in nitrogen overpressure. The commissioning of the CUORE cryostat is complete. The installation of the CUORE readout electronics and data acquisition system is planned for the first months of 2016, and it will be followed by the installation of the bolometer towers in the cryostat. The detector cool down is foreseen for the second half of 2016.

To check the effectiveness of the CUORE detector assembly procedure, a first CUORE-like tower made of 52 bolometers, named CUORE-0 [2], was operated in the former Cuoricino [3] cryostat from 2013 to 2015. From a study of the CUORE-0 background spectrum and with the aid of Monte Carlo simulations, it could be possible to evince that the CUORE background goal of 10^{-2} counts/(keV·kg·y) is within reach. CUORE-0 measured an average energy resolution slightly better than 5 keV FWHM on the 2615 keV photoelectric peak from ^{208}Tl, in perfect agreement with the energy resolution goal of CUORE. Finally, CUORE-0 set a limit on the neutrinoless double beta decay half life of ^{130}Te, $T^{0\nu}_{1/2} > 2.7 \times 10^{24}$ y at 90% C.L. [4]. When combined with the previous result from Cuoricino, the most stringent limit available for this decay is obtained, $T^{0\nu}_{1/2}(^{130}\mathrm{Te}) > 4.0 \times 10^{24}$ y at 90% C.L.

3. CUORE computing model and the role of CNAF

The CUORE raw data consist in Root files containing events in correspondence with energy releases occurred in the bolometers. Each event contains the waveform of the triggering bolometer and of those geometrically close to it, plus some ancillary information. The non event-based information is stored in a PostgreSQL database that is also accessed by the offline data analysis software. The data taking is organized in runs, each run lasting about one day. Raw data are transferred from the DAQ computers to the permanent storage area at the end of each run. In CUORE-0 about 500 GB/y of raw data were produced, while for CUORE about 20 TB/y of raw data are expected.

The CUORE data analysis flow consists in two steps. In the first level analysis the event-based quantities are evaluated, while in the second level analysis the energy spectra are produced and studied. The analysis software is organized in sequences. Each sequence consists in a collection of modules that scan the events in the Root files sequentially, evaluate some relevant quantities and store them back in the events. The analysis flow consists in several fundamental steps that can be summarized in pulse amplitude estimation, detector gain correction, energy calibration and search for events in coincidence among multiple bolometers.

The CUORE-0 data analysis and simulations were run on a computing cluster located at the Roma1 division of INFN. In view of the start of the CUORE data taking, since 2014 a transition phase has started to move the CUORE analysis and simulation framework to CNAF. The framework is now installed and it is ready for being used. In 2015 most of the CUORE Monte Carlo simulations were run at CNAF, and some tests of the data analysis framework were performed using mock-up data. In 2016, when the first CUORE data will be available, a more intense usage of the CNAF resources is expected, both in terms of computing resourced and storage space.

References

[1] Artusa D *et al.* (CUORE) 2015 *Adv.High Energy Phys.* **2015** 879871 (*Preprint* 1402.6072)

[2] Artusa D *et al.* (CUORE) 2014 *Eur.Phys.J.* **C74** 2956 (*Preprint* 1402.0922)

[3] Andreotti E *et al.* 2011 *Astropart.Phys.* **34** 822–831 (*Preprint* 1012.3266)

[4] Alfonso K *et al.* (CUORE) 2015 *Phys. Rev. Lett.* **115** 102502 (*Preprint* 1504.02454)

DAMPE data processing and analysis at CNAF

G. Ambrosi[1], **D. D'Urso**[1,2,*], **M. Duranti**[1,3], **F. Gargano**[4,5], **S. Zimmer**[6]

[1] INFN, Sezione di Perugia, I-06100 Perugia, Italy
[2] ASDC, I-00133 Roma, Italy
[3] Università di Perugia, I-06100 Perugia, Italy
[4] INFN, Sezione di Bari, I-70125 Bari, Italy
[5] Università di Bari, I-70125 Bari, Italy
[6] University of Geneva, Departement de physique nuclaire et corpusculaire (DPNC), CH-1211, Genève 4, Switzerland
DAMPE experiment `http://dpnc.unige.ch/dampe/`, `http://dampe.pg.infn.it`

E-mail: `domenico.durso@pg.infn.it`

Abstract. DAMPE (DArk Matter Particle Explorer) is one of the five satellite missions in the framework of the Strategic Pioneer Research Program in Space Science of the Chinese Academy of Sciences (CAS). DAMPE has been launched the 17 December 2015 at 08:12 Beijing time into a sun-synchronous orbit at the altitude of 500 km. The satellite is equipped with a powerful space telescope for high energy gamma-ray, electron and cosmic ray detection. The main scientific objective of DAMPE is to measure electrons and photons with much higher energy resolution and energy reach than achievable with existing space experiments in order to identify possible Dark Matter signatures. It has also great potential in advancing the understanding of the origin and propagation mechanism of high energy cosmic rays, as well as may enable new discoveries in high energy gamma-ray astronomy.

1. Introduction

DAMPE is a powerful space telescope for high energy gamma-ray, electron and cosmic ray detection. In Fig. 1 a scheme of the DAMPE telescope is shown. The top, the plastic scintillator strip detector (Psd) consists of two layers of scintillating plastic strips that serve as anti-coincidence detector, followed by a silicon-tungsten tracker-converter (STK), which is made of 6 tracking layers. Each tracking layer consists of two layers of single-sided silicon strip detectors measuring the two orthogonal views perpendicular to the pointing direction of the apparatus. Three layers of Tungsten plates with thickness of 1 mm are inserted in front of tracking layer 2, 3 and 4 to promote photon conversion into electron-positron pairs. The STK is followed by an imaging calorimeter of about 31 radiation lengths thickness, made up of 14 layers of Bismuth Germanium Oxide (BGO) bars which are placed in a hodoscopic arrangement. The total thickness of the BGO and the STK correspond to about 33 radiation lengths, making it the deepest calorimeter ever used in space. Finally, in order to detect delayed neutron resulting from hadron showers and to improve the electron/proton separation power, a neutron detector (NUD) is placed just below the calorimeter. The NUD consists of 16, 1 cm thick, boron-doped plastic scintillator plates of 19.5 × 19.5 cm^2 large, each read out by a photomultiplier.

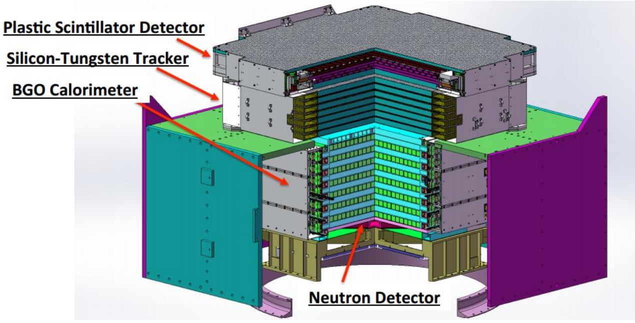


Figure 1. DAMPE telescope scheme:a double layer of the plastic scintillator strip detector (PSD); the silicon-tungsten tracker-converter (STK) made of 6 tracking double layers; the imaging calorimeter with about 31 radiation lengths thickness, made of 14 layers of Bismuth Germanium Oxide (BGO) bars in a hodoscopic arrangement and finally the neutron detector (NUD) placed just below the calorimeter.

The primary scientific goal of DAMPE is to measure electrons and photons with much higher energy resolution and energy reach than achievable with existing space experiments. This will help to identify possible Dark Matter signatures but also may advance our understanding of the origin and propagation mechanisms of high energy cosmic rays and possibly lead to new discoveries in high energy gamma-ray astronomy.

DAMPE was designed to have an unprecedented sensitivity and energy reach for electrons, photons and cosmic rays (proton and heavy ions). For electrons and photons, the detection range is 2 GeV-10 TeV, with an energy resolution of about 1.5% at 100 GeV. For cosmic rays, the detection range is 100 GeV-100 TeV, with an energy resolution better than 40% at 800 GeV. The geometrical factor is about 0.3 m^2 sr for electrons and photons, and about 0.2 m^2 sr for cosmic rays. The expected angular resolution is 0.1° at 100 GeV.

2. DAMPE Computing Model and Computing Facilities

As Chinese satellite, DAMPE data are collected via the Chinese space communication system and transmitted to the China National Space Administration (CNSA) center in Beijing. From Beijing data are then transmitted to the Purple Mountain Observatory (PMO) in Nanjing, where they are processed and reconstructed.

2.1. Data production

PMO is the deputed center for DAMPE data production. Data are collected 4 times per day, each time the DAMPE satellite is passing over Chinese ground stations (almost every 6 hours). Once transferred to PMO, binary data, downloaded from the satellite, are processed to produce a stream of raw data in ROOT [1] format (1B data stream, $\sim$ 15 GB/day), and a second stream that include the orbital and slow control information (1F data stream, $\sim$ 15GB/day). The 1B and 1F streams are used to derive calibration files for the different subdetectors ($\sim$ 400MB/day). Finally, data are reconstructed using the DAMPE official reconstruction code, and the so-called 2A data stream (ROOT files, $\sim$ 70 GB/day) is produced. The total amount of data volume produced per day is $\sim$ 100 GB.

2.2. Monte Carlo Production

Analysis of DAMPE data requires large amounts of Monte Carlo simulation, to fully understand detector capabilities, measurement limits and systematics. In order to facilitate easy workflow handling and management and also enable efficient monitoring of a large number of batch jobs in various states, a NoSQL metadata database using MongoDB [2] is being developed with a prototype currently running at the Physics Department of Geneva University. Database access is provided through a web-frontend and command tools based on the flask-web toolkit [3] with a client-backend of cron scripts that will run on the selected computing farm. The design and implementation of this workflow system is heavily influenced by the implementation of the Fermi-LAT data processing pipeline [4] and the DIRAC computing framework [5].

Fundamentally, each production consists of repeated batch submissions with minimal differences across different batch jobs. To this end, a MC shifter creates a *Job* which contains the basic information regarding the type of task (MC generation, digitization, reconstruction, etc.) and each *Job* contains a number of *JobInstances*. Web-frontend and command line tools provide means to manage and access these Jobs and JobInstances. On the computing farm a set of cron jobs establish connections with the web-frontend (and by extension the database) and request JobInstances of type *New*. Once a new instance is pulled from the database, a batch job is created and submitted. Upon submission the batch ID is reported back to the database and used for further tracking. This cron job is supplemented by a second job which acts as a watchdog that monitors running jobs in the batch farm and requests jobs to be terminated if they exceed memory or cpu requirements and otherwise reports running jobs back to the database. The termination of jobs by the watchdog ensures that failed job status are reported back with maximum transparency to the exact reason of their failure. To this extent, the database utilizes a combination of major and minor status. The major status is mapped to one or more of the status that are provided by the batch system while the minor status is set at application level.

Once submitted, each batch job continuously reports its status to the database through outgoing http requests. To that end, computing nodes need to allow for outgoing internet access. Each batch job implements a workflow where in- and output data transfers are being performed (and their return codes are reported) as well as the actual running of the payload of a job (which is defined in the metadata description of the job). Dependencies on productions are implemented at the framework level and jobs are only submitted once dependencies are satisfied. While log files are stored on the computing farm, an external rsync script ensures permanent storage on a server in Geneva.

This framework is currently under test on the Geneva computing farm and at CNAF.

3. CNAF contribution

The CNAF computing center is the mirror of DAMPE data outside China and will be the main data center for Monte Carlo production. To keep available DAMPE data to the European DAMPE Collaboration, DAMPE data are transferred from PMO to CNAF using the gridFTP [6] protocol. Every time a new 1B, 1F or 2A data files are available at PMO, they are copied to a server at CNAF, gridftp-plain-virgo.cr.cnaf.infn.it, to the DAMPE storage area. The connection to China is passing through the Orientplus [7] link of the Géant Consortium [8]. The data transfer rate is currently limited by the connection of the PMO to the China Education and Research Network (CERNET), that has a bandwidth of 100 Mb/s.

On the user interface at CNAF, every hour a copy of each stream is triggered to the Geneva computing farm by means of lsf jobs. Dedicated lsf jobs are submitted once per day to asynchronously verify the checksum of new transferred data from PMO to CNAF and from CNAF to Geneva. In addition we foresee to have a set of cron jobs running on the user interface which will be a trusted host to connect to the web server in Geneva.

4. Activities in 2015

DAMPE has been launched on December 17$^{\text{th}}$ 2015, so only some implementation test have been performed in the last days of 2015. The implementation of all the data transferring pipeline has been then completed in the first months of 2016. Currently, data are copied daily at CNAF from PMO and a test of the Monte Carlo production framework is on-going. From 2016, DAMPE is officially supported by the CNAF center with 380 HS06 and 24 TB of storage disk.

5. Outlook for 2016

Due to the late launch of DAMPE, data transfer, analysis and MC production activities started only in Q1/2016. During the first three months critical tools have been implemented and it is now possible to fully exploit the CNAF in order to significantly contribute to DAMPE science. Over the next few months we will have a clearer picture on the need for resources (CPU and storage) and file an updated request accordingly.

6. Acknowledgments

We thank the DAMPE groups of the Purple Mountain Observatory and of the Geneva University for their collaboration in the definition and the implementation of the DAMPE Computing Model.

References

[1] Antcheva I. *et al.* 2009 *Computer Physics Communications* **180** 12, 2499 - 2512, https://root.cern.ch/guides/reference-guide.
[2] https://www.mongodb.org
[3] http://flask.pocoo.org
[4] Dubois R. 2009 *ASP Conference Series* **411** 189
[5] Tsaregorodtsev A. et al. 2008 *Journal of Physics: Conference Series* **119** 062048
[6] Allcock, W.; Bresnahan, J.; Kettimuthu, R.; Link, M. (2005). "The Globus Striped GridFTP Framework and Server". ACM/IEEE SC 2005 Conference (SC'05). p. 54. doi:10.1109/SC.2005.72. ISBN 1-59593-061-2. http://www.globus.org/toolkit/docs/latest-stable/gridftp/
[7] http://www.orientplus.eu
[8] http://www.geant.org
[9] http://www.cernet.edu.cn/HomePage/english/index.shtml

The EEE Project activity at CNAF

C. Aiftimiei, E. Fattibene, A. Ferraro, B. Martelli, D. Michelotto, F. Noferini, M. Panella, V. Sapunenko, C. Vistoli, S. Zani

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

E-mail: francesco.noferini@cnaf.infn.it

Abstract. The Extreme Energy Event (EEE) experiment is devoted to the search of high energy cosmic rays through a network of telescopes installed in about fifty high schools distributed throughout the Italian territory. INFN-CNAF hosts the data management infrastructure that receives data registered in stations very far from each other and allows a coordinated analysis. This infrastructure runs on the INFN-CNAF Cloud facility based on the OpenStack open-source Cloud framework, that provides Infrastructure as a Service (IaaS) for its users. During 2015 EEE used it for collecting, monitoring and reconstructing the data acquired in all the EEE stations. The synchronization between the stations and the INFN-CNAF infrastructure is performed through BitTorrent Sync, a free peer-to-peer software designed to optimize data syncronization between distributed nodes. All data folders are syncronized with the central repository in real time to allow an immediate reconstruction of the data and their publication in a monitoring webpage. In this paper we describe the system architecture and the data workflow, such as trasfer, reconstruction, monitoring and analysis.

1. Introduction

The EEE experiment setup is very peculiar and requires an *ad hoc* solution for data management. The CNAF Cloud facility (Cloud@CNAF) provides a flexible environment based on OpenStack [1] open-source Cloud framework, offering Infrastructure as a Service (IaaS), such as on-demand resources that can be adapted to the users need. This facility has been designated to host the EEE data management infrastructure, that aims to collect data from the telescopes which are distributed in a wide territory and provide reconstruction, monitoring and analysis services. In the CNAF cloud infrastructure a project (tenant) was provided to deploy all the virtual services requested by the EEE experiment.

2. Data Trasnfers

After a pilot run in 2014, the EEE project performed a first global run, Run-1, involving 35 schools in a coordinated data aquisition. During the run[1] all the schools were connected/authenticated at CNAF in order to transfer data using a BitTorrent technology. To realize this goal a btsync client (Win OS) is installed in each school and a front-end at CNAF is dedicated to receive all the data with a total required bandwidth of 300 kB/s, to collect the expected 5–10 TB per year. All the data collected are considered as custodial and for this reason they are stored also on tape. In Fig. 1 the general architecture for the EEE data flow is reported.

[1] Pilot run from 27-10-2014 to 14-11-2014 and Run-1 from 02-03-2015 to 30-04-2015.

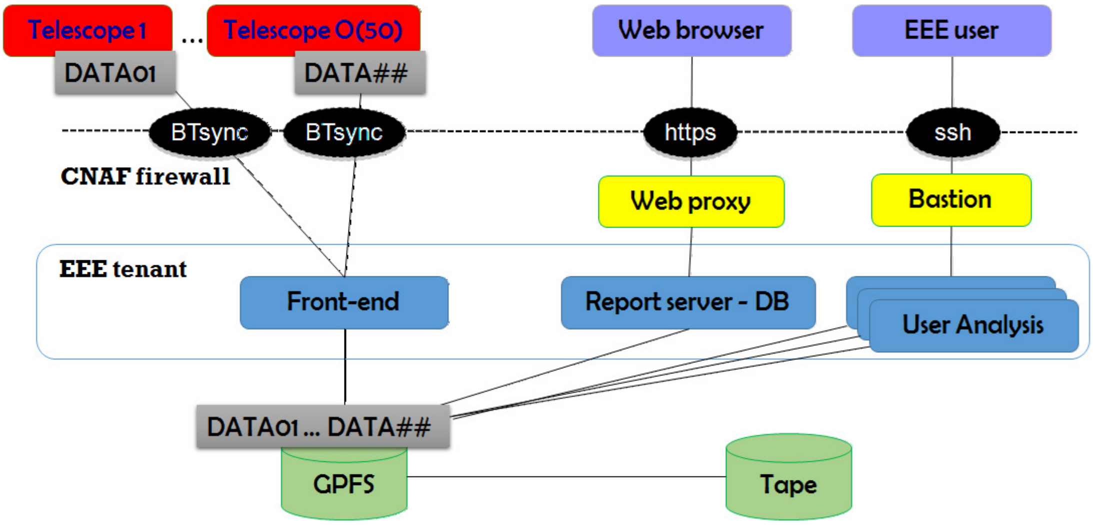


Figure 1. Architecture of the EEE tenant at CNAF.

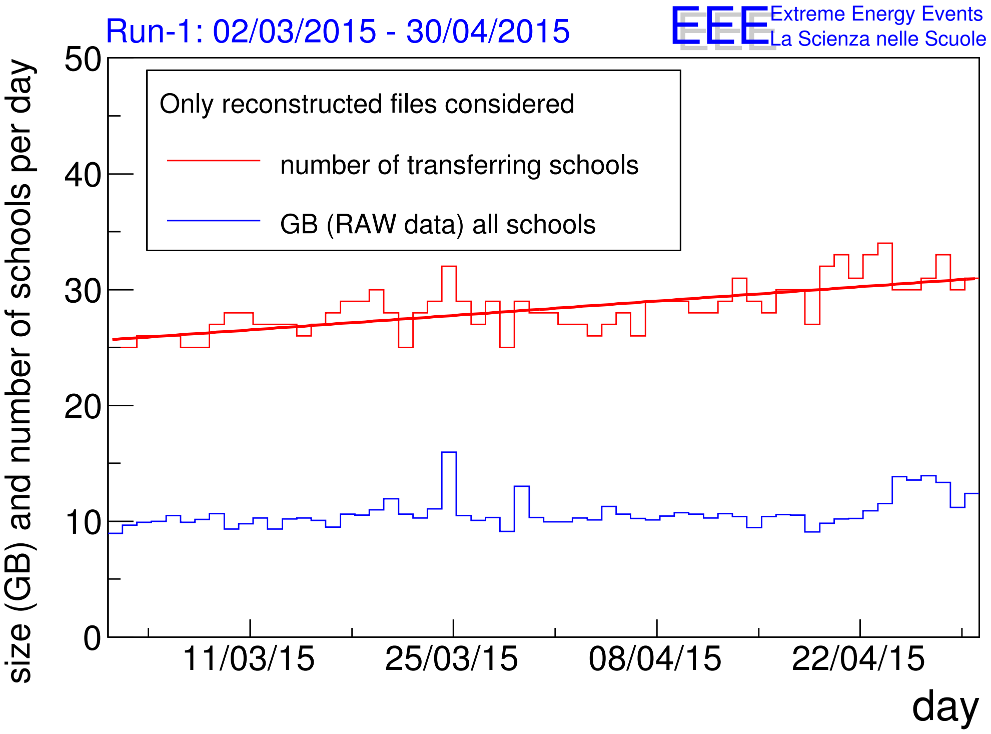
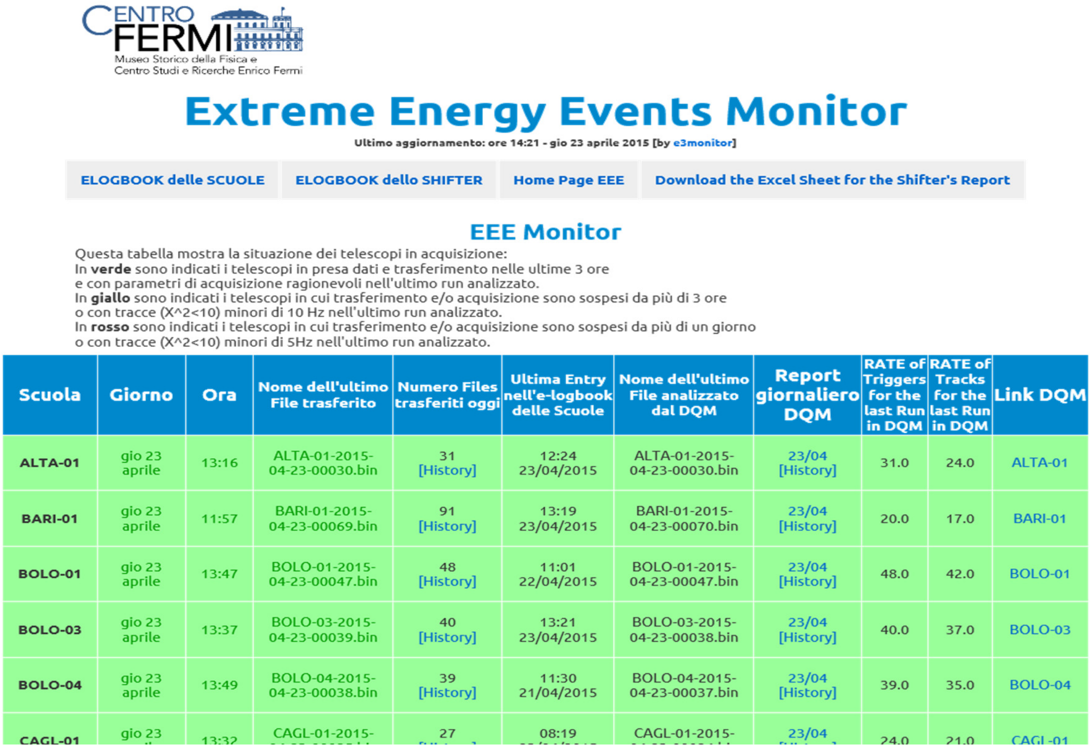


Figure 2. Statistics for the EEE Run-1 in 2015. For each day the number of schools transferring data and the amount of data collected at CNAF (in GB) are reported.

In the period including the pilot run and Run-1 we collected about 7 billion cosmic rays, corresponding to 2 TB data transferred at CNAF. In the same period also 3 TB from past years were also transferred. In Fig. 2 a summary of the data flow performances during Run-1 is reported.



CENTRO FERMI
Museo Storico della Fisica e
Centro Studi e Ricerche Enrico Fermi

Extreme Energy Events Monitor

Ultimo aggiornamento: ore 14:21 - gio 23 aprile 2015 [by e3monitor]

EEE Monitor

Questa tabella mostra la situazione dei telescopi in acquisizione:
In **verde** sono indicati i telescopi in presa dati e trasferimento nelle ultime 3 ore
e con parametri di acquisizione ragionevoli nell'ultimo run analizzato.
In **giallo** sono indicati i telescopi in cui trasferimento e/o acquisizione sono sospesi da più di 3 ore
o con tracce (X^2<10) minori di 10 Hz nell'ultimo run analizzato.
In **rosso** sono indicati i telescopi in cui trasferimento e/o acquisizione sono sospesi da più di un giorno
o con tracce (X^2<10) minori di 5Hz nell'ultimo run analizzato.

Scuola	Giorno	Ora	Nome dell'ultimo File trasferito	Numero Files trasferiti oggi	Ultima Entry nell'e-logbook delle Scuole	Nome dell'ultimo File analizzato dal DQM	Report giornaliero DQM	RATE of Triggers for the last Run in DQM	RATE of Tracks for the last Run in DQM	Link DQM
ALTA-01	gio 23 aprile	13:16	ALTA-01-2015-04-23-00030.bin	31 [History]	12:24 23/04/2015	ALTA-01-2015-04-23-00030.bin	23/04 [History]	31.0	24.0	ALTA-01
BARI-01	gio 23 aprile	11:57	BARI-01-2015-04-23-00069.bin	91 [History]	13:19 23/04/2015	BARI-01-2015-04-23-00070.bin	23/04 [History]	20.0	17.0	BARI-01
BOLO-01	gio 23 aprile	13:47	BOLO-01-2015-04-23-00047.bin	48 [History]	11:01 22/04/2015	BOLO-01-2015-04-23-00047.bin	23/04 [History]	48.0	42.0	BOLO-01
BOLO-03	gio 23 aprile	13:37	BOLO-03-2015-04-23-00039.bin	40 [History]	13:21 23/04/2015	BOLO-03-2015-04-23-00038.bin	23/04 [History]	40.0	37.0	BOLO-03
BOLO-04	gio 23 aprile	13:49	BOLO-04-2015-04-23-00038.bin	39 [History]	11:30 21/04/2015	BOLO-04-2015-04-23-00037.bin	23/04 [History]	39.0	35.0	BOLO-04
CAGL-01	gio 23	13:32	CAGL-01-2015-	27	08:19	CAGL-01-2015-	23/04	24.0	21.0	CAGL-01

Figure 3. A screenshot of the EEE monitor page. Data Quality Mointor (DQM) plots are provided in real time as well the status of the connection of each school.

3. Data Reconstruction/Monitor/Analysis

The chain to reconstruct data at CNAF is fully automated [3]. This point is really crucial because all the schools have to be monitored also remotely to act promptly in case of problems. This point is addressed throught automatic agents, running in a CNAF node dedicated to this issue, which are able to identify the arrival of a new file and then to trigger the reconstruction. A MySql database is deployed to trace all the actions performed on each single file (run) and the main parameters resulting from the reconstruction. Once the run is reconstructed a DST (Data Summary Tape) output is created and some quality plots are made available and published in the web page devoted to monitoring [4] (Fig. 3).

On parallel, a cluster of analysis nodes is reserved to EEE users via virtual nodes constructed on a dedicate image of the Operating System selected for the experiment (SL6). The EEE users authenticated at CNAF can access data (both RAW and DST files) via a GPFS filesystem as well the software of the experiment. The analysis activity [5] at CNAF resources is currently focused on several items like coincidences searches (two-three-many stations), rate vs. time (rate monitor+pressure correction), East-West asymmetry, cosmic ray anisotropy, upward going particles and the observation of the moon shadow.

4. Conclusion

From 2014 the EEE experiment entered in the phase of a coordinated activity between its telescopes. Such a step is realized with the creation of a data collector center at CNAF which at the same time provide the resources needed for the user analsyis. The centralization of the EEE activities gave a big boost both in the scientific program and in the participation of the high schools students. This "joint venture" between EEE and CNAF is still young and it will increase in the next months with the development of other services which are currently under study. In the future, CNAF staff planned to provide an Infrastructure-as-a-Service to EEE users to make the access to the resources even more flexible according to the cloud paradigm (user will be able to instantiate VMs on demand for the analyses) and to submit jobs to a dedicated LSF queue of CNAF "Tier1" data center. Several solutions to release the most relevant data using consolidated OpenData frameworks are under investigation (CKAN, OpenDataKit, etc.). Easy-to-use access mechanism to CNAF data will be deployed (Webdav/Swift/Torrent) as well as a structure of data replicated and stored among schools, exploiting torrents benefits.

References

[1] OpenStack, *http://www.openstack.org/*.

[2] BitTorrent Sync, *https://www.getsync.com/intl/it/*.

[3] F. Noferini, The computing and data infrastructure to interconnect EEE stations, Nucl. Inst. & Meth. A (2015), doi:10.1016/j.nima.2015.10.069

[4] INFN-CNAF, EEE monitor, *https://www.centrofermi.it/monitor/*.

[5] M. Abbrescia et al., The EEE Project: Cosmic rays, multigap resistive plate chambers and high school students, JINST 7 (2012) 11011.

The FAMU experiment: measurement of the proton Zemach radius

Emiliano Mocchiutti on behalf of the FAMU Collaboration

National Institute for Nuclear Physics (INFN), Sezione di Trieste, via A. Valerio 2, 34127 Trieste, Italy

E-mail: `Emiliano.Mocchiutti@ts.infn.it`

Abstract. The FAMU experiment main goal is the measurement of the proton Zemach radius using muonic hydrogen. In order to extract the Zemach radius, the FAMU collaboration aims at measuring the hyperfine splitting of the μp ground state, since the effect of the proton finite size affects the hyperfine transition energy. The proposed experimental method requires a detection system which is suited for time resolved X-ray spectroscopy. Signals recorded using HPGe and scintillating crystals is recorded up to 20 μs using a 500MHz digitizer to measure both the energy and the time spectrum of the detected events. Data are processed online at the Rutherford Appleton Laboratory for a quicklook analysis and immediately transferred to CNAF for storage and the subsequent full data analysis.

1. The proton radius puzzle

Protons are the most common particles in the universe and one of the building blocks of ordinary matter. Since their discovery, in 1919 by Ernest Rutherford, the proton properties have been widely studied. Until the first decade of the 2000, the electromagnetic structure of the proton was studied with very high accuracy by elastic scattering of electrons and positrons on hydrogen nuclei and by accurate measurements of the Lamb shift in the hydrogen atom spectrum. The latter type of experiments permit to take into account contributions like recoil corrections and the effects due to the nuclear structure, like its finite size. The accuracy level needed to measure the nuclear size, that is the root mean square of the charge distribution, was reached in the 1990s by the Lamb shift experiments and it was in excellent agreement with the same result obtained by scattering experiments with an average data value of $r_p = 0.8751(61)$ fm [1].

In order to increase the accuracy of the this value, a measurement of the Lamb shift in the 2S–2P transition of muonic hydrogen (μp), i.e. the bound state formed by a proton and a negative muon, was proposed. In fact the muon, being about 200 times heavier than the electron is also 200 times closer to the proton, and thus much more sensitive to its structure. The results of this measurement were published in 2010 and provided a value of the proton charge radius $r_p = 0.84184(67)$ fm that was about an order of magnitude more accurate than previous measurements but totally inconsistent [2]. This result was confirmed by other measurements published in 2013 [3], but nobody has been able to explain the discrepancy between "electronic" and "muonic" measurements and all the hypotheses are still on the table, ranging from experimental errors to effects of unconsidered contributions in the calculations or to hints for new physics [4].

2. The FAMU experiment

The FAMU experiment goal is the measurement of the Zemach radius of the proton (R_p), convolution of the charge and magnetic moment density [5], which can be extracted performing a precise measurement of the hyperfine splitting of the μp ground state [6, 7]. This quantity has been already measured using ordinary hydrogen, and a comparison with the value extracted from muonic hydrogen may either reinforce or delimit the proton radius puzzle.

The apparatus consists of a gas target filled with a mixture of hydrogen and heavier gasses surrounded by the detectors. In the 2015 set-up, show in Fig. 1, in front of the cryogenic gas target a hodoscope was used to measure the shape and timing of the muon beam, four HPGe and nine $LaBr_3$(Ce) detectors were used to identify X-rays coming from the de-excitation of muonic atoms. Six other detectors (CeCAAG and PrLuAg crystals) were placed below the apparatus to study their response in this experimental environment. The experiment takes place at the Rutherford Appleton Laboratories (UK) where the RIKEN–RAL muon complex [8] is located, the only facility in the world able to provide the pulsed muon beam which is needed for the experiment.

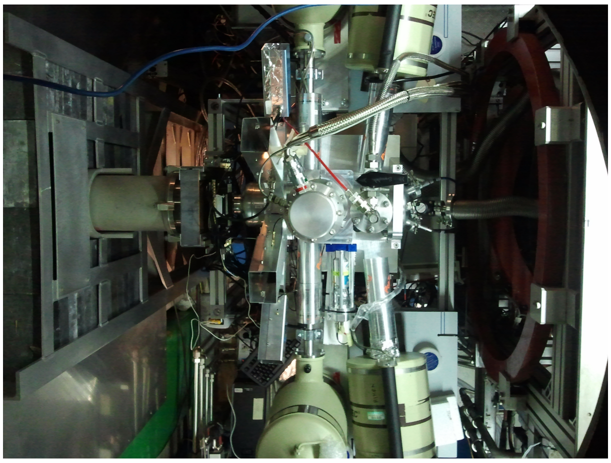


Figure 1. Top view of the experimental set-up in operations at RIKEN-RAL in 2015. The beam pipe can be seen on the left, while the gas target is on the center surrounded by the detectors.

3. Data taking in 2014 and 2015

The FAMU project foresees a progressive approach to the final measurement of the 1S state hyperfine transition on the muonic hydrogen atom. In the first phase a preliminary experimental layout is set up to perform measurements of the collision energy dependence of the μ^- transfer rate in various gases. In particular, the objective of the first phase of the experiment is a complete study of the muon transfer to gases for which there are experimental evidences or theoretical hints for a pronounced energy dependence of the transfer rate in the epithermal range. The results will set the required firm ground for all following activities and will be used to determine the optimal temperature, pressure, laser shot timing and the chemical composition (admixture gas and its concentration) of the gas target for the measurement of the HFS in muonic hydrogen.

A first test of the apparatus was performed in the summer of 2014 and studied the detector response in environment of the muon beam at RIKEN-RAL through the measurement the muon transfer rate at room temperature [9].

In december 2015 a first data acquisition with a cryogenic target was taken in order to study the temperature and gas dependence of the muon transfer rate. Data analysis is currently in progress.

4. FAMU computing model

Signals from all detectors (except the hodoscope) were digitized at 500 MHz, using a CAEN D5730 digitizer; the 64 hodoscope output channels were read by two CAEN V792 QADC. Data were recorded real time event by event using a trigger signal. The trigger signal was given by the beam line and the acquisition started about 250 ns before the time at which the center of the first muon pulse reached the target. Then the digitizer samples, for each detector, the output signal every 2 ns in a chosen time window from 5 to 20 μs depending on the detector.

For each trigger the syncronized output of all channels from digitizer and QDC was recorded on a local disk. Sets of data from 15000 trigger events were acquired as a single run and stored in n-tuples. During seven days of experimental operations about 1300 runs have been recorded, for a total of more than 10^7 triggered events.

Data are temporary stored and processed on site at RAL in order to check online the status of the detectors and the housekeeping informations.

Real time the raw files are also copied to CNAF via GridFTP.

5. FAMU at CNAF

The CNAF data center is used by FAMU as main data repository and processing site.

Data at CNAF are first converted from n-tuples to rootples. Then the first task of the analysis is to identify the peaks due to the characteristic X-rays of the muonic atoms in the energy spectrum obtained by all the detectors. Wave forms are analyzed by mean of a C++ program suite based on ROOT [10] classes. Time of arrival and energy are reconstructed for each photon peak by fitting the pulse; in case of pile-up a multiple fitting procedure is applied to correctly determine the desired parameters.

Processed data are saved into new root files that are used for detector calibration and data analysis which is performed at CNAF and at local computing resources in institutes and universities.

CNAF is also used as Monte Carlo production site by the FAMU collaboration. In 2015 simulations were run in order to determine the best target configuration for the muon transfer rate data acquisition system.

6. Conclusions

CNAF plays a major role in the computing of the FAMU experiment. Both the storage and the computing resources needed by the experiment are provided by this facility.

Acknowledgments

The research activity presented in this paper has been carried out in the framework of the FAMU experiment funded by Istituto Nazionale di Fisica Nucleare (INFN). The use of the low energy muons beam has been allowed by the RIKEN RAL Muon Facility. We thanks the RAL staff for the help and precious collaboration in the set up of the experiment at RIKEN-RAL port 4.

We gratefully recognize the help of T. Schneider, CERN EP division, for his help in the optical cutting of the scintillating fibers of the hodoscope detector and the linked problematics and N. Serra from Advansid srl for useful discussions on SiPM problematics.

[1] Mohr P J, Newell D B and Taylor B N 2015 Codata recommended values of the fundamental physical constants: 2014 URL http://dx.doi.org/10.5281/zenodo.22826

[2] Pohl R et al 2010 *Nature* **466**

[3] Antognini A et al 2013 *Science* **339**

[4] Pohl R, Gilman R, Miller G A and Pachucki K 2013 *Ann. Rev. Nucl. Part.* **S. 63**

[5] Zemach A C 1956 *Phys. Rev.* **104**

[6] Bakalov D, Milotti E, Rizzo C, Vacchi A and Zavattini E 1993 *Phys. Lett. A* **172**

[7] Bakalov D, Adamczak A, Stoychev L and Vacchi A 2012 *Nucl. Instr. Meth. Phys. Res. B* **281**

[8] Matsuzaki T, Ishida K, Nagamine K, Watanabe I, Eaton G and Williams W 2001 *Nucl. Inst. Meth. Phys. Res. A* **465**

[9] Adamczak A et al 2016 *accepted for publication on J. Inst.*

[10] Brun R and Rademakers F 1996 *Nucl. Inst. Meth. Phys. Res. A* **389**

The *Fermi*-LAT experiment at the INFN CNAF Tier 1

M Kuss[1], F Longo[2], S Viscapi[3] and S Zimmer[4], on behalf of the *Fermi* LAT collaboration

[1] Istituto Nazionale di Fisica Nucleare, Sezione di Pisa, I-56127 Pisa, Italy
[2] Department of Physics, University of Trieste, via Valerio 2, Trieste and INFN, Sezione di Trieste, via Valerio 2, Trieste, Italy
[3] Laboratoire Univers et Particules, Université de Montpellier II Place Eugène Bataillon - CC 72, CNRS/IN2P3, F-34095 Montpellier, France
[4] Departement de Physique Nucléaire et Corpusculaire, Université de Genève, 24 Quai Ernest Ansermet, CH-1211 Geneva, Switzerland

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..

So far, the LAT has registered 460 billion trigger (1800 Hz average trigger rate). An on-board filter analyses the event topology and discards about 80%. Of the 92 billion events that were transferred to ground 863 million were classified as photons. All photon data are made public almost immediately. Downlink, processing, preparation and storage take about 24 hours.

2. Scientific Achievements Published in 2015

In 2015, 62 collaboration papers (Cat. I and II) were published, keeping the pace of about 60 per year since launch. Independent publications by LAT collaboration members (Cat. III) amount to 42. Also external scientists are able to analyse the *Fermi* public data, resulting in 229 external publications.

In 2015, 3 Fermi papers triggered NASA press releases: the first detection of a pulsar located in another galaxy (the Large Magellanic Cloud) in gamma rays [2, 3], the discovery of periodic gamma-ray flux variations in an active galaxy [4, 5], and the implementation of a new improved analysis algorithm (called Pass 8) [6].

While the press release for Pass 8 was on 7 January 2016 (thus not in 2015), the work started already in January 2009, when analyzing the first few months of on-orbit data. By mid of 2015, the standard analysis pipeline switched to Pass 8, and all data taken since the launch was reprocessed. The improvement of Pass 8 over the previous Pass 7 code can already been seen in the number of photons reconstructed: at the end of 2014 (6.5 years of data) 400 million Pass 7 photons [7] were in the public data base, but now, with 7.5 years of data and Pass 8, they amount to 863 million, thus doubling the number of photons on essentially the same data set. In particular, Pass 8 now also allows for a reliable reconstruction of photons with energies below 100 MeV. Also the high energy end gained from an improved reconstruction of the original photon direction, as well as from a better energy reconstruction, extending it up to a few TeV. The paper detailing the methods used in Pass 8 and its performance is still in preparation. However, many sources are being restudied based on Pass 8 data, with some science results published already [8, 9, 10, 11], and many more in preparation.

3. The Computing Model

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 [12] 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). The available resources are used for specific tasks: the SLAC batch farm for data processing, high level analysis, and smaller MC tasks, the batch farm of the CC-IN2P3 at Lyon and the grid resources for large MC campaigns. The grid resources [13] are accessed through a DIRAC (Distributed Infrastructure with Remote Agent Control) [14] interface to the LAT data pipeline [15]. This setup is in production mode since April 2014.

The jobs submitted through DIRAC (c.f. Fig. 1) constitute a substantial fraction of the submitted jobs, with CNAF-T1 being the fourth largest contributor (and marginally below resources in Pisa). However, we also exploit the possibility to submit jobs directly using the grid middleware. Figure 2 shows the usage of grid resources in 2015. About 11% of the jobs were run at the INFN Tier 1 at CNAF as shown by Fig. 3. The total usage in 2015 was 3317 HS06. Assuming 10 HS06 per core, this is equivalent to about 330 CPU-years.

4. Conclusions and Perspectives

The prototype setup based on the DIRAC framework described in the INFN-CNAF Annual Report 2013 [16] proved to be successful. In 2014 we transitioned into production mode, but since mid of 2015 we have been suffering from disk (SE) problems. Initially, a failure in our job pipeline caused temporary files not to be removed from SEs after they were transferred to a permanent storage space at SLAC. This issue has since been mitigated and while the reliability of the system has improved overall, we continue experiencing issues, in particular related to the DIRAC server and the involved SEs.

References

[1] Atwood W B et al. 2009 *Astrophysical Journal* **697** 1071

[2] NASA press release 2015 November 12, http://www.nasa.gov/feature/goddard/nasas-fermi-satellite-detects-first-gamma-ray-pulsar-in-another-galaxy/

[3] Ackermann M et al. 2015 *Science* **350***, 801*

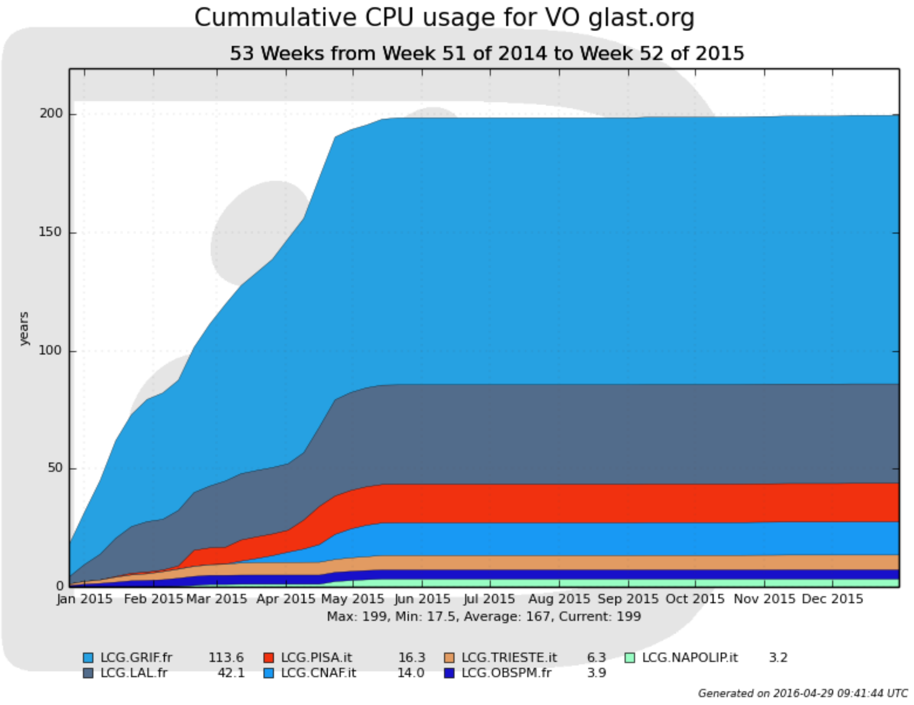


Figure 1. DIRAC usage plot for VO glast.org in 2015

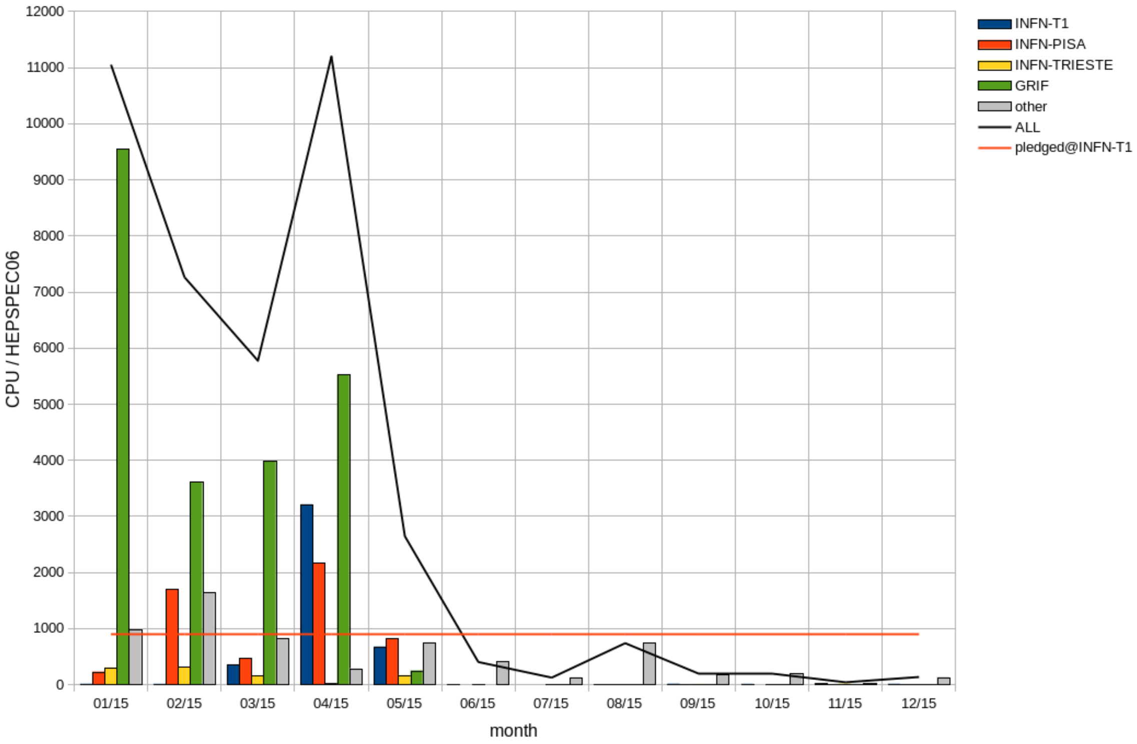


Figure 2. Usage of grid sites by the VO glast.org in 2015

[4] NASA press release 2015 November 13, http://www.nasa.gov/feature/goddard/nasas-fermi-mission-finds-hints-of-gamma-ray-cycle-in-an-active-galaxy/

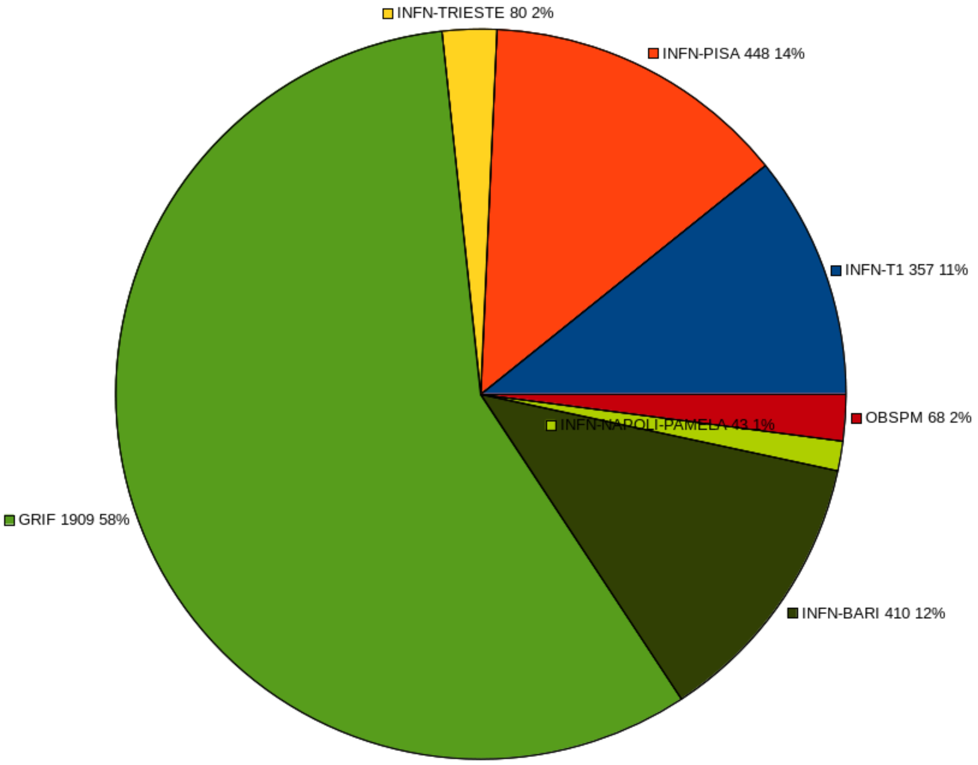


Figure 3. Usage (in HEP-SPEC06) of grid sites by the VO glast.org in 2015

[5] Ackermann M et al. 2015 *Astrophysical Journal Letters* **813** *no.2, 41*

[6] NASA press release 2016 January 7, http://www.nasa.gov/feature/goddard/2016/nasas-fermi-space-telescope-sharpens-its-high-energy-vision/

[7] Kuss M et al. 2014 *INFN-CNAF Annual Report 2014*, edited by L. dell'Agnello, F. Giacomini, and L. Morganti, pp. 54

[8] Ackermann M et al. 2015 *Physical Review Letters* **115** *no.23, 231301*

[9] Ackermann M et al. 2015 *Physical Review* **D91** *no.12, 122002*

[10] Ajello M et al. 2016 *Astrophysical Journal* **819** *no.2, 98*

[11] Ackermann M et al. 2016 *Astrophys.J.* **819** *no.2, 149*

[12] Dubois R 2009 *ASP Conference Series* **411** 189

[13] Arrabito L et al. 2013 CHEP 2013 conference proceedings arXiv:1403.7221

[14] Tsaregorodtsev A et al. 2008 *Journal of Physics: Conference Series* **119** 062048

[15] Zimmer S et al. 2012 *Journal of Physics: Conference Series* **396** 032121

[16] Arrabito L et al. 2014 *INFN-CNAF Annual Report 2013*, edited by L. dell'Agnello, F. Giacomini, and C. Grandi, pp. 46

[17] Atwood W B et al. 2013 *2012 Fermi Symposium: eConf Proceedings C121028* arXiv:1303.3514

The GERDA experiment

E. Medinaceli on behalf of the Gerda collaboration

Università di Padova, via Marzolo 8, 35100 Padova, Italy

E-mail: medinaceli@pd.infn.it

Abstract. The GERmanium Detector Array (Gerda) experiment searches for neutrino-less double beta ($0\nu\beta\beta$) of ^{76}Ge. It is located at the INFN's Gran Sasso National underground Laboratory with an overburden of 3500 m.w.e.. Gerda uses high purity Germanium detectors enriched in the isotope ^{76}Ge, which are deployed bare in a liquid Argon (LAr) cryostat. The cryostat is embedded inside a water tank that serves as active water Cerenkov muon veto and as a passive gamma and neutron shield. The signature searched for is a monochromatic line at $Q_{0\nu\beta\beta} = 2039\,\mathrm{keV}$.

Gerda Phase I was concluded in 2013, after two years of data taking, and since December 2015 Gerda has been run with an increased active mass (20 kg added) in a second experimental Phase. With the help of a new instrumentation of the LAr and an increased Pulse Shape Discrimination (PSD) capability of the additional detectors, the aim is to reduce the background index by one order of magnitude with respect to Phase I; arriving at $\sim 10^{-3}\,\mathrm{cts/(keV\,kg\,yr)}$ in order to reach a sensitivity of $T^{0\nu}_{1/2} \sim 10^{26}$yr with an exposure of $\sim 100\,\mathrm{kg\,yr}$.

In the following a short description of the Gerda experimental setup is given, and the background modeling and other activities related to the data analysis including Monte Carlo simulations performed at CNAF are presented.

1. The Gerda experiment and status

The GERmanium Detector Array (Gerda) experiment is searching for neutrino-less double beta decay ($0\nu\beta\beta$) of ^{76}Ge. This process would violate lepton number conservation by two units, and is possible only if neutrinos have a Majorana mass component, thus, being their own anti-particles. $0\nu\beta\beta$ decay is predicted by a large number of extensions of the Standard Model of particle physics.

With a decay half-live exceeding 10^{25} yr $0\nu\beta\beta$ decay, if existing, would be a rare phenomenon requiring an ultra-low background in order to reach an adequate experimental sensitivity. The decay half-life is connected to the effective Majorana neutrino mass and can give a handle on the absolute neutrino mass scale.

Gerda uses an array of high-purity germanium (HPGe) detectors, which are enriched to $> 86\%$ with the $0\nu\beta\beta$ candidate isotope ^{76}Ge. They serve as detector and source simultaneously, thus, maximizing the detection efficiency.The concept design consists in the operation of bare HPGe detectors in a $64\,\mathrm{m}^3$ liquid Argon (LAr) cryostat which is surrounded by an ultra-pure water shield to suppress external backgrounds [1, 2]. The water tank is instrumented with photo-multipliers and Cerenkov light emission is used to detect cosmic ray muons [1].

During Gerda Phase I, the LAr volume was only used as a passive shield; whereas in Phase II, the LAr was equipped with a hybrid instrumentation of photo-multipliers and light-guiding fibers coupled to Silicon photo-multipliers (SiPMs). Tagging the scintillation light from LAr it is possible to suppress background emitted in the vicinity or on the surface of the detectors.

Furthermore, Gerda Phase II implements 40 detectors (about 20 kg increase of detector mass [6]) of which 30 have enhanced pulse shape discrimination properties [3] which are used for further background suppression.

The searched for signal is a monochromatic line at the end-point of ^{76}Ge double beta decay ($2\nu\beta\beta$) at an energy of 2039 keV. HPGe detectors, being made from a semiconductor, have intrinsically an excellent energy resolution which is $\sim 0.1\%@Q_{\beta\beta}$.

Between November 2011 and December 2013, Gerda Phase I collected a total of 21.6 kg yr of exposure [4]. A background model was developed describing the observed energy spectrum outside a blinded region of ± 40 keV around $Q_{\beta\beta}$. It takes several contributions into account on the basis of expectations from material screening and characteristic structures in the energy spectrum which are compared to Monte Carlo (MC) simulation. The model was used to predict the level and the spectral shape of the background in the region of interest (ROI) around $Q_{\beta\beta}$, before unblinding. A background index of 10^{-2}cts/(keV kg yr) was predicted for the ROI [5]. After unblinding, no signal was observed in the ROI and a lower limit of $T^{0\nu}_{1/2} > 2.1 \times 10^{25}$ yr (at 90% C.L.) was derived [4].

Since December 2015 Gerda has been running Phase II and aims to reach an experimental sensitivity of about $T^{0\nu}_{1/2} > 10^{26}$ yr by collecting an exposure of about 100 kr yr with a one order of magnitude lower background with respect to Phase I.

Also, a new background model is being developed taking the full Phase II detector array into account.

2. Gerda at CNAF

The Gerda analysis chain is based in a hierarchical structure of the data; where data is stored in different formats according to the level of complexity of the data, corresponding to different stages of the event reconstruction. Each dataset is stored in a subset called Tier.

Raw data (Tier0) coming from the DAQ is stored in a custom binary format and contains all the digital information (waveforms and digitizer data) coming from the detector. The next level (Tier1) contains the same information of the previous level but the region of interest around $Q_{\beta\beta}$ is removed (2039 ± 25 keV in Phase II), in order to perform a blind analysis. At this level data is translated to the Gerda data-format and is stored in Root files [8]; the change of data-format is done using the MGDO libraries [7]. This procedure allows to standardize data quality, and decouple the high-level analysis from the specific binary format of the raw data. Data at this level of the reconstruction is saved and stored in a public and accessible area. In the next reconstruction step, each waveform is analyzed individually using software packages from the Gelatio [9] framework. Custom analysis chains using different analysis modules can be used to obtain the required observables needed for the user analysis. The usage of common analysis modules guarantees uniformity at low level introducing common systematics in the reconstruction of the variables; custom analysis chains provide flexibility to the user. The output (Tier2) is stored also in a Root file, and is used for high-level analysis. In higher levels of the reconstruction, energy calibrations and pulse shape discrimination analysis are performed. Several energy reconstruction algorithms are provided in the Gelatio framework. Results are stored in Tier3 and Tier4 Root files. The final Gerda physics analysis is performed using Tier4 files.

Gerda uses the storage resources of CNAF, accesing the GRID through the virtual organization (VO) `gerda.mpg.de@lfc.italiangrid.it`. Currently the amount of data stored at CNAF is around 10 TB, mainly data from Phase I, which is saved and registered in the Gerda catalogue. To perform data analysis CNAF provides a dedicated machine $ui-gerda.cr.cnaf.infn.it$ to Gerda users. A complete suit of the Gerda analysis software is available in this machine. Git technology of subversioning is used to manage and keep updated the different development branches of the software, stored in a github repository. The SWMOD package (based on a set of custom scripts) regulates the different versions of the analysis software MGDO, MAGE [10], GELATIO, resolving all the dependencies with third-party applications *e.g.* Root, Geant4, CHLEP, among others; and sets the environmental variables. Each instance of the software is labeled and SWMOD provides the capability to load different instances of the same software with their dependencies with just one command. The complete data reconstruction pipeline of batch jobs is handled by the *Luigi* python module (*https://github.com/spotify/luigi*) [11], which offers a web interface to search and filter among all the tasks.

With Gerda being a low background experiment, the amount of data produced during physics runs is modest. Furthermore, the usage of Root files allows to achieve a data reduction of a factor of ~ 2. The rate of data taking in Phase I was about 4 GB/day (Tier0), plus around 20 GB/week for calibrations. About the same amount of data is being produced in the on-going Phase II. Data is stored at the experimental site in INFN's Gran Sasso National Laboratory (LNGS) cluster. The policy of the Gerda collaboration requires three backup copies of the raw data (Italy, Germany and Russia). CNAF is the Italian backup storage center, and raw data is transferred from LNGS to Bologna in specific campaigns of data exchange. Data is copied and registered in the Gerda catalogue at the VO.

The Gerda collaboration uses the CPU provided by CNAF to process higher level data reconstruction (TierX), and to perform dedicated user analysis. Besides, CPU is mainly used to perform MC simulations. A very important requirement of the Gerda experiment is to model the background as well as the $2\nu\beta\beta$ spectrum of ^{76}Ge, to be able to provide a background index in the ROI.

The background model identifies the contributions from different sources of background in the energy spectra. The Gerda MC simulations are done with MAGE [10] which is based on GEANT4 [12]. The background model includes intrinsic radioactive impurities of the detectors, structural materials and the surrounding environment.

The main contributions come from the natural radioactivity of the detector holders and high voltage and signal cables, being the main source of ^{228}Th, ^{226}Ra and ^{40}K; the LAr, introducing the ^{42}Ar radioactive decay chain; background from contaminations of the detector surfaces with nuclei from the ^{238}U, ^{232}Th and the ^{222}Rn decay chains; and the cosmogenically induced isotopes ^{68}Ge and ^{60}Co in the detector bulk material [5].

Figure 1 shows the Gerda Phase I background model, containing all individual background contributions as well as the $2\nu\beta\beta$ continuum; the plot shows the fit considering a data set labeled GOLD-Coax (17.9 kg year measured with enriched semi-coaxial detectors). This global model describing the background spectrum was obtained by fitting simulated spectra of different contributions to the measured spectrum using the Bayesian Analysis Toolkit (BAT) [13], details about the statistical procedure can be found in [5]. The same plot, at the lower part, shows the residuals of the data with respect to the MC global background model; the differences are

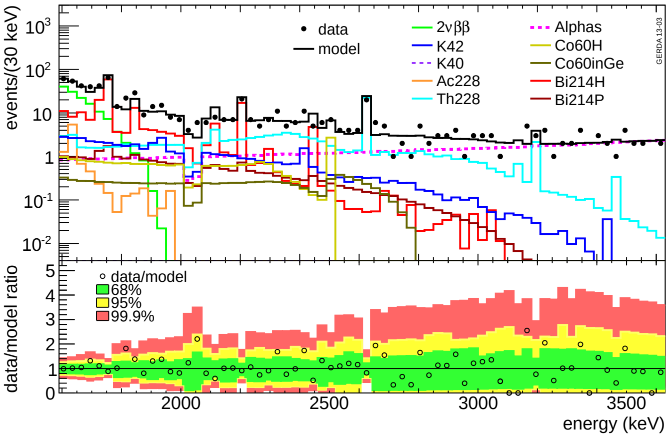


Figure 1. Background decomposition to the best fit minimum model of the GOLD-coax data set, shown with a black solid line. The contribution of different isotopes is also shown. The $2\nu\beta\beta$ contribution is drawn with a green solid line, and the experimental energy spectrum is represented with black dots. The lower panel in the plot shows the ratio between the data and the prediction of the best fit model together with the smallest intervals of 68% (green band), 95% (yellow band) and 99.9% (red band) probability for the ratio assuming the best fit parameters [5].

represented as the ratio of the data and the fit assuming the best set of parameters.

For Gerda Phase II the global model of the background is being evaluated using new simulations, considering the list of isotopes present in the detector array and the LAr. The simulation load is distributed among different computing centers of member institutions of the Gerda collaboration, among them CNAF. In particular, at CNAF the simulation of the decay chain of LAr considering the isotope ^{42}K is being carry on, taken into account the interaction at the surface of the detectors and inside the Mini-Shroud. Predictions for the $2\nu\beta\beta$ continuum are simulated and other internal detector bulk contaminations due to ^{68}Ge and ^{60}Co are being evaluated. Furthermore, global fits of the simulated backgrounds to experimental data, using BAT, are being performed at CNAF.

References

[1] K.-H. Ackermann *et al.* (Gerda Collaboration), Eur. Phys. J. C (2013) 73:2330
[2] M. Agostini *et al.* (Gerda Collaboration), Eur.Phys.J. C73 (2013)2330
[3] M. Agostini *et al.* (Gerda Collaboration), EPJ C 73(2013)2583.
[4] M. Agostini *et al.* (Gerda Collaboration), Phys. Rev. Lett. 111 (2013) 122503
[5] M. Agostini *et al.* (Gerda Collaboration), Eur. Phys. J. C (2014) 74:2764
[6] M. Agostini *et al.* (Gerda Collaboration), Eur. Phys. J. C (2015) 75:39
[7] M. Agostini *et al.*, J. of Phys.: Conf. Ser. 375 (2012) 042027
[8] R. Brun and F. Rademakers, Nucl. Inst. & Meth. in Phys. Res. A 389 (1997) 81
[9] M. Agostini *et al.*, J. of Phys.: Conf. Ser. 368 (2012) 012047.
[10] M. Boswell *et al.*, IEEE Trans. Nucl. Sci. 58 (2011) 1212
[11] *http://luigi.readthedocis.io/en/stable/*
[12] GEANT collaboration, S. Agostinelli *et al.*, Nucl. Inst. Methods A 506, 250(2003)
[13] A. Caldwell *et al.*, Comput. Phys. Atom. Nucl. 63, 1282(2000)

The KM3NeT neutrino telescope network and CNAF

Cristiano Bozza

Department of Physics of the University of Salerno and INFN Gruppo Collegato di Salerno,
via Giovanni Paolo II 132, 84084 Fisciano, Italy

E-mail: `cbozza@unisa.it`

Abstract. The KM3NeT Collaboration is building a new generation of neutrino detectors in the Mediterranean Sea. The scientific goal is twofold: with the ARCA programme, KM3NeT will be studying the flux of neutrinos from astrophysical sources; the ORCA programme is devoted to investigate the hierarchy of neutrino mass eigenstates. The unprecedented size of detectors will imply PByte-scale datasets and calls for large computing facilities and high performance data centres. The data management and processing challenges of KM3NeT are reviewed as well as the computing model. Specific attention is given to describing the role and contributions of CNAF.

1. Introduction

Water-Cherenkov neutrino telescopes have a recent history of great scientific success. Deep-sea installations provide naturally high-quality water and screening from cosmic rays from above. The KM3NeT Collaboration [1] aims at evolving this well-proven technology to reach two scientific goals in neutrino astronomy and particle physics, by two parallel and complementary research programmes [2, 3]. The first, named ARCA (Astroparticle Research with Cosmics in the Abyss), envisages to study the neutrino emission from potential sources of high-energy cosmic rays, like active galactic nuclei, supernova remnants and regions where high fluxes of gamma rays originate (including supernovae), and received a boost of interest after the IceCube report of diffuse neutrino flux at energies exceeding 1 PeV. The goals of directional identification of the source of high-energy neutrinos and good energy resolution require a detector with a total volume beyond 1 km^3. The second line of research is devoted to studying the hierarchy of neutrino mass eigenstates (Oscillation Research with Cosmics in the Abyss - ORCA). A detector technical identical to the ARCA one but with smaller spacing between sensitive elements will be used to detect atmospheric neutrinos oscillating while crossing the Earth volume: the modulation pattern of the oscillation is influenced by a term that is sensitive to the hierarchy (normal or inverted), hence allowing discrimination between the models. The KM3NeT detector technology originates from the experience of previous underwater Cherenkov detectors (like ANTARES and NEMO), but it takes a big step forward with the new design of the digital optical modules (DOM), using strongly directional 3-inch photomultiplier tubes to build up a large photocatode area. The whole detector is divided for management simplicity in *building blocks*, each made of 115 *Detection Units* (DU). A DU is in turn made of 18 DOM's, each hosting 31 photomultipliers (PMT). Hence, a building block will contain 64,170 PMT's. With an expected livetime of at least 15 years, and a single photoelectron rate of a few kHz per PMT, online, quasi-online and offline computing are challenging activities themselves. In addition, each detector installation will include instruments that will be dedicated to Earth and Sea Science (ESS), and will be operated by the KM3NeT Collaboration. The data flow from these instruments is negligible compared to optical data and is not explicitly accounted for in this document. The first phase of the ARCA detector is under construction and the first DU has been put in place in the Italian site in December 2015 and is taking data smoothly and steadily. More detection units are expected to be deployed soon.

2. Computing model

The computing model of KM3NeT is modelled on the LHC experiment strategy, i.e. it is a three-tier architecture, as depicted in Fig. 1.

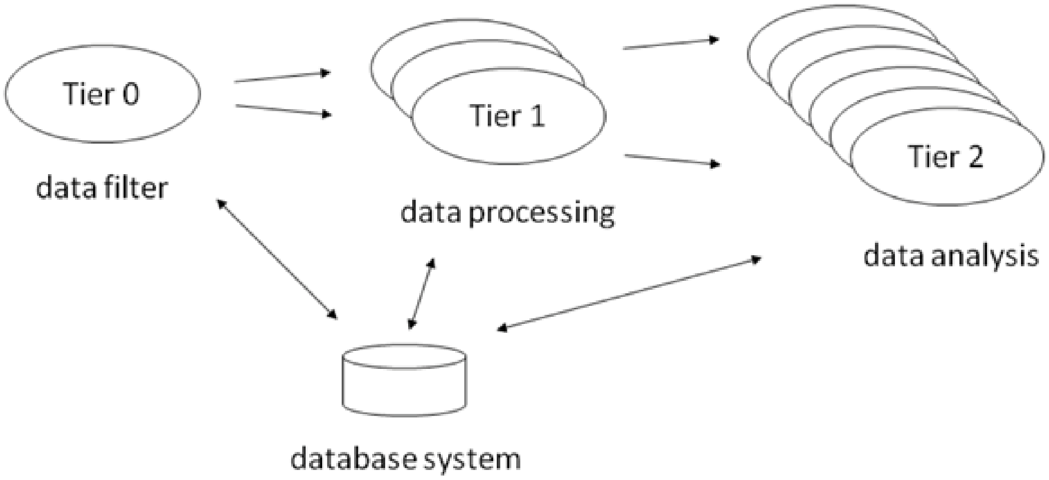


Figure 1. Three-tier model for KM3NeT computing.

With the detector on the deep seabed, all data are transferred to data acquisition (DAQ) control stations on the shore. Tier 0 is composed of a computer farm running triggering algorithms on the full raw data flow with a reduction from $5GB/s$ to $5MB/s$ per *building block*. Quasi-on-line reconstruction is performed for selected events (alerts, monitoring). The output data are temporarily stored on a persistent medium and distributed with fixed latency (typically less than few hours) to various computing centres, which altogether constitute Tier 1, where events are reconstructed by various fitting models (mostly searching for shower-like or track-like patterns). Reconstruction further reduces the data rate to about $1MB/s$ per *building block*. In addition, Tier 1 also takes care of continuous detector calibration, to optimise pointing accuracy (by working out the detector shape that changes in because of water currents) and photomultiplier operation. Local analysis centres, logically allocated in Tier 2 of the computing model, perform physics analysis tasks. A database system interconnects the three tiers by distributing detector structure, qualification and calibration data, run book-keeping information and slow-control and monitoring data.

KM3NeT exploits computing resources in several centres and in the GRID, as sketched in Fig. 2. The conceptually simple flow of the three-tier model is then realised by splitting the tasks of Tier 1 to different processing centres, also optimising the data flow and the network path. In particular, CNAF and CC-IN2P3 will be mirrors of each other, containing the full data set at any moment.

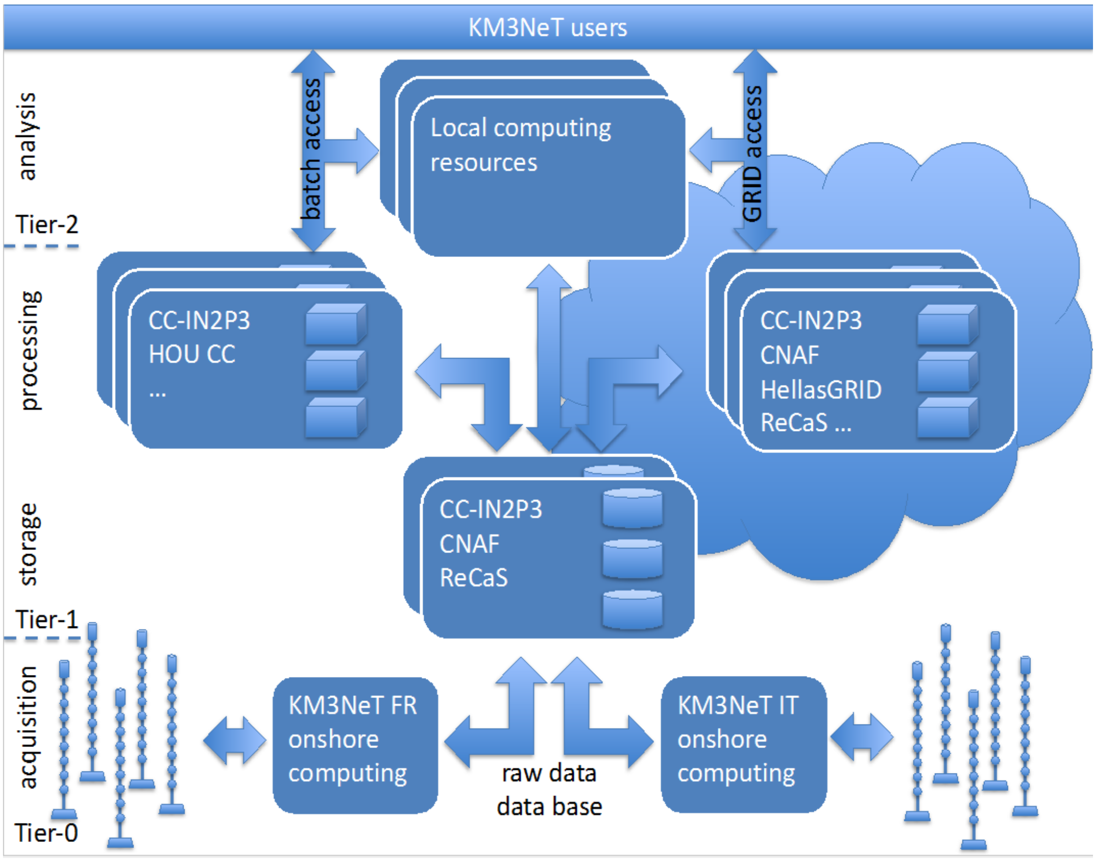


Figure 2. Data flow in KM3NeT computing model.

Table 1. Task allocation in Tier 1.

Computing Facility	Main Task	Access
CC-IN2P3	general processing, central storage	direct, batch, GRID
CNAF	central storage, simulation, reprocessing	GRID
ReCaS	general processing, simulation, interim storage	GRID
HellasGRID	reconstruction	GRID

3. Data size and CPU requirements

Calibration and reconstruction work in batches. The raw data related to the batch are transferred to the centre that is in charge of the processing before it starts. In addition, a rolling buffer of data is stored at each computing centre, e.g. the last year of data taking.

Simulation has special needs because the input is negligible, but the computing power required is very large compared to the needs of data-taking: indeed, for every year of detector livetime, KM3NeT envisages to produce 10 years of simulated events (with exceptions). Also, the output of the simulation has to be stored at several stages. While the total data size is dominated by real data, the size of reconstructed data is dictated mostly by the amount of simulated data.

Table 1 details how the tasks are allocated.

Thanks to the modular design of the detector, it is possible to quote the computing

Table 2. Yearly resource requirements per *building block*.

Processing stage	Storage (TB)	CPU time (MHS06.h)
Raw Filtered Data	300	-
Monitoring and Minimum Bias Data	150	150
Calibration (+ Raw) Data	1500	48
Reconstructed Data	300	238
DST	150	60
Air shower sim.	50	7
Atmospheric muons	25	0.7
Neutrinos	20	0.2

requirements of KM3NeT per *building block*, having in mind that the ARCA programme corresponds to two *building blocks* and ORCA to one. Not all software could be benchmarked, and some estimates are derived by scaling from ANTARES ones. When needed, a conversion factor about 10 between cores and HEPSpec2006 (HS06) is used in the following.

KM3NeT detectors are still in an initial construction stage, although the concepts have been successfully demonstrated and tested in small-scale installations[4]. Because of this, the usage of resources at CNAF has been so far below the figures for a *building block*, but are going to ramp up as more detection units are added in 2016 and the following years. Nevertheless, CNAF has already added relevant contributions to KM3NeT in terms of know-how for IT solution deployment e.g. the Jenkins continuous integration server.

4. References

[1] KM3NeT homepage: http://km3net.org

[2] ESFRI 2016 *Strategy Report on Research Infrastructures*, ISBN 978-0-9574402-4-1

[3] KM3NeT Collaboration: S. Adrián-Martínez et al. 2016 *KM3NeT 2.0 Letter of Intent for ARCA and ORCA*, arXiv:1601.07459 [astro-ph.IM]

[4] KM3NeT Collaboration: S. Adrián-Martnez et al. 2016 *The prototype detection unit of the KM3NeT detector*, *EPJ C* **76** 54, arXiv: 1510.01561 [astro-ph.HE]

ICARUS Experiment

A. Rappoldi, on behalf of the ICARUS Collaboration
INFN, Sez. di Pavia, via Bassi, 6, 27100 Pavia, Italy
E-mail: andrea.rappoldi@pv.infn.it

Abstract. The ICARUS experiment concluded in June 2013 a very successful, long duration run with the T600 detector at the LNGS underground laboratory, taking data with both the CERN neutrino beam directed to the Gran Sasso Laboratory (CNGS) and cosmic rays. Many relevant physics and technical results, achieved during the three years long run at CNGS, have been already presented in several reports to the meetings of LNGS Scientific Committee [1, 2].

In this short report, some results of the currently ongoing analysis performed by the ICARUS Collaboration are presented, namely the analysis of CNGS interactions and the analysis of cosmic events.

In the next future the ICARUS detector will be one of three detectors exposed to ~ 0.8 GeV FNAL Booster neutrino beam, within an experiment proposed in the framework of the Short Baseline Neutrino Oscillation Program (SBN) at Fermilab as the definitive answer to the ”sterile neutrino puzzle” [3]. Presently, the activities on the detector overhauling and upgrading are proceeding according the schedule at CERN, within the common INFN and CERN effort (WA104 project) [4].

1. Update of CNGS events study

In order to identify CNGS ν_μ and $\overline{\nu_\mu}$ charged current (CC) interactions in the ICARUS detector, identified by a minimum of 2.5 m long muon track in the event, the following cut has been applied to the fiducial volume: events with the vertex in the last 2.5 m of the detector have not been considered, in order to allow the muon momentum reconstruction using multiple scattering method. The total number of identified ν_μ and $\overline{\nu_\mu}$ CC events satisfying this additional requirement is equal to 1285. All these events have been visually scanned and analysed, validating the event energy reconstruction. The two main components in the ν_μ CC events can be easily distinguished, namely: the muon track, and the hadronic shower accompanying it. Of course, muon and hadronic energy depositions should sum to the muon neutrino incident energy, however one has to correct for undetected (mostly due to neutrals) or not contained particles. Such correction will be obtained from reconstruction of Monte Carlo ν_μ CC events. Energy of the ionizing hadronic part can be measured directly from the collected charge in the anode wires. Deposited energy has to be corrected for charge quenching related to electrons recombination in liquid argon, and charge attenuation related to the purity of liquid argon, which has influence on electrons lifetime. The quenching factor is obtained with a full FLUKA Monte Carlo simulation of CNGS ν_μ CC events in the T600 detector. So far about 300 events, out of 1285, have been visually reconstructed following the standard procedures developed by the ICARUS Collaboration. The Fig. 1 shows the distribution of muon track length obtained from reconstructed collected CNGS ν_μ and $\overline{\nu_\mu}$ CC events compared with the Monte Carlo expectations.

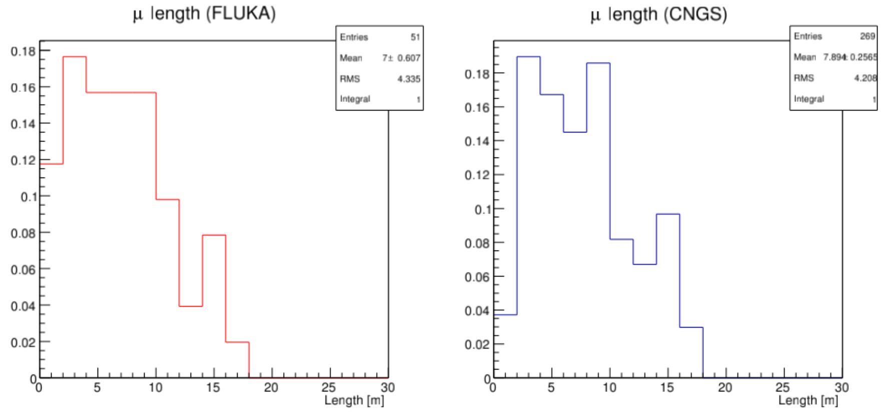


Figure 1. Reconstructed muon track length from ν_μ and $\overline{\nu_\mu}$ CC interactions: FLUKA Monte Carlo (left) and analysed data (right).

2. Update of the cosmic event study

The ICARUS T600 detector collected, apart from the CNGS ν beam events, also cosmic ray data with a total exposure of 0.73 kt/yr. The events collected with the cosmic ray trigger were analyzed also aiming at the study of atmospheric neutrino interactions inside the detector. According the Monte Carlo calculations, about 200 atmospheric neutrinos are expected for 0.73 kt/yr exposure. Atmospheric neutrino interactions are selected, rejecting at the same time events due to incoming cosmic muons, by means of an automatic classification procedure complemented by careful visual scanning. The automatic procedure, already described in the November 2015 report to LNGS SC, includes: the rejection of empty events with a 20 MeV threshold on the deposited energy, the grouping of hits in clusters, the 3D track and vertex reconstruction, the rejection of cosmic muons. An algorithm for automatic search of an interaction vertex allowed to discard "one prong" (i.e. one isolated track) and to identify "multi prong" event topologies for further analysis based on visual scanning by requiring:

- a minimum of 25 consecutive hits to define a track;
- an angle smaller than 170° is between two tracks; for events with only two tracks an angle smaller than 165° is required;
- at least 35 % of the hits within a distance of 100 mm from the vertex has to be assigned to identified tracks;
- the agreement of the drift coordinates for the reconstructed interaction vertex in different views better than 20 mm.

The performance of the automatic selection when applied to MC atmospheric neutrino events is summarized in Table 1. The crossing cosmic muons are automatically rejected with ~ 95 % efficiency, whereas only ~ 9 % of the atmospheric neutrinos signal is lost. The identification efficiency of the neutrino interactions determined with the Monte Carlo events is ~ 86 % for ν_μ CC, ~ 85 % for ν_e CC and ~ 36 % for NC. Overall 45 ± 6 atmospheric neutrino events with at least two charged particles and 67 ± 8 with only one charged particle are expected in the full T600 exposure.

So far, ~ 14 % of the whole data have been completely analysed, including the final time–consuming visual scanning stage. In this analysed sample, 2.9 ± 1.2 muon-like, 2.2 ± 1.0 electron-like and 1.0 ± 0.4 NC-like multi-prong atmospheric neutrino events are expected. The recorded out–of–spill event sample corresponding to 262148 triggers was filtered resulting in 10551 multi-prong neutrino event candidates to be visually studied, finding 3, 0, and 5 event candidates in

	ν_μ CC	ν_e CC	ν NC	**Total**
Empty (< 20 MeV) / cosmic μ	10 ± 1	8 ± 1	45 ± 6	63 ± 8
Small cluster candidate (E $<$ 400 MeV)	8 ± 1	12 ± 1	2 ± 1	22 ± 3
One prong ν candidate	33 ± 4	19 ± 3	15 ± 2	67 ± 8
Multi-prong ν candidate	21 ± 3	16 ± 2	8 ± 1	45 ± 6
Total	72 ± 8	55 ± 7	70 ± 8	197 ± 14

Table 1. MonteCarlo expectations for 0.73 kt/yr exposure after the selection criteria: in total, 197 atmospheric neutrino interactions are expected.

each of these categories, respectively. It should be noted that a residual neutron background, which could affect the NC-like events has not been addressed requiring a deeper dedicated analysis. Fig. 2 shows a muon-like atmospheric neutrino event candidate identified in the analysis, with a 230 cm length muon track exiting from the detector and two charged particles, with a total measured deposited energy of about 630 MeV.

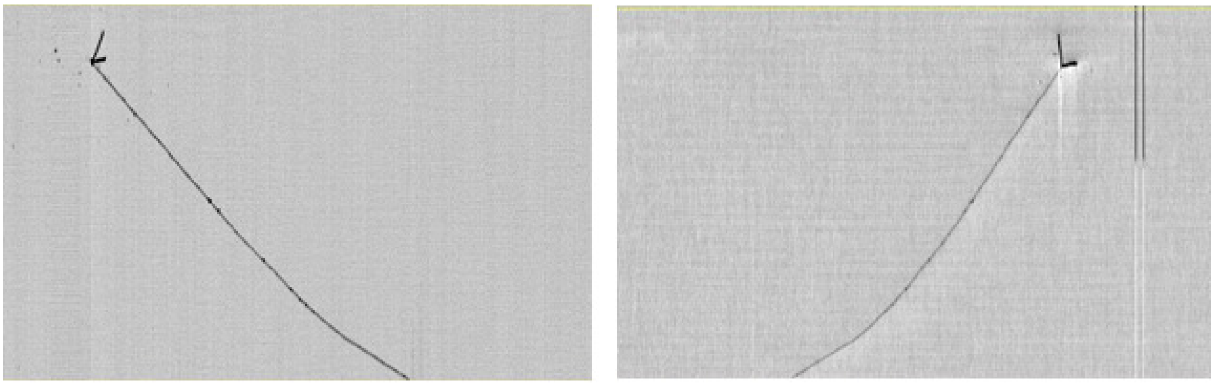


Figure 2. A muon-like atmospheric neutrino event candidate with total deposited energy E_{dep} = 630 MeV. Collection (left) and Induction 2 views (right) are shown. Both the exiting muon and the hadronic tracks are clearly visible. The muon escapes through the detector bottom.

The collaboration has successfully investigated the possibility to reorganize the analysis of cosmic ray events, more heavily exploiting ICARUS computer cluster at LNGS to perform the initial automatic data reduction phase. To this end the collected data presently stored at CNAF, as described in Sec. 4, will be transferred back to the LNGS, where appropriate computer codes for event reconstruction have been already installed and validated. The strongly reduced amount of data preselected by this automatic procedure will then be sent to the scanners for the second phase of visual analysis. This should speed up the process of data preparation for visual scanning removing some limitation in computer and bandwidth resource in external laboratories.

3. Computing model

The ICARUS detector is a 600 tons liquid Argon Time Projection Chamber (LAr TPC) with a total of about 53000 sensitive wires that give an electric signal proportional to the charge released in LAr volume by ionizing particles. This large amount of signals is read by means of a set of 96 readout units (equipped crates mounted on the top of the detector) that perform the analog to digital conversion (with 10 bit ADCs) and send the digital data to the *event-builder* unit, using a dedicated Ethernet network [5], where each readout crate is connected with a 100

Mbps link to the aggregation switches, which in turn are connected to the control room via optical links at 1 Gbps.

A full event, consisting of the digitized charge signals of each wire (4096 samples per wire), has a size of about 220 MB, which may be reduced by a factor 4 with appropriate compression. Since the average rate of data acquisition is about 0.36 mHz, the amount of recorded data is about 170 GB/day.

The data produced by the 600 tons liquid Argon Time Projection Chamber (LAr TPC) of the ICARUS experiment can be thought as the images of an electronic bubble chamber, as clearly depicted in Fig. 2.

This means that the data analysis is carried out mainly by means of interactive graphics programs, by means of which the various tracks that compose each event (seen from different angles) are identified and reconstructed in order to identify the type of interaction that has taken place within the detector, and to measure its kinematic parameters.

Depending on the type of analysis to be performed, the raw data are first selected by means of appropriate filter procedures, generally executed in batch mode.

4. Role and contribution of CNAF

As discussed above (Sec. 3), both the selection procedure and the interactive analysis of the events are conveniently performed locally at the laboratories of the institutions involved in the project.

However, due to the rather large size required for the data, about 220 TB for the data collected in three years, it was advantageous to use the storage system available to CNAF.

Therefore, all the most relevant data collected at LNGS during the detector operation was duplicated into the storage system of the CNAF, where they can be easily retrieved from any of the institutions that are part of the ICARUS collaboration. The repository of the data stored at CNAF amounts to about 3.7 milion of file (divided into 3415 *runs*, acquired with different condition and different trigger logic), for a total of 223 TB.

Concerning to the data recovery, the experience has shown that, since the data are divided into small files (about 55 MB each one), the procedure is much more efficient if performed in two phases: in the first one the online recall (from tape to disk) of a a relatively large number of files (order of 50000) is performed, while in the second the actual transfer of all the files via the network is accomplished. In this way it is possible to recover (at any site where the analysis is performed) up to 15000 event/day, corresponding to about 5 days of data taking of the detector.

References

[1] C. Farnese, *Some recent results from ICARUS*, AIP Conf. Proc. **1666** (2015) 110002. doi:10.1063/1.4915574

[2] J. Kisiel [ICARUS T600 Collaboration], *ICARUS T600 - a large Liquid Argon Time Projection Chamber*, J. Phys. Conf. Ser. **650** (2015) no.1, 012004. doi:10.1088/1742-6596/650/1/012004

[3] A. Zani [ICARUS Collaboration], *ICARUS and Sterile Neutrinos*, Nucl. Part. Phys. Proc. **265-266** (2015) 330. doi:10.1016/j.nuclphysbps.2015.06.084

[4] M. Bonesini [ICARUS/WA104 Collaboration], *The WA104 Experiment at CERN*, J. Phys. Conf. Ser. **650** (2015) no.1, 012015. doi:10.1088/1742-6596/650/1/012015

[5] S. Amerio et al., *Design, construction and tests of the ICARUS T600 detector*, Nucl. Instr. and Meth. **A** 527 (2004) 329

LHCb Computing at CNAF

C. Bozzi
CERN, EP/LBD, CH-1211 Geneve 23, Switzerland, and
INFN Sezione di Ferrara, via Saragat 1, 44122 Ferrara, Italy
E-mail: Concezio.Bozzi@fe.infn.it

V. Vagnoni
CERN, EP/LBD, CH-1211 Geneve 23, Switzerland, and
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 2015 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. More than 330 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.

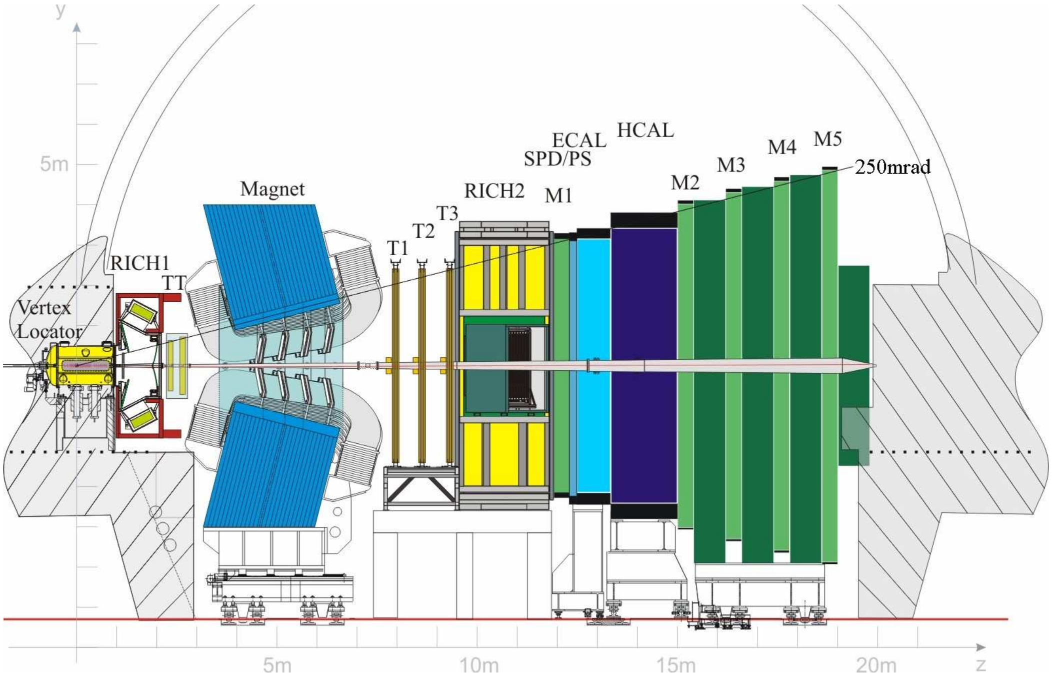


Fig. 1: Sketch of the LHCb detector.

2. Recent evolutions of the LHCb computing model

The initial LHCb computing model, described in the LHCb Computing TDR [2], had a number of shortcomings that made it expensive on resources. As discussed in previous reports, these limitations were addressed during Run1 data taking and during the first Long Shutdown (2013-2014) of the LHC.

In addition, LHCb has implemented in Run2 a new trigger strategy, by which the high level trigger (HLT) is split in two parts. The first one, synchronous with data taking, writes events at a 150kHz output rate in a temporary buffer. Real-time calibrations and alignments are then performed and used in the second high-level trigger stage, where event reconstruction algorithms as close as possible to those run offline are applied. This split HLT concept avoids completely the need for reprocessing campaigns.

Events passing the high level trigger selections are sent to offline, either via a "FULL" stream of RAW events which are then reconstructed and processed as in Run1, or via a "TURBO" stream which outputs the results of the online reconstruction in a micro-DST format which does not require further processing and can be used right away for physics analysis.

The LHCb computing model assumes 10kHz of FULL stream and 2.5kHz of TURBO stream in proton collisions. In 2015, rates of up to 18kHz and 5kHz were measured for FULL and TURBO, resulting in a throughput rate of up to 1.4GB/s and a transfer rate to the Tier0 of up to 750MB/s. Given the LHC live time in 2015 being considerably lower than initially foreseen, the offline resources were enough to cope with these higher rates. These rates were later readjusted in 2016 to 8kHz (FULL) and 4kHz (TURBO).

In 2015, the RAW data associated to the TURBO stream were written to storage for validation reasons. These RAW data will be dropped when the TURBO stream will be fully validated.
Other than proton collisions, LHCb also took data in ion collisions during the last weeks of 2015, and proton-ion collisions in a fixed-target configuration. The associated datasets account for less than 20% of the total.

As in previous years, LHCb continued to make use of opportunistic resources, which are not pledged to WLCG, which significantly contributed to the overall usage. The most significant unpledged contribution were due to the LHCb HLT farm and resources from the Yandex company. Small scale production use has been made of virtual machines on clouds and other infrastructures, including HPC resources (Ohio Supercomputing Center) and volunteer computing through the BOINC framework. This integration of non-WLCG resources in the LHCb production system is made possible by the DIRAC framework [3] for distributed computing.

The use of storage resources has been optimised by reducing the number of disk-resident copies of the analysis data. Disk provisions at certain large Tier-2s (T2D), introduced in 2013, continued to grow, thereby allowing users to run analysis jobs at this sites, and further blurring the functional distinction between Tier-1 and Tier-2 sites in the LHCb computing model.

3. Resource usage in 2015

Table 1 shows the resources pledged for LHCb at the various tier levels for the 2015 period.

2015	CPU	Disk	Tape
	(kHS06)	(PB)	(PB)
Tier0	36	5.5	11.2
Tier1	139	14.0	28.1
Tier2	61	2.0	
Total WLCG	**236**	**21.5**	**39.3**

Tab. 1: LHCb 2015 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. 2. The same data is presented in tabular form in Tab. 2. It must be emphasised that CNAF is the second highest CPU power contributor, slightly lower than the IN2P3 computing center. The CNAF contribution is about 20% higher than the pledged one. This achievement has been possible owing to great stability, in particular of the storage system, leading to maximal efficiency in the overall exploitation of the resources.

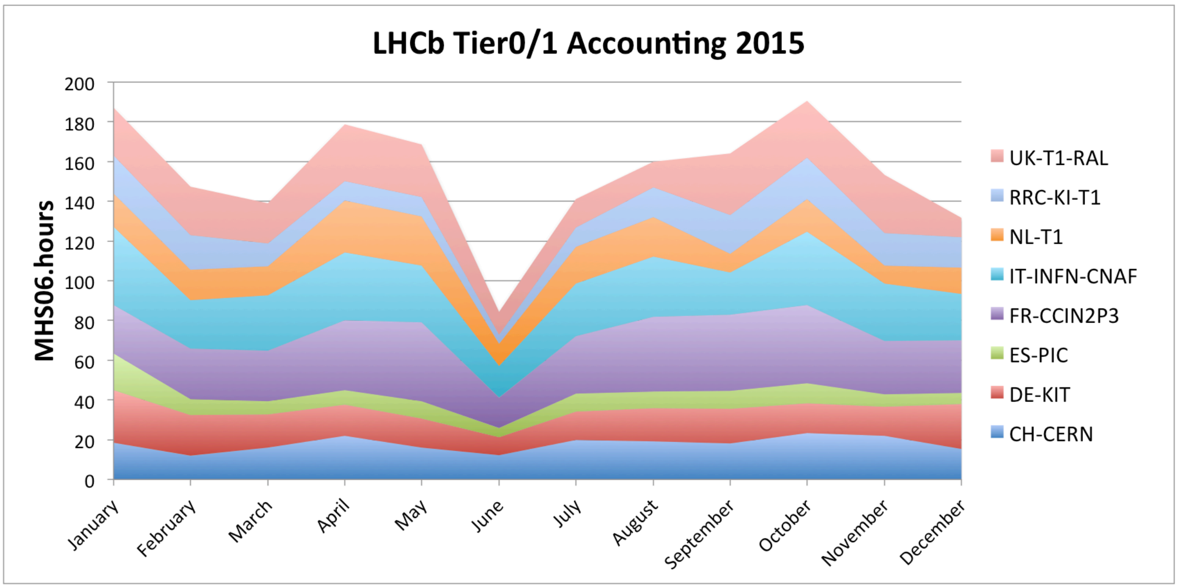


Fig. 2: Monthly CPU work provided by the Tier-0 and Tier-1s to LHCb during 2014.

<Power>	Used (kHS06)	Pledge (kHS06)
CH-CERN	17.9	36
DE-KIT	17.0	19.6
ES-PIC	8.5	7.7
FR-CCIN2P3	30.2	23
IT-INFN-CNAF	28.1	23.6
NL-T1	16.3	15.7
RRC-KI-T1	14.0	14.2
UK-T1-RAL	21.8	35.4
Total	**153.8**	**175.2**

Tab. 2: Average CPU power provided by the Tier-0 and the Tier-1s to LHCb during 2015.

The number of running jobs at Tier-0 and Tier-1s is detailed in Fig. 3. As seen in the top figure, 70% of the CPU work is due to Monte Carlo simulation.

The usage of the Storage is the most complex part of the LHCb computing operations. Tape storage grew by about 6 PB. Of these, 4PB were due to RAW data taken in the last four months of 2015. The rest was equally shared among RDST and ARCHIVE, the latter due to the archival of Monte Carlo productions, the legacy stripping of Run 1 data, and new Run2 data. The total tape occupancy as of December 31st 2015 is 21.6 PB, 10 PB of which are used for RAW data, 5.6 PB for RDST, 6 PB for archived data. The total tape occupancy at CNAF is 3.4PB.
Table 3 shows the situation of disk storage resources at the Tier-0 and Tier-1s at the end of December 2015. Despite the lower disk pledges, CNAF has been the second Tier-1 in terms of disk storage made available to LHCb. The available disk space was sufficient to cover the various activities foreseen until the end of the 2015 WLCG year.

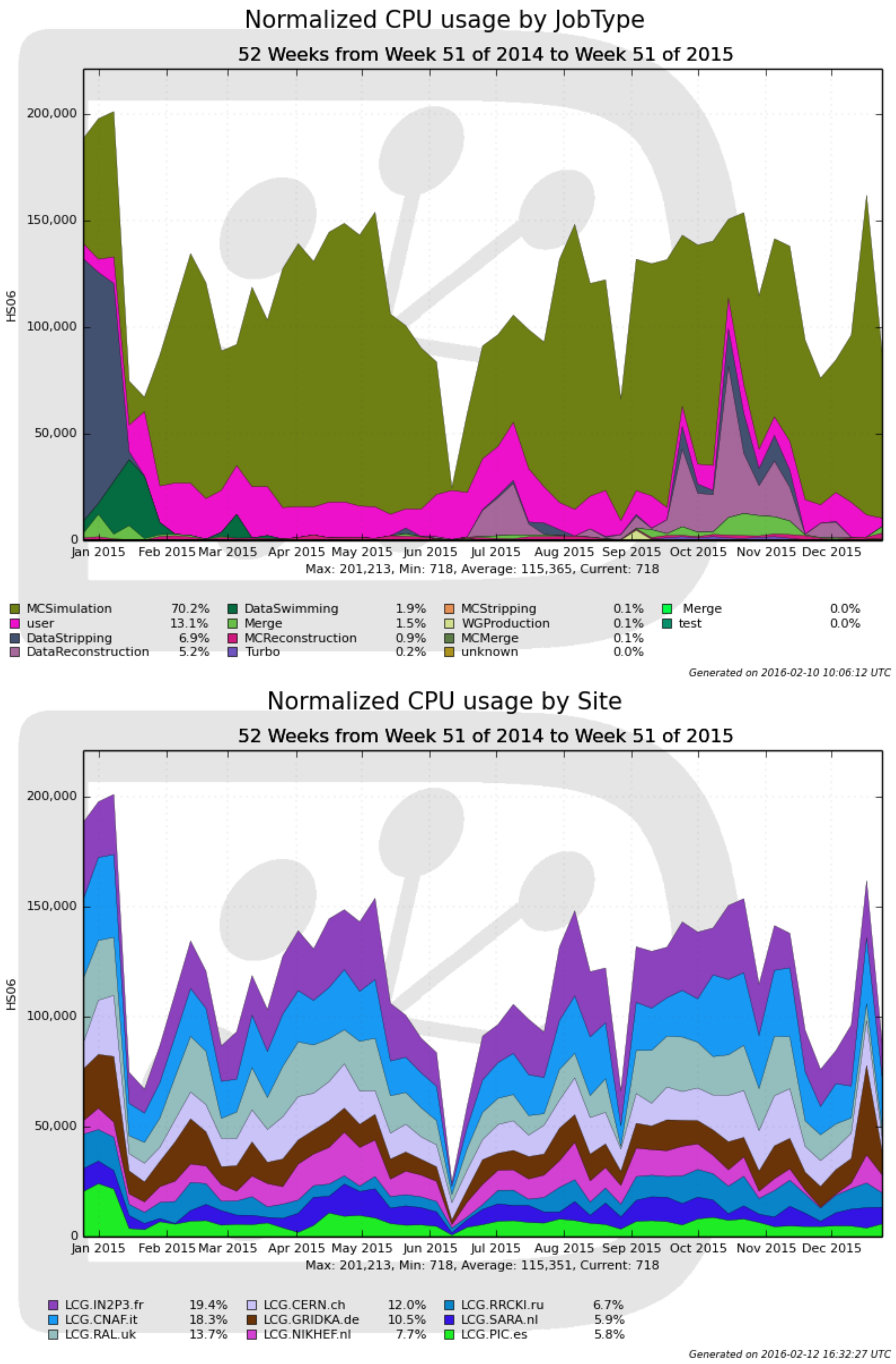


Fig. 3: Usage of LHCb resources at Tier-0 and Tier-1s during 2015. The top plot shows the usage of resources for the various activities, whereas the bottom plot shows the contributions from the different countries.

Disk (PB)	CERN	Tier1s	CNAF	GRIDKA	IN2P3	PIC	RAL	RRCKI	SARA
LHCb accounting	**3.98**	**11.46**	**2.40**	**1.92**	**1.56**	**0.71**	**3.13**	**0.55**	**1.20**
Disk used	4.08	11.48	2.43	1.91	1.56	0.71	3.12	0.55	1.20
Disk free	0.72	2.37	0.17	0.22	0.22	0.08	0.06	0.71	0.35
Staged area (used+free)	0.91	3.23	2.72	0.14	0.03	0.01	0.20	0.01	0.03
Disk total	**5.61**	**17.08**	**5.32**	**2.27**	**1.81**	**0.80**	**3.94**	**1.35**	**1.58**
Pledge 2015	***5.50***	***14.04***	***2.72***	***2.34***	***1.88***	***0.76***	***3.51***	***1.26***	***1.57***

Tab. 3: Situation of disk storage resource usage as of December 31st 2015, available and installed capacity, and 2015 pledge. The contribution of each Tier1 site is also reported.[1]

In summary, the usage of computing resources in the 2015 has been quite smooth for LHCb.

The legacy stripping of the Run1 dataset was completed at the end of January. A “swimming” activity was run shortly afterwards. The workflow for the Run2 data taking was exercised in the spring, and successfully put to work as soon as LHC started providing collisions.

The ramp-up of the LHC luminosity resulted in a corresponding increase of the data throughput to offline resources. Due to somewhat relaxed trigger thresholds, throughputs of up to 1.3GB/s were input to the offline system, which was able to keep up without major problems.

Simulation has been running at almost full speed using all available resources, being the dominant activity in terms of CPU work. Additional unpledged resources, as well as clouds, on-demand and volunteer computing resources, were also successfully used.

Storage resources were not a concern. The storage resources provided to LHCb were not saturated by the end of the WLCG year in March 2016, due to the reduced LHC live time with respect to the initial expectations.

The usage of datasets produced for physics analysis is constantly monitored, the subsequent analysis of data popularity having allowed LHCb to free a significant amount of disk space.

[1] The total allocated space is 3PB above the 2015 pledges, 2.7PB of which are due to an artifact in the assignment of buffer space at CNAF.

4. Expected resource growth

In terms of CPU requirements, Tab. 4 presents for the different activities the CPU work estimates for 2016 and 2017. The last row shows the power averaged over the year required to provide this work.

CPU Work in WLCG year (kHS06.years)	**2016**	**2017**
Prompt Reconstruction	31	34
First pass Stripping	13	14
Full Restripping	17	14
Incremental Restripping	2	8
Simulation	153	218
VoBoxes and other services	4	4
User Analysis	20	24
Total Work (kHS06.years)	**240**	**316**
Efficiency corrected average power (kHS06)	**285**	**375**

Tab. 4: Estimated CPU work needed for the various LHCb activities in 2016-2017. Proton physics.

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. 5. The table also shows resources available to LHCb from sites that do not pledge resources through WLCG.

Power (kHS06)	**Forecast 2016**	**Forecast 2017**
Tier 0	48	61
Tier 1	146	189
Tier 2	81	106
Total WLCG	**275**	**356**
HLT farm	10	10
Yandex	10	10
Total non-WLCG	**20**	**20**

Tab. 5: CPU power requested at the different Tiers in 2016-2017. Minimal additional resources with respect to the ones shown in Tab. 4 are requested for heavy ions physics.

Tables 6 and 7 present, for the different data classes, the forecast total disk and tape space usage at the end of the years 2016-2017. These disk and tape estimates are then broken down into fractions to be provided by the different Tiers. These numbers are shown in Tables 8 and 9. These tables also include small additional resources needed for heavy ions physics. These additional resources are quoted in Table 10. 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. Some mitigation measures, such as the archival of a single copy of all derived datasets, and the removal of raw banks from the reconstruction output, were put in place to keep the growth in the tape storage requirement to a manageable level.

LHCb Disk storage usage forecast (PB)	**2016**	**2017**
Stripped Real Data	11.1	14.7
Simulated Data	6.4	10.4
User Data	1.1	1.2
MDST.DST	1.2	0.0
RAW and other buffers	1.0	1.0
Other	0.5	0.6
Total	**21.3**	**27.9**

Tab. 6: Breakdown of estimated disk storage usage for different categories of LHCb data (proton physics)

LHCb Tape storage usage forecast (PB)	**2016**	**2017**
Raw Data	18.3	28.3
FULL.DST	8.0	11.3
MDST.DST	3.8	6.8
Archive	9.6	13.3
Total	**39.7**	**59.7**

Tab. 7: Breakdown of estimated tape storage usage for different categories of LHCb data (proton physics)

LHCb Disk (PB)	**2016 Forecast**	**2017 Forecast**
Tier0	5.8	8.7
Tier1	14.9	17.7
Tier2	2.8	3.8
Total	**23.5**	**30.2**

Tab. 8: LHCb disk request for each Tier level (proton + heavy ions physics). 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)	**2016 Forecast**	**2017 Forecast**
Tier0	15.0	22.0
Tier1	25.8	38.8
Total	**40.8**	**60.8**

Tab. 9: LHCb tape request for each Tier level (proton + heavy ions physics)

Resources for heavy ion running	**2016 Forecast**	**2017 Forecast**
CPU (kHS06)	10.1	0.8
Disk (PB)	2.2	2.3
Tape (PB)	1.1	1.1

Table 10: LHCb requests for heavy ions physics in 2016 and 2017.

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 until 2017. It has been shown that CNAF has been in 2015 the second most important LHCb computing centre in terms of CPU power made available to the collaboration. This achievement 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] [LHCb collaboration], CERN-LHCC-2005-019.

[3] F. Stagni *et al.*, J. Phys. Conf. Ser. **368** (2012) 012010.

The PAMELA experiment

A. Bruno
on behalf of the PAMELA collaboration
INFN and University of Bari, Bari, Italy

Abstract. PAMELA has been in orbit studying cosmic rays for more 9 years. Its operation time will continue in 2016. In this paper we will present some of the latest results obtained by PAMELA in flight, describing 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 galactic, solar and magnetospheric radiation in the near-Earth environment.

2. Results obtained in 2015

The PAMELA satellite experiment is providing comprehensive observations of the cosmic ray radiation in low Earth orbits. Thanks to its identification capabilities and the semi-polar orbit, PAMELA is able to precisely measure the energetic spectra and the angular distributions of the cosmic-ray populations in different regions of the terrestrial magnetosphere. In particular, PAMELA reported accurate measurements of the geomagnetically trapped protons in the inner Van Allen belt, extending the observational range for protons down to lower altitudes (South Atlantic Anomaly – where the inner belt makes its closest approach to the Earth's surface), and up to the maximum kinetic energies corresponding to the trapping limits (a few GeV) [1].

PAMELA also provided detailed measurements of the re-entrant albedo populations generated by the interaction of cosmic ray from the interplanetary space with the Earth's atmosphere [2]. On the basis of a trajectory tracing simulation, the analyzed protons were classified into quasi-trapped, concentrating in the magnetic equatorial region, and un-trapped spreading over all latitudes, including both short-lived (precipitating) and long-lived (pseudo-trapped) components. In addition, features of the penumbra region around the geomagnetic cutoff were investigated in detail.

PAMELA's observations comprise the Solar Energetic Particle (SEP) events between solar cycles 23 and 24. Specifically, PAMELA is providing the first direct observations of SEPs in a large energetic interval (>80 MeV) bridging the low energy measurements by in-situ spacecrafts and the ground level enhancement data by the worldwide network of neutron monitors. Its unique observational capabilities include the possibility of measuring the flux angular distribution and thus investigating possible anisotropies associated to SEP events. Results were supported by an accurate back-tracing analysis based on a realistic description of the Earth's magnetosphere, which was exploited to estimate the SEP fluxes as a function of the asymptotic direction of arrival [3].

The trajectory tracing approach was also exploited in the search for large-scale anisotropies in the arrival directions of high-energy (>10 GeV) cosmic ray positrons, allowing a full sky investigation with sensitivity to global anisotropies in any angular window of the celestial sphere. The resulting distributions of arrival directions were found to be isotropic: starting from the angular power spectrum, a dipole anisotropy upper limit of 0.076 at the 95% confidence level was determined [5].

PAMELA's measurements of the electron component of the cosmic radiation were used to investigate the effects of propagation and modulation of galactic cosmic rays in the heliosphere, particularly significant for energies up to at least 30 GeV. Solar modulation effects were studied using data acquired between 2006 July to 2009 December over six-month time intervals, placing significant constraints on the theoretical transport models [4].

Finally, PAMELA performed a direct search for strange quark matter in cosmic rays [6]. If this state of matter exists it may be present in cosmic rays as particles, called strangelets, having a high density and an anomalously high mass-to-charge (A/Z) ratio. PAMELA's observations in space are complementary to those from ground-based spectrometers, and take the advantage of being potentially capable of directly identifying these particles, without any assumption on their interaction model with Earths atmosphere and the long-term stability in terrestrial and lunar rocks.

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.

4. SEP trajectory reconstruction

The analyses described in Section 2 are supported by accurate simulations of particle trajectories in the terrestrial magnetosphere. Using spacecraft ephemeris data (position, orientation, time), and the particle rigidity (R = momentum/charge) and direction provided by the PAMELA tracking system, the trajectories of all selected down-going protons were reconstructed by means of a tracing program based on numerical integration methods [7], and implementing realistic models of internal and external geomagnetic fields. In particular, the trajectory approach was applied to the study of all SEP events registered by PAMELA, allowing the estimate of proton fluxes as a function of the asymptotic direction of arrival, thus the investigation of possible anisotropies associated to SEP events. Because the directional response of the apparatus varies with satellite position/orientation and particle rigidity, the calculation was performed for 1-sec time steps along the orbit and 22 rigidity values between 0.39 – 4.09 GV, for a total of $\sim 8 \times 10^7$ trajectories for each polar pass.

References

[1] Adriani, O., et al., Trapped Proton Fluxes at Low Earth Orbits Measured by the PAMELA Experiment, ApJ 799 L4, 2015, doi:10.1088/2041-8205/799/1/L4.

[2] Adriani, O., et al., Re-entrant albedo proton fluxes measured by the PAMELA experiment, J. Geophys. Res. Space Physics, 120, 2015, doi:10.1002/2015JA021019.

[3] Adriani, O., et al., PAMELA's Measurements of Magnetospheric Effects on High Energy Solar Particles, ApJ 801 L3, 2015, doi:10.1088/2041-8205/801/1/L3.

[4] Adriani, O., et al., Time Dependence of the e- Flux Measured by PAMELA during the July 2006-December 2009 Solar Minimum, ApJ, 810, 142, 2015, doi:10.1088/0004-637X/810/2/142.

[5] Adriani, O., et al., Search for Anisotropies in Cosmic-Ray Positrons Detected by the PAMELA Experiment, ApJ 811 21, 2015, doi:10.1088/0004-637X/811/1/21.

[6] Adriani, O., et al., New Upper Limit on Strange Quark Matter Abundance in Cosmic Rays with the PAMELA Space Experiment, Phys. Rev. Lett. 115, 111101, 2015, doi:10.1103/PhysRevLett.115.111101.

[7] Smart, D. F., & Shea, M. A., Final Report, Grant NAG5–8009, Center for Space Plasmas and Aeronomic Research, University of Alabama in Huntsville, 2000.

XENON computing activities

G. Sartorelli, F. V. Massoli

INFN e Università di Bologna

E-mail: Gabriella.Sartorelli@bo.infn.it; massoli@bo.infn.it

1. The XENON project

A lot of astrophysical and cosmological observations support the hypothesis that a considerable amount of the energy content of the Universe is made of cold dark matter. Recently, more detailed studies of the Cosmic Microwave Background anisotropies have deduced, with remarkable precision, the abundance of dark matter to be about 25% of the total energy in the Universe. Dark matter candidate particles share some basic properties, mainly: they must be stable or very long lived; they have to be weakly interacting and colorless and they have to be not relativistic. Due to such characteristics, they are identified under the generic name of Weakly Interacting Massive Particles (WIMPs). Among the various experimental strategies to directly detect dark matter, detectors using liquid Xenon (LXe), as XENON100 and LUX, have demonstrated the highest sensitivities over the past years. The XENON collaboration is focused on the direct detection of WIMP scattering on a LXe target. Currently, the XENON100 detector is running at LNGS. It set the most stringent limit, for the 2012, on the spin-independent WIMP-nucleon elastic scattering cross section for WIMP masses above 8 GeV/c^2, with a minimum at $2 \cdot 10^{-45}$ cm^2 at 55 GeV/c^2 (90% CL) [1]. In parallel, since 2011, the XENON1T project started. It will be the largest dual phase (LXe/GXe) Xe-based detector ever realized and, after its approval from all funding agencies in 2011, it is now under construction and installation in Hall B at the Gran Sasso Underground Laboratory (LNGS). Both XENON100 and XENON1T detectors are based on the same detection principles. The target volume is hosted in a dual phase (LXe/GXe) Time Projection Chamber (TPC) that contains Xenon in liquid phase (LXe) with gaseous phase (GXe) on top. The TPC is enclosed by two meshes: the cathode (at negative voltage) on the bottom and the gate mesh (grounded) on top. This structure contains the LXe active region, called the sensitive volume that represents the volume used to detect the interactions. A particle interacting in LXe produces a prompt scintillation signal (S1) through excitation and recombination of ionization electrons. The electrons that do not recombine are drifted towards the liquid-gas interface where they are extracted into the GXe to produce the secondary scintillation signal (S2). Two PMT arrays, one on top of the TPC inside the GXe and one at its bottom below the cathode, in LXe, are used to detect the scintillation light. The x-y position of the events is determined from the PMTs hit, while from the time dfference between S1 and S2 signals the z coordinate is inferred. Hence a 3D vertex reconstruction is possible. The knowledge of the interaction point allows the selection of the events in the inner part of the LXe, usually called “fiducial volume” since the majority of background events are expected to be found outside of it. With respect to its predecessor, XENON1T will use a larger amount of LXe: about 3.3 tonnes 2 of which will represent the sensitive volume available for the WIMP interactions. Its goal is to lower the current limits on the WIMP interaction cross section of about two orders of magnitude. To reach such a result a severe screening campaign has been required in order to choose the materials with the lowest contaminations, and a also a MC study has been developed, through simulations with the GEANT4 toolkit, in order to optimize the detector design and to evaluate the expected background. Due to the large amount of simulations required to perform that research, the GRID is the most appropriate facility to be used.

2. XENON100

To acquire data, the XENON100 detector uses a DAQ machine equipped with a storage buffer of 1.1 TB. That data are then moved on the above-ground facility and stored on 5 disks server with a total capacity of about 214 TB. Raw data are also stored in tapes as backup copy. Data are processed by a dedicated 32-cores server and by U-LITE LNGS batch system which provides shared CPU. For the analysis, there are two 8-cores machines dedicated, plus the availability of our 32-cores server in case of needs. Another 4-cores machine with 2.1 TB of disk space is used to provide several services: home space, web server hosting, SVN repository for code (data processing and Monte Carlo) and documents, run database and the XENON wiki. In the latest published scientific run (2011- 2012), XENON100 collected 225 days of Dark Matter search (light-weight data). For that scientific run, a total amount of data of 17 TB, 53 TB and 1 TB for dark matter search, gamma and neutron calibrations have been collected. The total amount of resources used so far at LNGS are: 210 TB of raw data, 10 TB of processed data and 76k CPU-hours per year.

3. XENON1T

The DAQ system for XENON1T is located in the experiment service building, underground. It has a 6 TB storage buffer that is enough to store data for few hours, at the maximum foreseen acquisition rate of 300 MB/s during calibration runs, or for few weeks in the case of DM data. The system communicates with the surface through a switch connected with one dedicated 10 Gbps optical fiber connection. Most of the raw data will be sent to the GRID storage to be processed and stored. Moreover, GRID is also currently used for massive production and storage of MC simulations. The remote data processing would be feasible only if a reasonable bandwidth to connect LNGS to GRID will be guaranteed. At the moment, we can count on a 1 Gbps connection shared with anyone using in LNGS that route. GARR has recently updated the line at 10 Gbps and LNGS already setup the hardware to support this new line. A dedicated 32-cores disk server is used as disk space for users to be used as home. The total available space is 10 TB which translates into about 100 GB of disk space for each user. A high memory 32-cores machine is used to host several virtual machines, each one running a dedicated service: code (data processing and Monte Carlo), documents repository on SVN/GIT, the run database, the on-line monitoring web interface, the XENON wiki and GRID UI. Virtual machines are created on a 1 TB disk, then three separate 5 TB disks are used to store data for SVN/GIT, Dokuwiki and MongoDB. Another 64-cores local server is used to access to the copy of processed data in LNGS. Moreover, it will work as bridge to access to LNGS homes meaning that this machine will be opened to outside to directly access to the xenon cluster. For data storage, three separates volumes are used: a 50 TB disk for temporary store calibration raw data before they are sent to the GRID storage, a 200 TB disk for DM raw data and a 100 TB volume to store the processed data, i.e. ROOT files. The 200 TB for DM data have been evaluated assuming 45 TB/year of data taking for more than 4 years. To transfer data, a 32-cores machine has been designed. It will run many daemons organized in such a way to perform only one read and one write operation at same time on each storage volume. As previously said, massive raw data processing (from calibration) will be done remotely with GRID. Moreover, the U-LITE LNGS batch system, more than 500 shared CPU in total, will be used mainly for DM data processing. Such a CPU power will be enough to process daily DM data.

For what concerns the CNAF resources, we have allocated 700HS06 and we are currently using about 30 TB, out of 60 TB, of disk space. We are extensively using CNAF GRID resources for data processing and to run (and to store outputs from) MC simulations aimed to the optimization of the detector design, background evaluation and signal simulations. Thanks to this work, we have recently published the results from MC studies [2] regarding the XENON1T background and its foreseen sensitivity to Dark Matter interactions with liquid Xenon. Due to the success of such a work, it is foreseen to continue to use the GRID to produce simulated data and to store them in the related disk storage. Moreover, the number of GRID users is increasing meaning that we will dramatically increase the use of such resources. There is also the possibility to store all the processed calibration data on GRID and to move there also the heaviest part of the analysis related to that kind of data. This will certainly increase the amount of CPU and disk space that is foreseen to be used during 2016 and later.

4. References

[1] Aprile E. et al (XENON Collaboration), *Dark Matter Results from 225 Live Days of XENON100 Data*, 2012, Phys. Rev. Lett. **109**, 181301

[2] Aprile E. et al (XENON Collaboration), *Physics reach of the XENON1T dark matter experiment*, 2016, JCAP **04**, 027

The INFN-Tier1 Center and National ICT Services

The INFN-Tier1

Luca dell'Agnello

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

1. Introduction

The CNAF Data Center 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 HVAC systems in 2008. Currently, over 30 scientific collaborations use computing and storage resources hosted at Tier1, including experiments at accelerator facilities (the above mentioned LHC experiments, CDF, AGATA, KLOE, LHCf, the "new entry" NA62 and, formerly, BaBar and SUPERB), astroparticle physics experiments (AMS, ARGO, Auger, Borexino, FERMI/GLAST, Gerda, ICARUS, MAGIC, PAMELA, Xenon100, VIRGO along with the "new entries" CTA, Opera, Darkside, Cuore, KM3 and LHAASO) and contacts are ongoing with others. The rapid growth of CPU and storage capacity in the last years has been mainly driven by the startup of LHC, and the same trend is confirmed also in the next years.

The Tier-1 staff, composed by about 25 people (including post docs) is structured in 5 groups:

- Farming unit (taking care of the computing farm and of the related grid services such as CEs and UIs);
- Data Management unit (taking care of databases and disk and tape storage devices and services);
- Network unit (managing the whole CNAF LAN and WAN connections)
- Facility management unit (taking care of Data Center infrastructure from electric power to conditioning system)
- User support (interface and support to users)

Our main challenge is to guarantee H24 support: to avoid H24 manpower requirements, all services are completely redundant and based on enterprise level hardware.

The inclusion of User Support group in the Tier-1 division (from 2015), allows better synergies and has increased the number of people able to actively operate on the services. Moreover, at the beginning of the year, the staff number has been increased with 3 new permanent positions (2 for the storage group and 1 for the farming unit). As a result, even if we still have some criticalities (mainly in the Network and in the Facility units), most of the main services (i.e. LSF, GPFS, TSM, and StoRM) are managed by permanent people.

To optimise the resource management effort, WLCG standard services are offered to all users/scientific collaborations. In particular:

- One general purpose farm with both grid (i.e. CEs) and local access supported;
- Posix parallel file-system (GPFS) for data with both grid (i.e. gridftp) and local access supported;
- Standard HSM service (TSM) to access the tape library.

2. The farm

The farm computing power in 2015 was of about 188 kHS06 (corresponding to about 14,000 job slots): every day 100 Kjobs have been processed. The largest share is assigned to LHC experiments (70% in total as shown in figure 1). In order to maximise the CPU usage, the computing resources are assigned according to fair share mechanism: in this way the farm is constantly fully utilised (excepting for maintenance interventions).

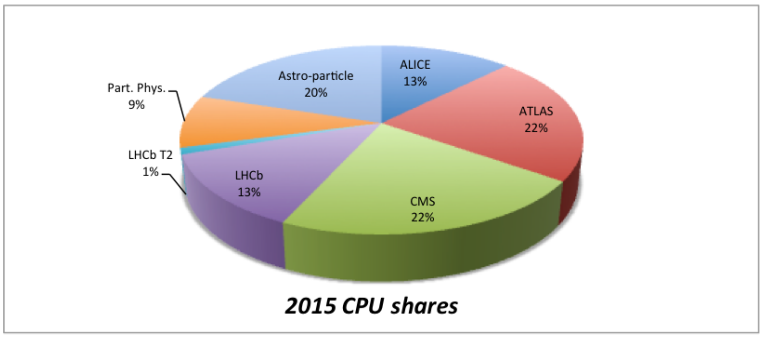


Figure 1. CPU shares at CNAF for 2015

During 2015 the farm has been migrated to the latest version of LSF (LSF 9) in order to cope with new requirements from some experiments (namely the support of ‘cgroups’ to enable users to specify the memory usage): the upgrade has been achieved installing a new LSF master and migrating all WNs from the old to the new farm. This process has been completed in about 6 months (see par. xxx for more details).

Another important change for the computing farm was the introduction of multicore jobs, i.e. jobs using several computing slots at the same time. This workflow, used mainly by Atlas and CMS experiments, requires , usually, 8 free slots on the same WN. To optimise the allocation of WNs a dynamic partitioning of the farm has been implemented to automatically allocate resources to this type of jobs according to the queued jobs.

Beside the main farm, a (small) HPC cluster is also available:

- 24 nodes, 800 cores (about 10 Tflops)
- 17 GPUs
- 3 Intel Xeon Phi

The HPC nodes, interconnected via Infiniband protocol, are operated with same tools as the generic farm (i.e. LSF and GPFS). The HPC cluster is used at 80% on average; the main users are from a theoretical physics group for particle acceleration and laser plasma acceleration simulations.

3. The storage

The storage, both disk and tape, is managed by GEMSS (Grid Enabled Mass Storage System) which is composed from GPFS and TSM ”glued” together by a software layer managing the data movements between the two systems. Beside file, the supported protocols are GridFTP, XrootD, and http/webdav. In 2015 about 18 net PB of disk and about 20 PB of tapes were installed and available to the experiments. As in the case of CPU resources, the largest shares are allocated to the LHC experiments (see figures 2 and 3).

An example of use of resources is depicted in figures 4 and 5, where a sustained throughput of 17 GB/s from CMS jobs during April is clearly visible.

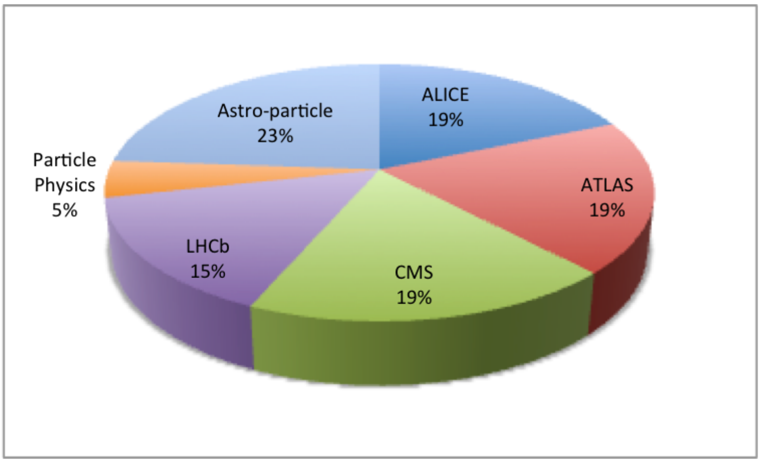


Figure 2. Disk storage shares at CNAF for 2015

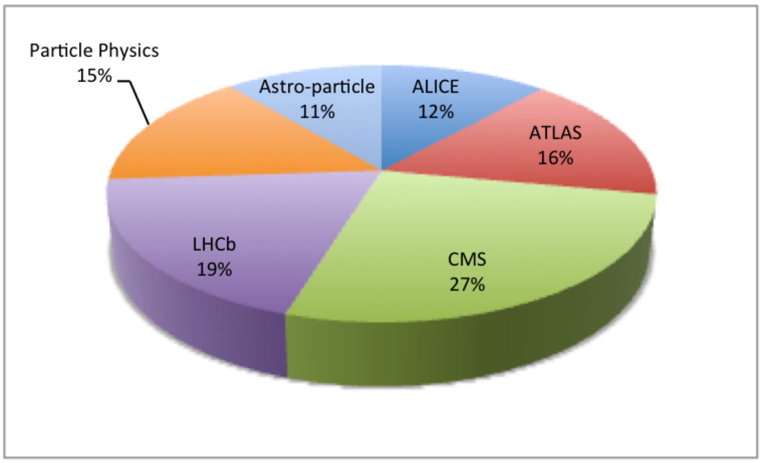


Figure 3. Tape storage shares at CNAF for 2015

On February 2015 a new infrastructure based on Scientific Linux 6 has been installed and configured in order to host the virtual instances of Storm servers: in this way an additional level of redundancy (i.e. the hardware) has been achieved. Moreover during this year, the migration from Quattor to Puppet/Foreman as installation and configuration framework has started: while for the farm resources, this has been completed in 2015, for the storage servers this process will finish in 2016.

3.1. Tape repack

During 2015 all the data present on the tape library have been repacked (i.e. moved) from tape B and C to tape D in order to free slots (see figure 7). The repack process took 10 months (from the beginning of the year to October). Since the repack involves actively the TSM server (all data flow through it), to speed up the operations, in May we had to upgrade of the Fibre Channel connectivity from 4 to 16 Gbps reaching a rate of about 80 TB repacked every day (see figure 6).

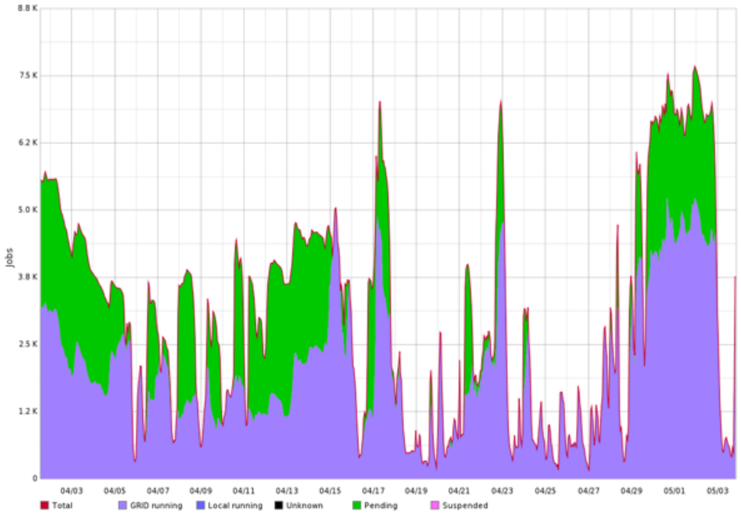


Figure 4. CMS jobs at CNAF during April 2015

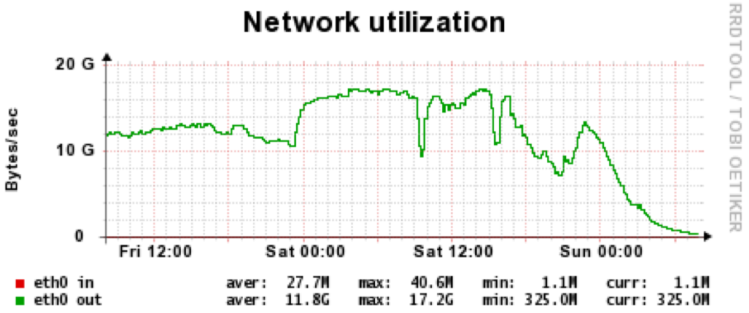


Figure 5. CMS jobs throughput at CNAF during April 2015

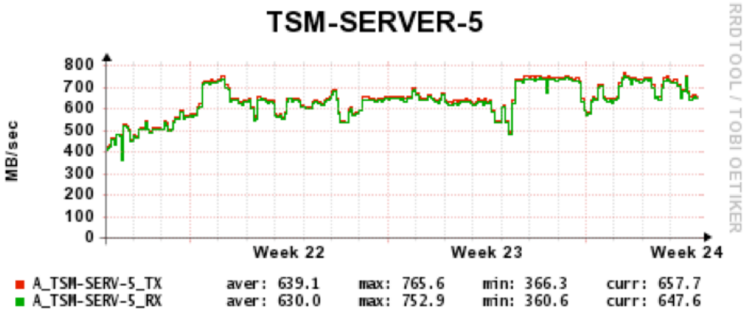


Figure 6. Repack rate

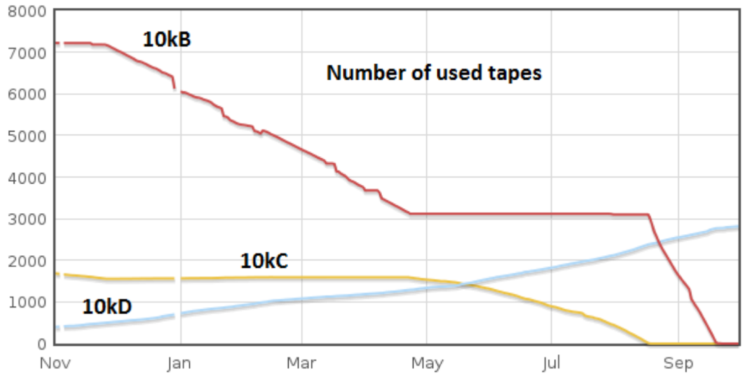


Figure 7. Tape repack during 2015

4. Long Term Data Preservation for CDF

An activity for the Long Term Data Preservation (or LTPD) for CDF data is ongoing since 2013 in collaboration with Fermi Lab and CDF group. There are two main areas of activity:

- Bit preservation – about 4 PB of data have been transferred from FNAL and archived on the tape library at CNAF (with a transfer rate up to 500 MB/s as shown in figure 8). An automated system to perform regular checks of data integrity and copy back from FNAL corrupted files is under development.
- Preservation of code and analysis frameworks – instantiation on demand services and analysis computing resources on pre-packaged VMs job submission to move from a dedicated portal (Eurogrid) to jobsub, to permit execution of legacy software on SL6 nodes. The metadata, accessed directly at FNAL through Squid servers, will be copied to a local DB to ensure complete independence from FNAL.

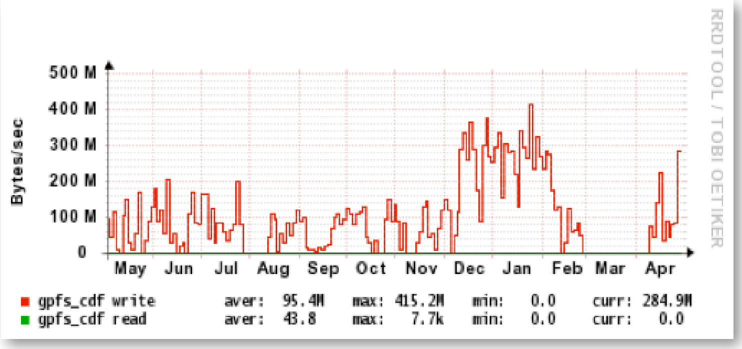


Figure 8. Rate of CDF data transfer from FNAL to CNAF

5. Tier-1 extension on external sites

Given the foreseen huge increase of resources during and after LHC Run-2 (especially for CPU), there is a strong interest in testing usage of remote resources for (dynamically) extend the Tier-1 farm. INFN participates to the HelixNebula Science Cloud, approved at the end of 2015. It is a Pre-commercial Procurement project, funded for 2/3 by the European Commission, aiming to build an hybrid cloud (i.e. using also resources on commercial clouds) for the science. To gain experience for this scenario, we have started a small scale test with one of the main Italian commercial cloud providers (Aruba). Aruba offered us the possibility to opportunistically use about 160 cores with the caveat that the processes could be frozen (but not killed) in case a paying user needed the resource. These resources are virtual servers accessed through VMWare APIs. For this test we decided to try the disk-less approach for the data, hence the resources have used only from experiments able to use XrootD to access the data (e.g. CMS). On the other hand, we needed to cache locally the LSF shared area in order to transparently extend the CNAF farm on Aruba resources. The network access is via the General Purpose Network: given the small scale of the test a dedicated network is not needed. The efficiency of the jobs on the Aruba resources is a bit lower (about 10% on average) excepting for the MonteCarlo jobs with have nearly the same efficiency: the main issue in general is that at site level there is generally no possibility to select the jobs to be submitted on remote resources.

During the last quarter of the year, we have started testing the use of remote resources at Bari-ReCaS: a number of servers with a computing capacity of about 20,000 HS06 has been leased from INFN and assigned to CNAF to use for LHC experiments. In this case, it is a static allocation of resources: these are completely dedicated to CNAF and directly managed by CNAf staff. As in case of the Aruba test, the servers have been statically inserted in CNAF farm with a cache of the LSF shared area at Bari. Since there resources should be used from all LHC experiments (i.e. not only from those XrootD enabled) and to minimise the network usage, a local storage is also needed. To avoid thwarting the saving of using there leased computing resources adding significative storage capacity, a cache system has been put in place to optimise the access to data present at CNAF. A dedicated L3 VPN ($2x10$ Gbps) has been set-up from GARR in order to extend LAN CNAF on Aruba (see figure 9): the link has been positively commissioned (see figure 10). Results are foreseen for mid 2016.

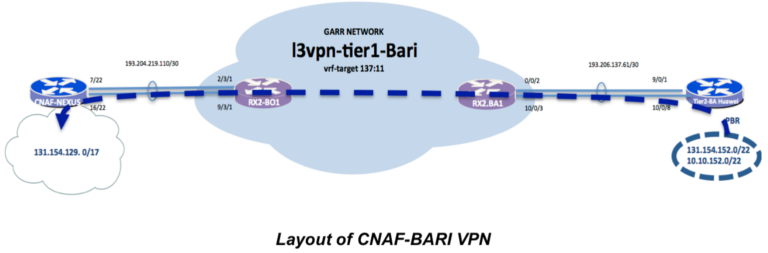


Figure 9. Layout of CNAF LAN extension on Bari-RECAS

Another interesting test has been started with CINECA for HPC: an INFN theoretical physics group uses HPC resources at CINECA and needs to write data on storage at CNAF. The proposed solution is to share a GPFS volume from CNAF using the AFM cache feature (see figure 11).

6. The power line incident

On August the 27th, one of the two main power lines, upstream to the continuity system, burnt due to a water leakage from the ceiling (see figure 12). The fire was immediately extinguished by the automatic system and the power on the line was secured by the diesel engine (the other line continued to work normally). To save gasoline, farm nodes were powered only by the other line. After almost 4 days the power line was partially restored without continuity and only at the end of September it was completely restored. We had only a disservice on the storage system for LHCb due to a bug in the firmware of the controllers not properly able to manage the power

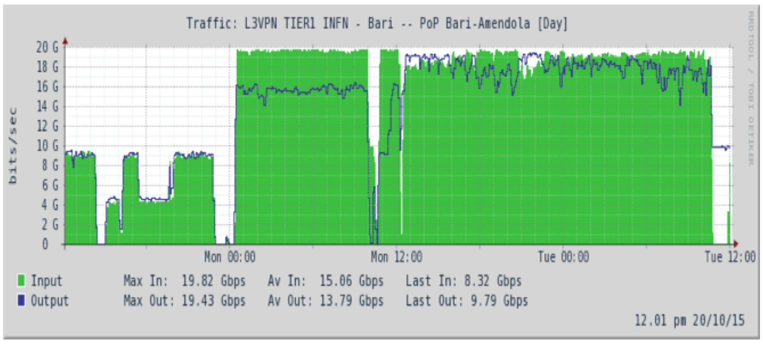


Figure 10. Commissioning of dedicated link CNAF-Bari

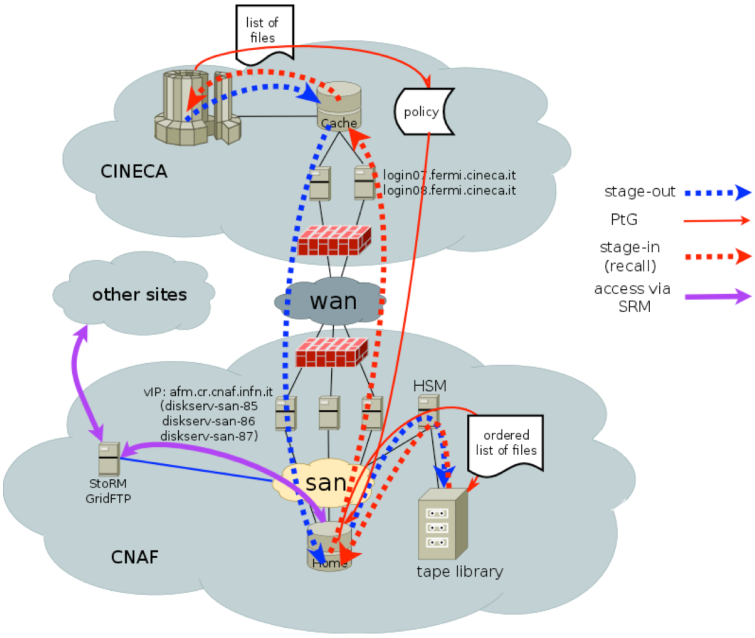


Figure 11. Layout of the test-bed CINECA-CNAF for distributed storage

loss of one line; all the other systems and services were not affected. We experiences some minor issues (e.g. systems not properly cabled) which were solved without interfering with the normal operativity of the center.

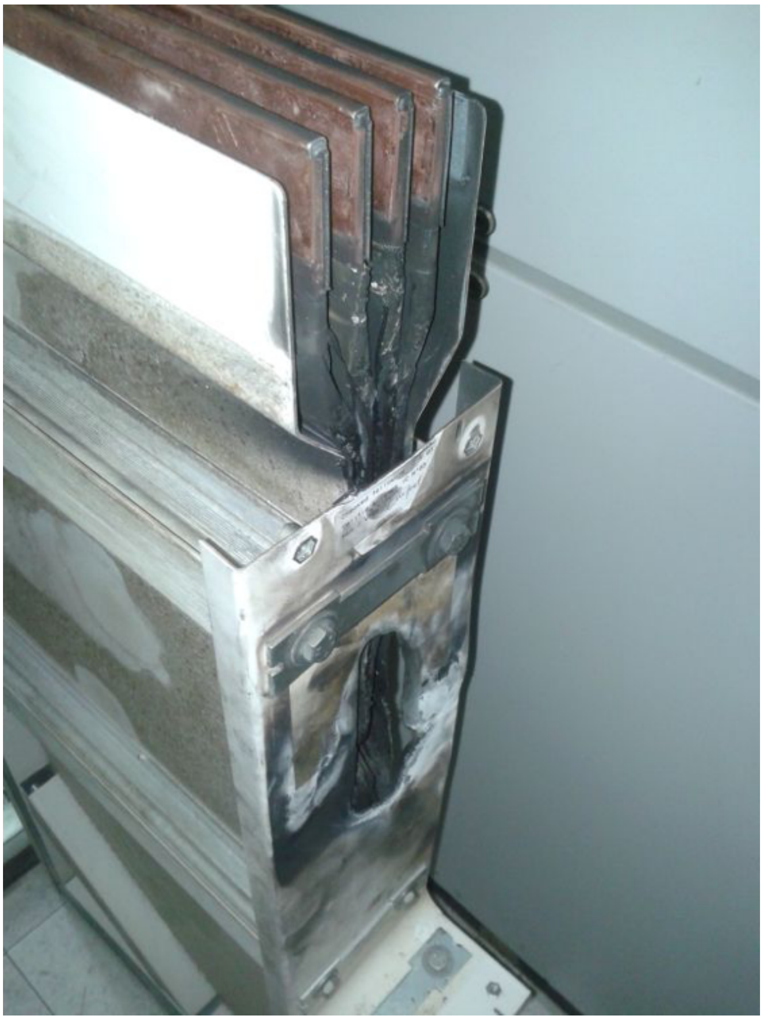


Figure 12. Part of the main power line after the incident

The INFN-Tier1: the computing farm

A Chierici, S Dal Pra and S Virgilio

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 grid interfaces, CE and site BDII). This implies the deployment of installation and configuration services, monitoring facilities and to fairly distribute the resources to the experiments that have agreed to run at CNAF.

2. Farming status update

Farm computing power in 2015 was set to 187.500 HS06, since we decommissioned old unsupported hardware. We acquired 2015 hardware through a government public tender: this time we got blade servers with Intel CPU. The computing power provided by one of these machines is roughly double the one provided by last year's, allowing us to improve rack space occupancy and power consumption. After summer holidays, some experiments asked for an extra pledge of roughly 30.000 HS06 hence we had to reconnect some old machines, that still perform well. Currently the extra pledge is still in place and the total power given to the experiments is 221.500 HS06. During this year we made significant improvements to the service infrastructure, that can be summarized in this list:

- Upgrade LSF to version 9
- Migrate from quattor to puppet+foreman
- Acquire new hardware for virtualization system (currently still to be fully deployed)
- Update the monitoring and alarming infrastructure through a newly created working group (BeBop)

We will describe the activities that have already been completed in more details in following sections of this paper.

2.1. Migration to LSF v9

LSF is the production batch system at our Tier1. We rely on it since the very beginning of the project (excluding a short period with torque+maui) and it proved to be rock solid and fully featured. Even if during the years new releases came out, we haven't felt the need to do any major upgrade, since no new requirements were pushing us to do so. During 2015 we were facing the request to support *linux cgroups*[1], possibly GPUs and were also reaching some scalability limits, that pushed us to upgrade LSF to the latest version, number 9.

2.1.1. The migration Upgrading the batch system required us to basically split our computing farm in two, one part with legacy LSF version and one with the new one. Initially the legacy partition was the whole farm, but gradually, migrating the nodes to the new one, it decreased until becoming empty. Before we could start, we had to create on the NFS shared area (required by all LSF installations) a new lsf9 subtree, with all the software and configuration files in place, and a Computing Element supporting the new name space. This has been the major problem since Tier1 users got confused by the fact that different CEs were seeing different nodes with different total computing power. In general the migration process of the Worker Nodes consisted of these steps:

(i) draining of a rack of nodes

(ii) changing the configuration, pointing to the new shared area, running LSF configuration script

(iii) reconfiguring the middleware

(iv) opening the node in the new cluster

2.2. Migrating provisioning from quattor to puppet and foreman

Configuration management is a well known task in a sysadmin every day life. Several servers in the computing room have to be installed, configured and patched when necessary. This task requires a lot of human effort, with repetitive actions to be taken on every node. Up until 2015 we solved this problem with quattor, a tool developed inside the DataGrid project that proved to be scalable and effective. The end of the project left the effort in development and maintenance basically up to the will of a group of few people; meanwhile the emerging of new tools, well documented and adopted both in research institutions and private companies, pushed us to the change. Among the possible solutions, we adopted **puppet**[2] and **foreman**[3], that were chosen not only by tier1 staff (like quattor) but also by other CNAF departments, like Information Services and R&D. For a detailed description of the infrastructure and the implementation choices see the corresponding article in this report. For what concerns farming department, we can say that the migration was smooth and did not cause any problem. All the nodes are managed by foreman through puppet; moreover we integrated our existing *ovirt*[4] infrastructure in order to be able to deploy virtual machines directly through the foreman web interface. Some legacy nodes still rely on quattor and we decided not to put any effort in their migration, because during 2016 they will all be decommissioned.

2.2.1. The migration During this year we converted all quattor profiles to puppet manifests, reinstalling the nodes where possible. The transition has been smooth and allowed us to review and remove stale configuration that were not used. These new tools made it possible to share among different CNAF departments common configurations, saving a lot of work. Now CNAF has a unique, full fledged and powerful installation and configuration system, well documented and supported by a large community of users and developers, giving us the confidence of a future-proof choice.

3. Low power solutions: advanced tests

During 2015 we carried on tests on low power CPU solutions, with Intel Avoton powered servers. Thanks to the collaboration with some vendors we were able to test “on premises” a Supermicro Microblade server and a HP Moonshot.

3.1. Supermicro Microblade

The microblade is a surprisingly compact solution: a single blade (half-size compared to standard blade solutions) carries on 4 separate Avoton motherboards each with a 2,5” disk. A single chassis can host 28 such blades, and this means that in 6 rack units (the size of a single chassis) one can host 112 Avoton CPUs, providing roughly 5k HS06. During the evaluation period at CNAF (unfortunately we received only a single blade, with 4 CPUs) we have had the opportunity to deeply test this unit, integrating it seamlessly into the farm, without any particular configuration or tweak: Avoton CPUs implement standard x86 architecture, allowing all the binaries to run without any recompilation. With only a single blade, it was difficult to separate its power consumption from the one taken by the chassis (which takes power for fans and network switch): for the same reason we felt not confident in evaluating a TCO. Anyway we can state for sure that power consumption of this kind of server is extremely low compared to standard server class CPUs, and that this solution looks promising, particularly taking into account the possibility to install standard xeon DP blades, or even to mix Avoton and Xeon blades.

3.2. HP Moonshot with external storage

In 2015 we tested HP Moonshot mounting m350 cards, in a real production environment. Last year we got from HP a Moonshot chassis with a few cards inside, but since the nodes have only a small flash disk it was impossible for us to test a real production job. After a short discussion with HP during supercomputing 2014, we got the possibility to do a further test, this time using an external disk server, providing enough storage space to run production jobs.

3.2.1. Hardware and software setup The hardware configuration we tested is the following:

Moonshot the Moonshot server was configured with a 10Gb/s switch, redundant fans and power supplies, and carried 15 m350 cards, each with 4 separate Avoton C2730 nodes. Every node comes with 16GB of ram and 64GB of disk space (running on a m.2 ssd).

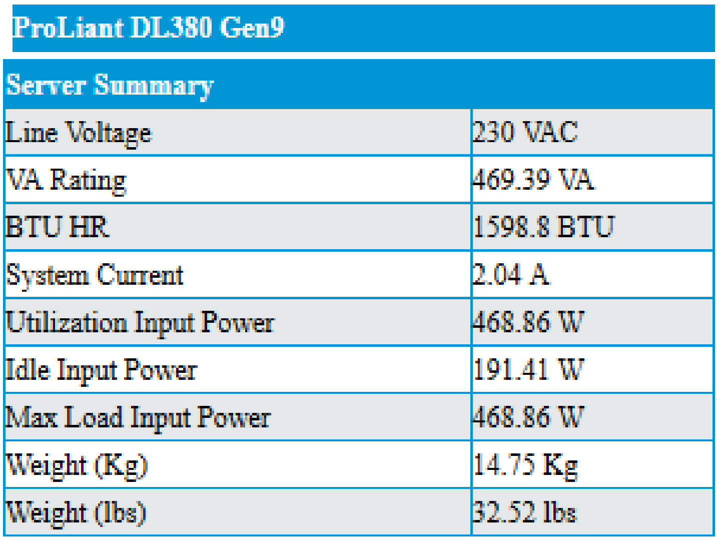
External Storage provided by a HP dl380 gen9 disk server, a 2U dual xeon server with 128GB of ram, mounting 24 SAS disks for a total of 20TB net space. The disk space was provided to the Moonshot via the iscsi protocol, serving 300GB per node.

The external storage was connected through a 10GB/s link directly to the moonshot switch, avoiding any latency. The moonshot was then connected to our core switch through a second 10GB/s link.

This setup allowed us to integrate all the 60 Avoton CPUs in the farm to perform a proper stress test (compared to the one done the last year) and to use them as standard worker nodes. We were particularly interested in understanding the load created on the external storage by the nodes and being able to evaluate the maximum number of worker nodes that a single external storage can handle. Given the peculiar characteristics of the Avoton CPU, we inserted the nodes into the multi-core queues, for both CMS and Atlas.

3.2.2. Power consumption To analyse the power consumption of the solution under test, we used a Voltech PM 300.

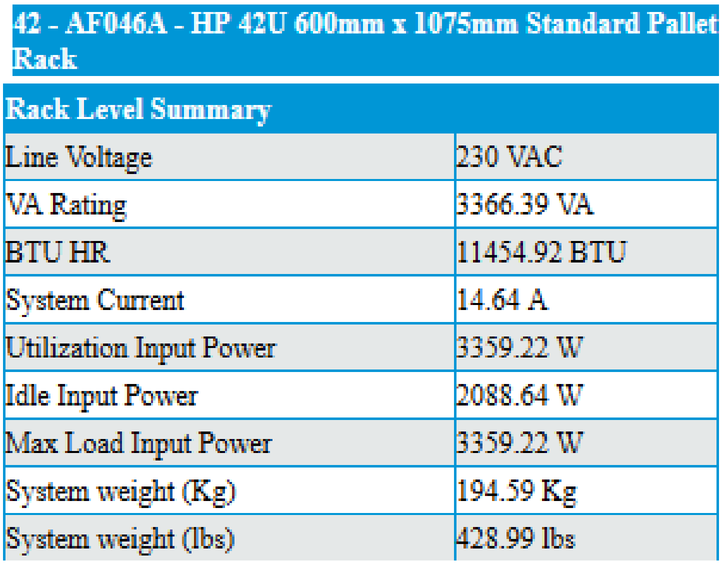
In figure 1 we see that the nominal power consumption of the external storage (provided by HP), under full load is 469W, with an idle of 191W. Figure 2 shows that a Moonshot chassis with all 45 cards takes up to 3360W, with an idle of 2088W (again, measure provided by HP). The unit under test had only 15 cards inside and the power consumption we measured was 1045W: this measure is totally compatible with what HP declares, taking into account switches and fans, that are part of the chassis and that take some Watts, no matter how many cards are inserted. Considering that a Moonshot fully populated sports 8190 HS06, a single chassis, with an external storage can provide 2,13 HS06/W. As a comparison, 2014 tender machines provide only 0,62 HS06/W, while 2015 nodes provide 1,4 HS06/W.



ProLiant DL380 Gen9

Server Summary	
Line Voltage	230 VAC
VA Rating	469.39 VA
BTU HR	1598.8 BTU
System Current	2.04 A
Utilization Input Power	468.86 W
Idle Input Power	191.41 W
Max Load Input Power	468.86 W
Weight (Kg)	14.75 Kg
Weight (lbs)	32.52 lbs

Figure 1. Power consumption of dl380 disk server



42 - AF046A - HP 42U 600mm x 1075mm Standard Pallet Rack

Rack Level Summary	
Line Voltage	230 VAC
VA Rating	3366.39 VA
BTU HR	11454.92 BTU
System Current	14.64 A
Utilization Input Power	3359.22 W
Idle Input Power	2088.64 W
Max Load Input Power	3359.22 W
System weight (Kg)	194.59 Kg
System weight (lbs)	428.99 lbs

Figure 2. Power consumption of the Moonshot under test

3.2.3. External storage load A key point of the solution under test is the external storage. During this evaluation we had 2 questions to answer:

(i) How many Moonshot cards can a single external storage support?

(ii) Is an external storage fast enough to support the needs of our jobs?

To answer the first question we monitored the disk server for different loads, given by increasing the number of clients (the computing nodes) accessing the storage.

Figure 3 shows the resources utilized on the disk server while 15 clients were active. CPU is practically idle, memory is mostly allocated by Operating System and the interesting measure

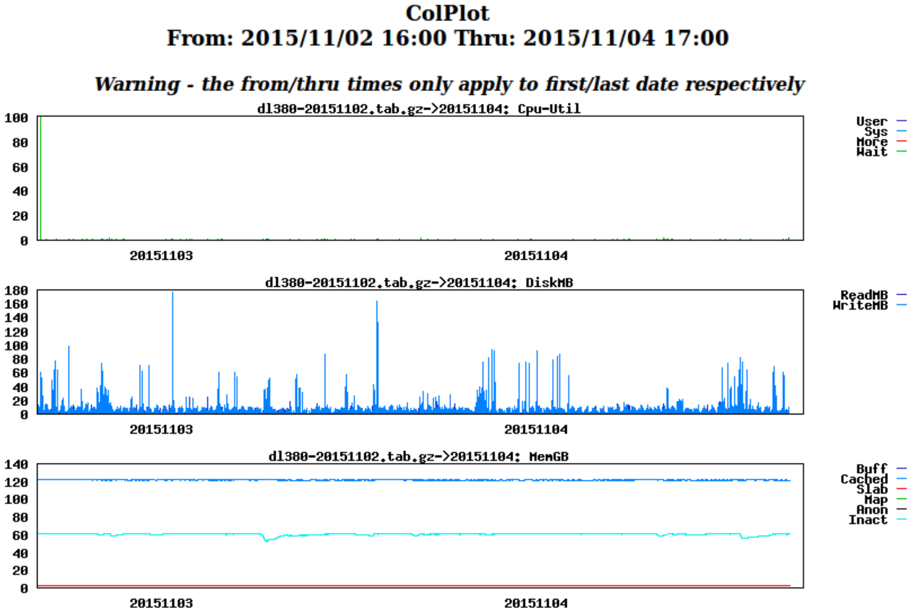


Figure 3. External storage load with 15 clients

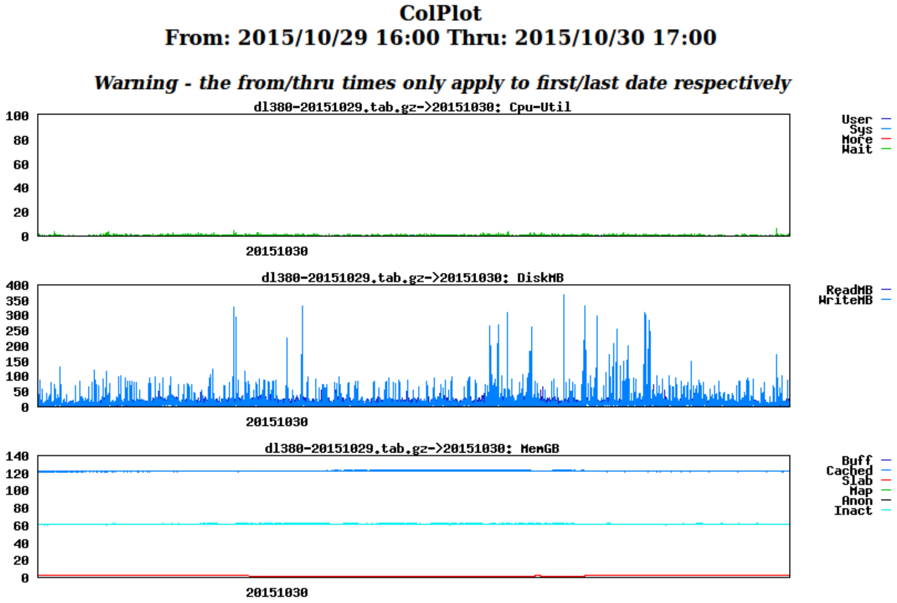


Figure 4. External storage load with 60 clients

comes from the disk I/O. This case shows that the average throughput is approximately 20MB/s, a value that does not concern.

In figure 4 we see the resources utilized while all the 60 clients are active. Even in this case, CPU is almost idle, memory still allocated by OS and the disk I/O average is below 50MB/s. Even this value is far from any concern, and allows

Queue	N.Jobs	WN Type	CPT_Days	WCT_Days	Eff
cms_mcore	60	moonshot	161,12	225,87	0,71
cms_mcore	69.190	standard	194.567,92	245.094,03	0,79
mcore	1.086	moonshot	1.295,16	1.454,90	0,89
mcore	127.455	standard	58.815,11	73.242,69	0,80

Table 1. Job efficiency comparison on cms and atlas multicore queues

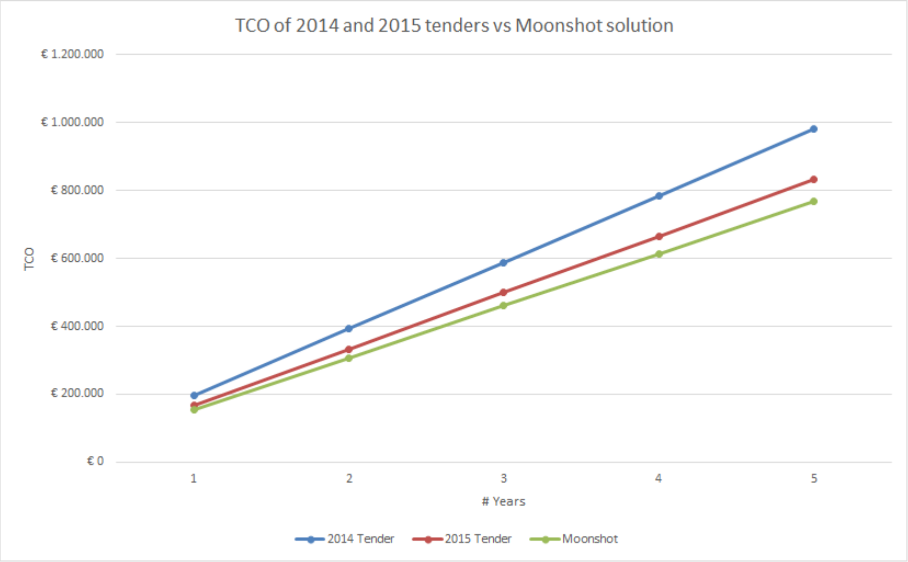


Figure 5. TCO Comparison

us to say that this server is capable of supporting 60 clients without any bottleneck. We can go further and say that a single disk server should be powerful enough to support a chassis with all the 45 cards inside, that means having 180 clients.

To answer the second question we monitored the data produced by our accounting system and noticed that the job efficiency is comparable to the one calculated for "standard" nodes. Table 1 shows the data.

As it can be clearly seen in last column, the efficiency of jobs running on Avoton nodes is comparable to the one calculated on standard nodes. The sample is rather small compared to standard nodes due to the fact that we tested these nodes only for a short period of time and also to the fact that even if with high efficiency, jobs on Moonshot nodes take longer time to end compared to standard nodes.

3.2.4. TCO Evaluation Given the fact that Moonshot nodes performances are comparable to standard nodes, we need to evaluate the TCO in order to understand if this solution may be a valid solution to the future challenges that our Computing Center must face. We evaluated the TCO of acquiring and maintaining 30k HS06, using the prices of the last 2 tenders (2014 and 2015), with this Moonshot solution. To be fair we have to admit that the last two tender prices were particularly good for us and that the purchase price of the Moonshot was just suggested

Years	2014	2015	Moonshot
1	€ 0	€ 29.574	€ 42.779
2	€ 0	€ 59.148	€ 85.558
3	€ 0	€ 88.722	€ 128.336
4	€ 0	€ 118.295	€ 171.115
5	€ 0	€ 147.869	€ 213.894

Table 2. TCO: savings compared to 2014 tender

by HP, not a tender price: this implies that the result could be a little biased by this and has to be taken into account, during evaluation. The TCO parameters include the cost of 0,2 €/kWh, the PUE of 1,6 of our center, and the desired life of the solution, that must be **at least 4 years**. The TCO of each solution is summarized in figure 5. As it can be seen, Moonshot solution is immediately cheaper than any other and gets better and better. See table 2 for details on figures.

3.2.5. Conclusion We proved that the solution proposed by HP "may" be convenient for us, allowing to save quite a lot of money. Anyway there is a major problem to this solution: the storage server suggested by HP is not in any way a reliable server and any fault on it would mean that all the moonshot nodes (worst case: 180) may loose the disk, this implying the loss of the computing job and of the whole node. Changing the server, adopting a more reliable one, implies bigger purchase costs, making the TCO not so different to the 2015 tender solution. At this point the administrative effort required to manage 7200 nodes instead of just 100 (considering the small number of HS06 provided by Avoton CPUs) does not seem so attractive. For this reason the evaluation of this solution is only partially positive, and requires a deep evaluation of the requirements and the possible scalability issues.

4. References

[1] cgroups on wikipedia: https://en.wikipedia.org/wiki/Cgroups
[2] Puppet webpage: http://puppetlabs.com
[3] Foreman webpage: http://theforeman.org
[4] Ovirt webpage: http://www.ovirt.org
[5] HP Moonshot system Website: http://www8.hp.com/us/en/products/servers/moonshot/

The INFN-Tier1: networking

S Zani, D De Girolamo and L Chiarelli

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 with two main physical links (see Fig. 1):

- General IP link has been upgraded to 20 Gb/s via GARR and GEANT
- The link to WLCG destinations has been upgraded to 40 Gb/s link shared between the **LHCOPN** Network for Tier0–Tier1 and Tier1–Tier1 traffic and **LHCONE** network mainly for T2 and T3 traffic.

2.1. Network capacity

The network capacity usage is constantly growing on General IP and on LHCOPN.

Fig. 2 shows the usage of the combined LHCOPN and LHCONE link. The average input traffic was 7.0 Gb/s, with a maximum of 34.6 Gb/s. The avarage output traffic was 4.5 Gb/s, with a maximum of 37.3 Gb/s.

Fig. 3 shows the usage of the general IP link. The average input traffic was 1.4 Gb/s, with a maximum of 15.1 Gb/s. The avarage output traffic was 1.8 Gb/s, with a maximum of 16.5 Gb/s.

2.2. Network capacity evolution

The General IP link has been upgraded to 20 Gb/s in 2015.

For the future the WLCG link can be upgraded any time to 60 Gb/s (6 × 10 Gb/s) and GARR has in its roadmap a 100 Gb/s link between GARR POP at CNAF and GEANT (the fibers are 100 Gb-ready but an evolution on GARR optical devices is needed).

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 LHCOPN 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.

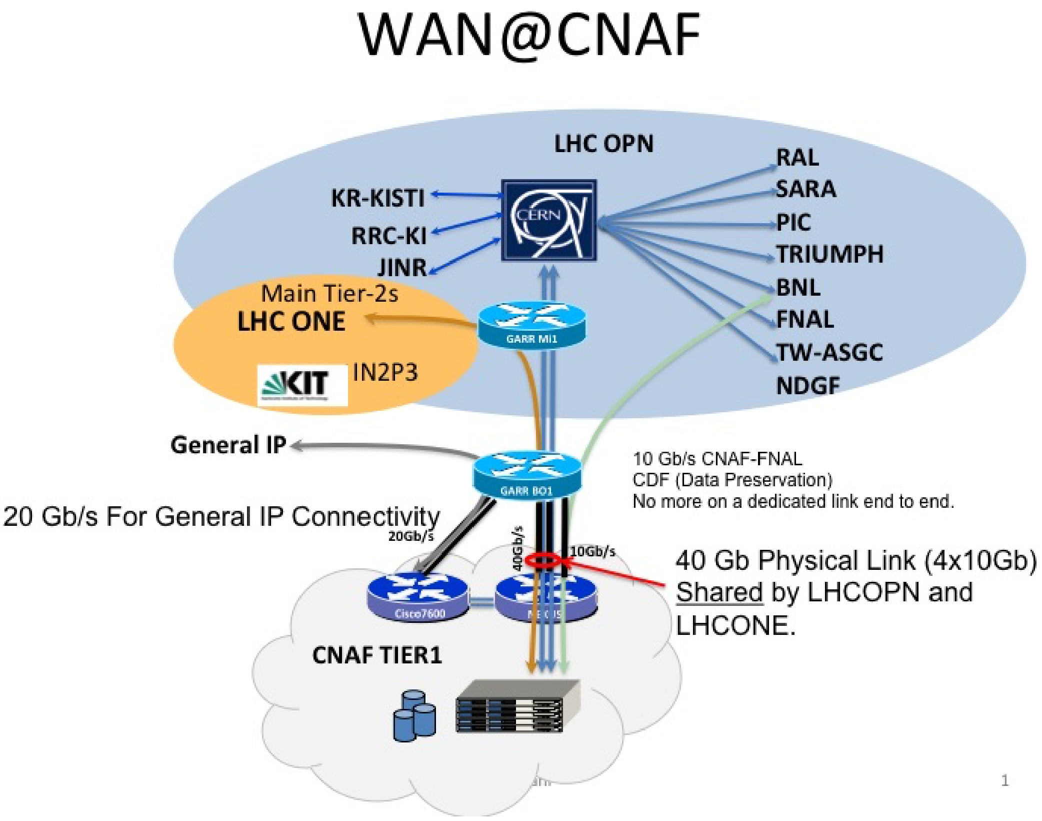


Figure 1. WAN connection schema

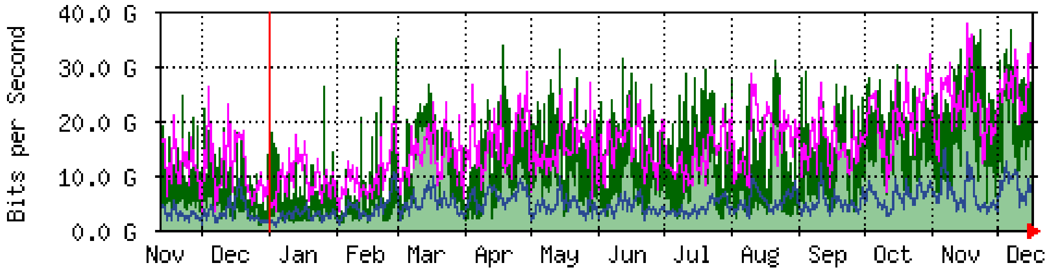


Figure 2. LHCOPN + LHCONE usage

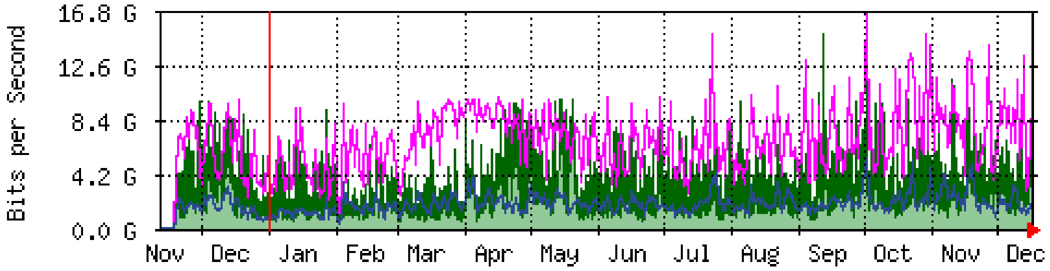


Figure 3. General IP link usage

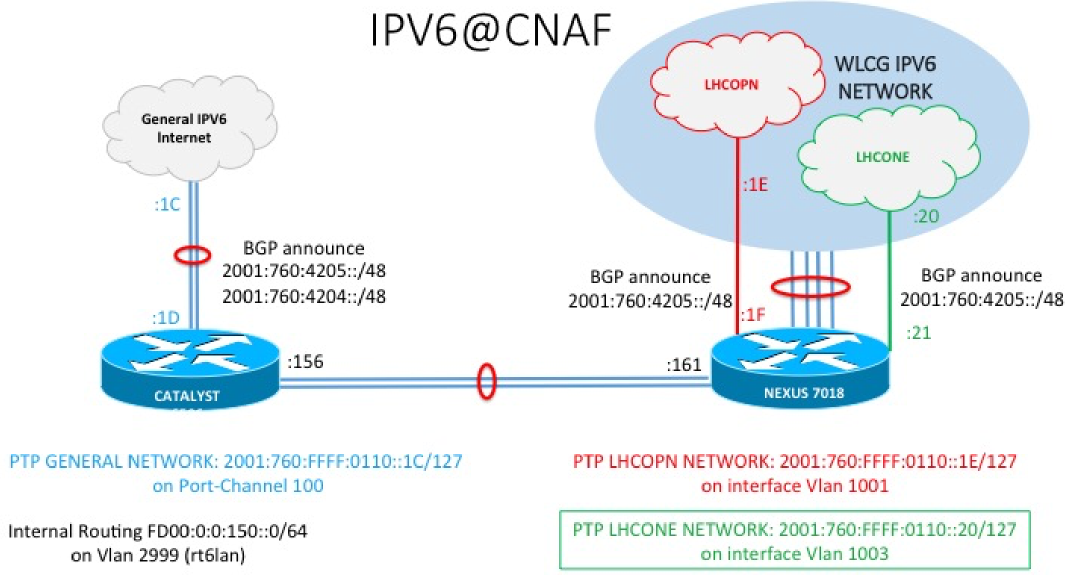


Figure 4. Implementation schema of IPv6 at CNAF

Disk-servers and GridFTP servers are directly connected to the core switch at 10 Gb/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 Cisco6506 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.

4. IPv6

Two /48 IPv6 prefixes are assigned to CNAF (2001:760:4204::/48 for CNAF General and 2001:760:4205::/48 for CNAF WLCG).

The IPv6 Infrastructure, shown in Fig. 4 has been implemented in March 2015 in the LHCOPN/ONE infrastructure and in the General IP Network.

The first dual stack production nodes are the perfSONAR servers:

perfsonar-ps.cnaf.infn.it has address 131.154.254.11

perfsonar-ps.cnaf.infn.it has IPv6 address 2001:760:4205:254::11

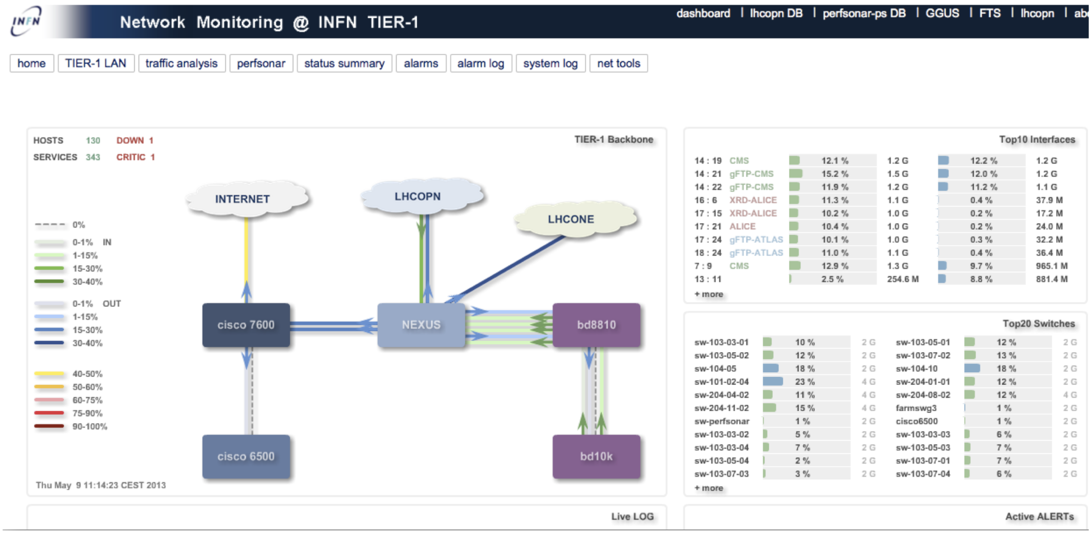
perfsonar-ow.cnaf.infn.it has address 131.154.254.12

perfsonar-ow.cnaf.infn.it has IPv6 address 2001:760:4205:254::12

5. Network monitoring and security

In addition to the perfSONAR-PS and the perfSONAR-MDM infrastructures required by WLCG, the monitoring system is based on several tools organized in the “Net-board”, a dashboard realized at CNAF (see Fig. 5). The Net-board integrates MRTG [2], NetFlow Analyser [3] and Nagios [4] 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.



HOSTS	130 DOWN 1
SERVICES	343 CRITIC 1

Thu May 9 11:14:23 CEST 2013

Top10 Interfaces

14 : 19	CMS	12.1 %	1.2 G	12.2 %	1.2 G
14 : 21	gFTP-CMS	15.2 %	1.5 G	12.0 %	1.2 G
14 : 22	gFTP-CMS	11.9 %	1.2 G	11.2 %	1.1 G
16 : 6	XRD-ALICE	11.3 %	1.1 G	0.4 %	37.9 M
17 : 15	XRD-ALICE	10.2 %	1.0 G	0.2 %	17.2 M
17 : 21	ALICE	10.4 %	1.0 G	0.2 %	24.0 M
17 : 24	gFTP-ATLAS	10.1 %	1.0 G	0.3 %	32.2 M
18 : 24	gFTP-ATLAS	11.0 %	1.1 G	0.4 %	36.4 M
7 : 9	CMS	12.9 %	1.3 G	9.7 %	965.1 M
13 : 11		2.5 %	254.6 M	8.8 %	881.4 M

+ more

Top20 Switches

sw-103-03-01	10 %	2 G	sw-103-05-01	12 %	2 G
sw-103-05-02	12 %	2 G	sw-103-07-02	13 %	2 G
sw-104-05	18 %	2 G	sw-104-10	18 %	2 G
sw-101-02-04	23 %	4 G	sw-204-01-01	12 %	2 G
sw-204-04-02	11 %	4 G	sw-204-08-02	12 %	4 G
sw-204-11-02	15 %	4 G	farmswg3	1 %	2 G
sw-perfsonar	1 %	2 G	cisco6500	1 %	2 G
sw-103-03-02	5 %	2 G	sw-103-03-03	6 %	2 G
sw-103-03-04	7 %	2 G	sw-103-05-03	7 %	2 G
sw-103-05-04	2 %	2 G	sw-103-07-01	7 %	2 G
sw-103-07-03	3 %	2 G	sw-103-07-04	6 %	2 G

+ more

Figure 5. Net-board screenshot

The network security policies are mainly implemented as hardware-based ACLs on the access router and on the core switches (with 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 credentials and known vulnerabilities of software and Grid middleware) cooperating with the involved parties.

6. Dynfarm: Dynamic farm extension

Dynfarm is a new software product developed in 2015. Its aim is to allow seamless integration into the existing Tier1 farm of computing resources physically located outside of the computing center. It thus allows to dynamically increase the number of resources available to better cope with times of peak usage, while putting only minimal requirements on the remote nodes themselves.

It has been used in production the first time on resources provided by Aruba (one of the main Italian public Cloud provider) running CMS jobs.

This project has been developed in a collaboration between R&D and Network Department (*Vincenzo Ciaschini* and *Donato De Girolamo*).

7. 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)

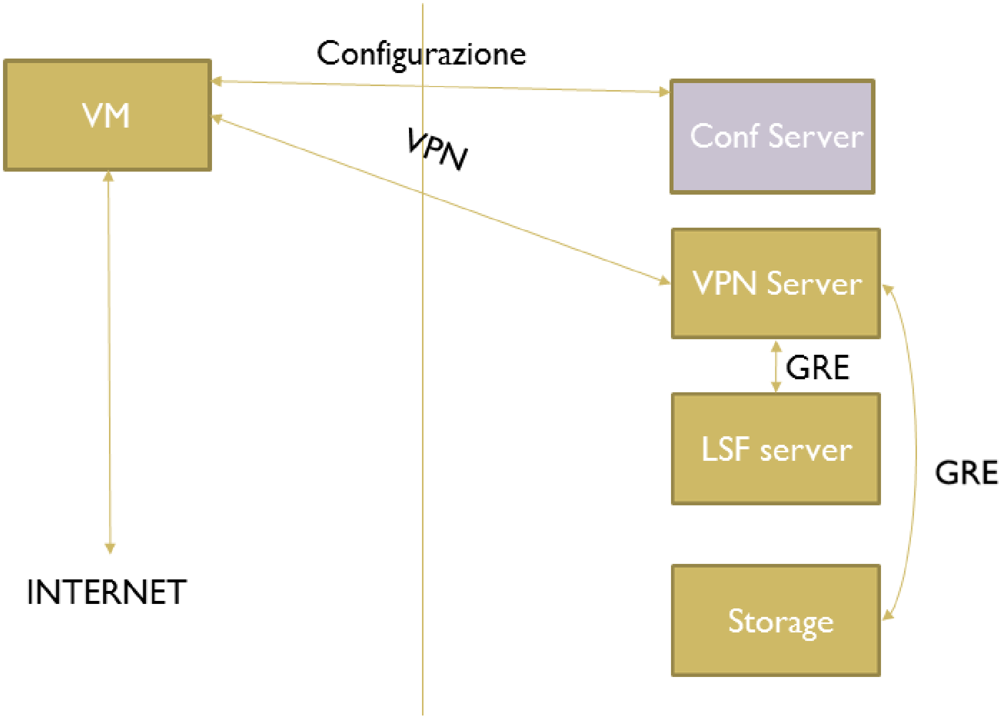


Figure 6. The Dynfarm architecture

Elastic CNAF DataCenter extension via opportunistic resources

S Dal Pra, V Ciaschini, L dell'Agnello, A Chierici, D De Girolamo, V Sapunenko and T Boccali

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

1. Introduction

During spring 2015, the CMS Experiment[2] support and the Tier–1 team at CNAF took contacts with Aruba[1], a major Italian host, web and cloud provider, at its site in Arezzo (Tuscany, Italy). The goal was to evaluate the possibility of remote computing resource usage from a commercial provider, in order to meet "elastically" the needs of High Energy Physics (HEP) computing, allowing for opportunistic resource utilisation as a transparent extension of CNAF during periods of time where activity from ordinary Aruba users is low. In the proposed solution, no check-pointing or failure handling is required on the user side. CMS has been used as initial test case, with the extension to all the LHC Experiments already ongoing.

2. The initial test case

In order to test the elastic setup in a real production environment, the CMS Experiment[2] has been selected. CMS is one of the biggest HEP users at CNAF, with a 2016 resource allocation (from Rebus[3]) of 48 KHS06 of CPU power, 4 PB of online Disk storage and 12 PB on Tape. CMS was selected as a CNAF test case mainly given the high level of local expertise.

The computational tasks CMS executes at CNAF cover various aspects of its activities, from Monte Carlo simulation and data reconstruction, to end–user analysis; the submission infrastructure uses pilots from glideinWMS[4] with late binding, with the site not being able to predict the type of running job before it starts.

The typical CMS workflow needs some basic requirement in order to succeed and, from the perspective of the Tier–1 it proceeds as follows:

- GlideinWMS submits *Grid pilot jobs* to one or more Computing Elements at the site, which in turn submit them to the Local Resource Manager System
- Each production CMS pilot submitted to INFN–T1 is currently a multi–core one, thus requiring to run on a Worker Node providing eight free slots and eight times more RAM than a single–core one (16 to 20 GB).
- The Batch System dispatches the pilot job to a suitable Worker Node (WN) and after the pre–execution check on the machine succeed, it starts and access the actual processing task as dispatched via HTCondor[5]; the executable and all the required libraries are distributed via CVMFS[6] and available through a local SQUID proxy;
- it then access conditions database, also available through a SQUID proxy;
- it access its input ROOT[7] files.

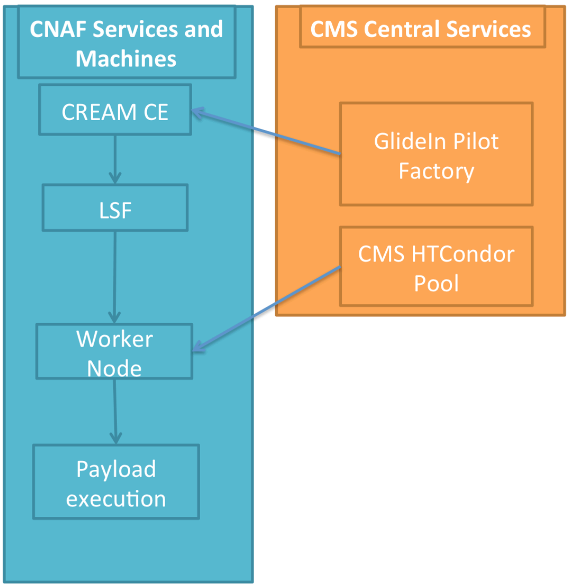


Figure 1. Interactions between CMS Central Services and CNAF Site when a pilot job arrives.

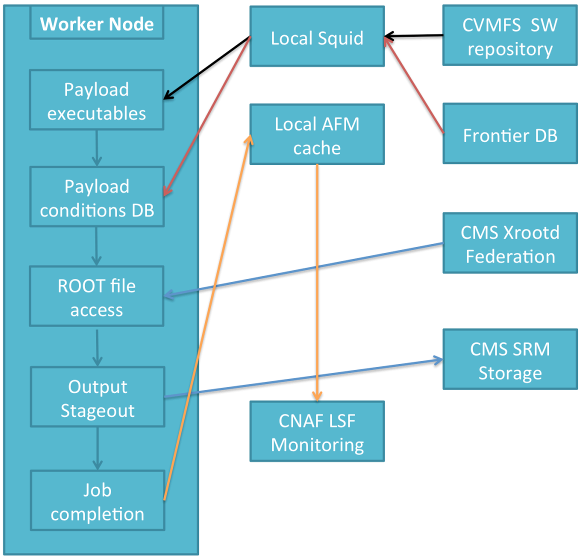


Figure 2. Interactions between the Worker Node and CMS services when an executable is started.

- stage output files and logs to a SRM enabled storage supporting CMS.

The interactions between the CMS Central Services and CNAF are sketched in Figure 1. Figure 2 shows the interactions needed by a CMS payload when running.

CMS differs from other experiments running at CNAF in that it exclusively submits multi–core pilots.

When a standard CMS job is run at CNAF, the prerequisites are satisfied by on–site services; namely, SQUIDs, LSF and Grid Computing Elements (CE) are accessible through the LAN, and the input files are stored on the locally mounted GPFS[8] filesystem. However, when the job runs on a remote instance, it is preferable to have some services (SQUIDs for Condition Database access and CVMFS) replicated and made locally available, in order to prevent excessive latencies and large geographical network I/O.

For what concerns the input ROOT files, LHC Experiments implement an internal mechanism which falls back to remote access using Xrootd Federations in case the local access fails. Using that, the input file is searched in all the CMS computing centres[10], and if successfully found, a direct remote read is performed. This failover mechanism has been optimised for the elastic extensions: the Xrootd services at CNAF are queried before other ones in the Federation.

3. Approach to opportunistic remote extension

The simplest way to extend CNAF Computing Center is making remote resources be seen as local. To achieve this several problems had to be overcome:

- The IPs of the remote VMs only have a private IP and are behind NAT and Firewall systems.
- LSF needs direct and reverse resolution (map hostname to IP and vice–versa) of its clients;
- all client nodes in a cluster need read access to a shared filesystem, which has to be provided over the general network, with all the associated network latency problems. This filesystems

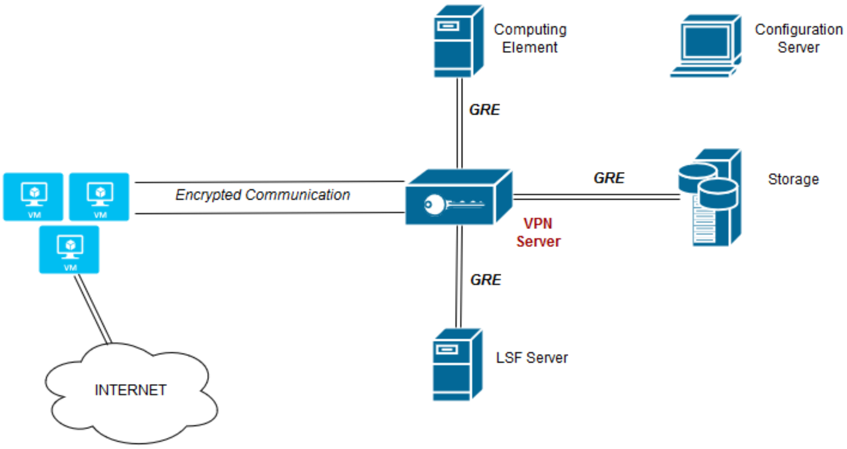


Figure 3. The dynfarm communication flow.

hosts LSF work area, and some experiment specific directories (those hosting site specific configuration for jobs).

In order to solve these problems, and to automate the process of deploying and configuring virtual worker nodes on remote machines a peice of software, *dynfarm*, has been internally implemented and used.

In its configuration at Aruba, the virtual machine, upon start-up, connect to the dynfarm server at CNAF. It authenticates connection requests coming from remote hosts and delivers the information needed in order to establish a VPN tunnel, used to communicate with a small subset of the local cluster in Bologna: the LSF batch system, the CREAM Computing Elements and the Authorization (Argus) service. This solves the first problem, since from the VM side the tunnel is an outbound connection that passes through the firewall; the fact that remote hosts get new addresses through the VPN makes them distinguishable even through a NAT.

In order to solve the second problem (remote hosts need a hostname discoverable by the LSF master), the `$LSF_ENVDIR/hosts` file is used. This file works just like the well known `/etc/hosts` but it is only used by LSF. By configuring that, we can simply map in advance a superset of known private IPs and hostnames.

Finally, in order to give each node access to the required shared filesystem, a read-only GPFS cache using the AFM plugin is deployed on a dedicate host at the Aruba computing centre.

Figure 3 shows how dynfarm interacts with local and remote services when the VPN is set-up.

4. Another remote setup: the case of Bari-ReCaS

The CNAF Tier-1, has also worked in the direction of the utilisation of remote resources at ReCaS[11], in BARI. The ReCaS project, financed by the MIUR (Italian Ministry for Education, University and Research) has deployed a large resource base in a computing center at the Physics Department in Bari, very close to the INFN Tier–2. The utilisation pattern is simpler than the one tested with Aruba:

- the machines are not virtualised. Once the bare metal is correctly deployed, connected to network and power lines, it is directly managed by CNAF staff both via IPMI and ssh connections.
- the machines have public and private (used for ipmi) IPs belonging to CNAF, and all the outgoing traffic is routed to CNAF via a Level 3 circuit, with a reserved bandwidth of 20 Gbit/s. No host-level VPN needs to be set-up;

- local squids are present on Site and managed by CNAF staff;
- online read access to GPFS disk is available through AFM replica. This has been set up for three LHC Experiments: ATLAS, CMS, LHCb with an overall disk size of 330TB. ALICE does not need it since it directly accesses Xrootd;
- The cluster shared filesystem (hosting configurations, LSF work area, ...) is also replicated via AFM and locally made available for read-only access to the WNs via NFS. Even thought this is a small filesystem with low traffic, Care have to be taken to keep its access latencies at a minimum: heavy data access should not interfere with the cluster management traffic, which depends on the readiness of the shared filesystem.

From the management point of view, these resources are seen by CNAF as local machines, are inside the LAN and seamlessly connected to LSF. No problem with firewalls neither a problem for host level VPNs exists.

Currently the "remote" cluster has roughly 2000 cores, and accepts jobs only by LHC Experiments, since only these have an Xrootd fallback mechanism in place.

5. Results

Results presented here for the Aruba testbed are the results of a few weeks of running.

mese	*queue*	*site*	*njobs*	*avg_eff*	*max_eff*	*avg_wct*	*avg_cpt*
2015-10	cms_mc	AR	2793	0.609	0.912	205.823	134.606
2015-10	cms_mc	T1	34342	0.713	0.926	120.886	96.147
2015-11	cms_mc	AR	232	0.752	0.871	320.182	241.498
2015-11	cms_mc	T1	22151	0.805	0.886	177.937	147.926
2016-01	cms_mc	AR	459	0.212	0.656	50.865	10.125
2016-01	cms_mc	T1	18577	0.774	0.852	122.470	96.862
2016-02	cms_mc	AR	2698	0.438	0.713	163.304	67.627
2016-02	cms_mc	T1	16367	0.793	0.868	146.713	117.404

Table 1. Comparative efficiency and duration of CMS jobs per site, by month. The average efficiency tends to be higher with longer jobs

The testbed Aruba has prepared for INFN is in the form of a VMware Virtual Data Center, where a given amount of RAM, Disk and CPU is offered. Its partitioning within different hosts is completely under CNAF control; in particular, the testbed allowed us to use 160 GHz of CPU (Intel 2697-v3). Assuming 2GHz cores, this enables in principle 80 active cores; in reality, more can be obtained, since the cap is not on allocated cores, but on used CPU. So, a larger WN base can be installed, and if at some point the total CPU utilization would surpass the cap, an automatic clock slowdown would happen, with the same mechanism used when opportunistic usage must be reduced. Figure 4 show a situation where cap is reached, and total CPU utilization is kept under 160 GHz.

Figure 5 shows a typical 1-week load result, where each different color represents a virtual host. Table 2 displays comparative information about efficiency (defined as $\sum \frac{CPUtime}{NUMcores \times RUNtime}$) and duration of finished jobs at the different *sub–sites* over a time period of a month.

6. Conclusions

The studies presented in this paper, even if at a level of test, show that an extension of a large WLCG Data Center like CNAF with remote resources is possible, either via proprietary resources

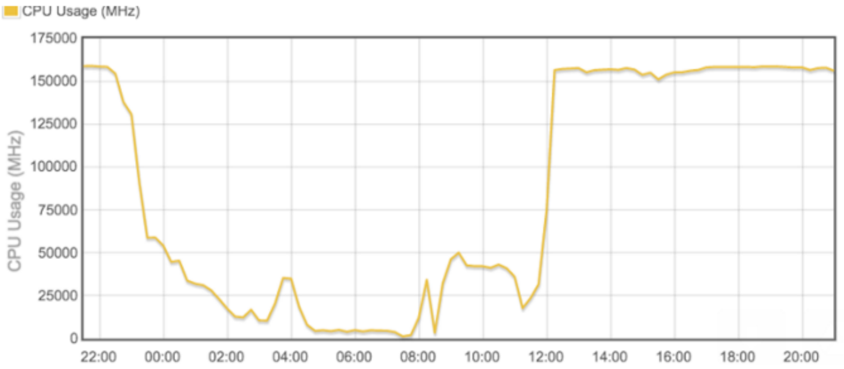


Figure 4. CPU capping at 160 GHz, when high load is present.

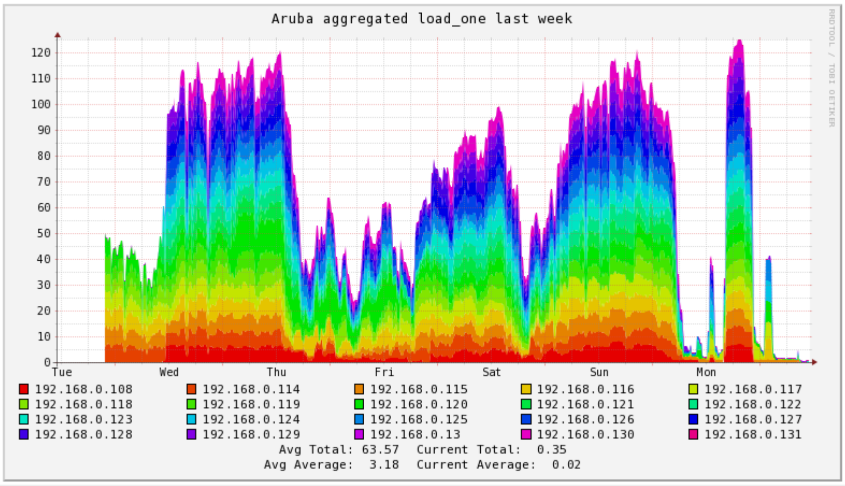


Figure 5. Load, one week. Each color represents a single Virtual Worker Node.

month	*queue*	*site*	*njobs*	*avg_eff*	*max_eff*	*avg_wct*	*avg_cpt*
2016-02	alice	BA	23153	0.725	0.966	16.071	11.765
2016-02	atlas	BA	451	0.942	0.999	2.833	2.721
2016-02	cms_mc	BA	287	0.516	0.865	388.130	194.689
2016-02	lhcb	BA	11269	0.916	0.967	15.319	13.082
2016-02	mcore	BA	23441	0.674	0.851	155.448	36.344
2016-02	cms_mc	AR	2698	0.438	0.713	163.304	67.627
2016-02	alice	T1	184292	0.697	0.958	15.936	11.284
2016-02	atlas	T1	1182474	0.823	0.999	2.219	1.936
2016-02	cms_mc	T1	16367	0.793	0.868	146.713	117.404
2016-02	lhcb	T1	78769	0.970	0.989	18.230	17.664
2016-02	mcore	T1	13857	0.649	0.972	22.071	16.776

Table 2. Comparative information about efficiency and duration at the different sub–clusters over one month.

(like in Bari, where computing for HEP has local expertise), or via commercial providers on a standard Cloud infrastructure. In the latter case, we have proved that a fully working setup, completely transparent to the rest of the CNAF infrastructure, can be obtained via a non intrusive VPN setup, automatised by the dynfarm tool. The setup is easily configurable, and offers a solution for fast-turnaround resource utilisation, even on the short time scale.

7. References

[1] https://www.aruba.it/en/home.aspx

[2] S. Chatrchyan et al. CMS Collaboration 2008 "The CMS experiment at the CERN LHC" J. Inst. 3 S08004.

[3] https://wlcg-rebus.cern.ch/apps/pledges/summary/

[4] Sfiligoi I, Bradley D C, Holzman B, Mhashilkar P, Padhi S and Würthwein F 2010 "The pilot way to grid resources using glideinwms" WRI World Congress 2 428-432.

[5] Thain D, Tannenbaum T and Livny M 2004 "Distributed computing in practice: the condor experience Concurrency" Pract. Exper. 17 323-356.

[6] Blomer J, Buncic P, Charalampidis I, Harutyunyan A, Larsen D, and Meusel R 2012 "Status and future perspectives of CernVM-FS" J. Phys.: Conf. Ser. 396 052013.

[7] Brun R and Rademakers F 1997 "ROOT - An Object Oriented Data Analysis Framework" Nucl. Instr. Meth. A389 pp 81-86.

[8] https://www.ibm.com/support/knowledgecenter/SSFKCN/gpfs_welcome.html

[9] http://ganglia.info/

[10] Kenneth Bloom and the Cms Collaboration "CMS Use of a Data Federation" Journal of Physics: Conference Series, Volume 513, Track 4.

[11] http://www.recas-bari.it/index.php/en/

Software Services and Distributed Systems

CNAF Provisioning system

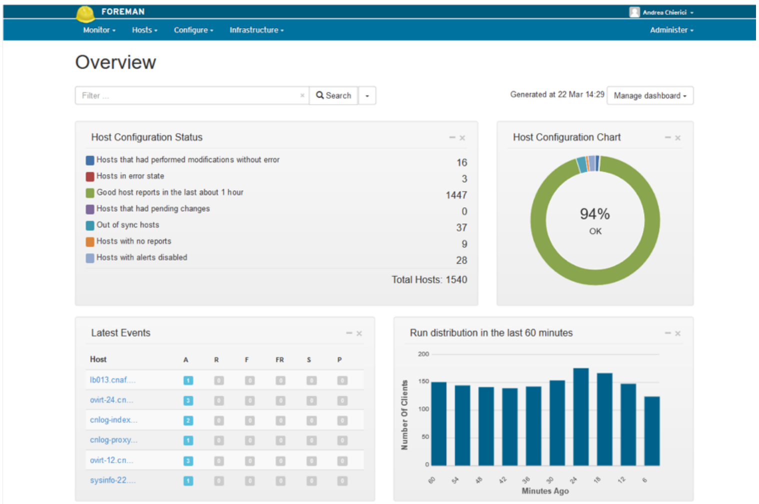
Stefano Bovina[1], **Andrea Chierici**[1], **Enrico Fattibene**[1], **Diego Michelotto**[1], **Giuseppe Misurelli**[1], **Saverio Virgilio**[1]

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

E-mail: stefano.bovina@cnaf.infn.it, andrea.chierici@cnaf.infn.it, enrico.fattibene@cnaf.infn.it, diego.michelotto@cnaf.infn.it, giuseppe.misurelli@cnaf.infn.it, saverio.virgilio@cnaf.infn.it

1. Abstract

The installation and configuration activity, in a big computing centre like CNAF, must take into account the size of the resources (roughly a thousand nodes to manage), the heterogeneity of the systems (virtual vs physical nodes, computing nodes and different type of servers) and the different working group in charge for their management. To meet this challenge CNAF implemented a unique solution, adopted by all the departments, based on two well known open source technologies: Foreman [1] for the initial installation, and Puppet [2] for the configuration of the machines.



Overview

Generated at 22 Mar 14:29

Host Configuration Status

🟦 Hosts that had performed modifications without error	16
🟥 Hosts in error state	3
🟩 Good host reports in the last about 1 hour	1447
🟪 Hosts that had pending changes	0
🟦 Out of sync hosts	37
🟧 Hosts with no reports	9
🟦 Hosts with alerts disabled	28

Total Hosts: 1540

Latest Events

Host	A	R	F	FR	S	P
lb013.cnaf....	1	0	0	0	0	0
ovirt-24.cn...	3	0	0	0	0	0
cnlog-index...	2	0	0	0	0	0
cnlog-proxy...	1	0	0	0	0	0
ovirt-12.cn...	3	0	0	0	0	0
sysinfo-22....	1	0	0	0	0	0

Figure 1. Foreman web frontend

2. The provisioning issue

CNAF is a Tier1 data center, serving the scientific community with a set of different services, as well as a huge number of computing resources counting for a total of roughly 2000 servers. In the past different groups, not only working for the Tier1, adopted different installation and configuration systems: among this the most used was quattor [3], a provisioning system developed inside the EGI project[4] funded by the EU. Other groups, with a small number of machines, didn't adopt the tool, considering the effort in learning and adopting it an overkill. Things got more complicated when virtualization and cloud infrastructures were widely adopted (added to the bare-metal installation) and this pushed towards the adoption of a more flexible, scalable and well documented set of common tools.

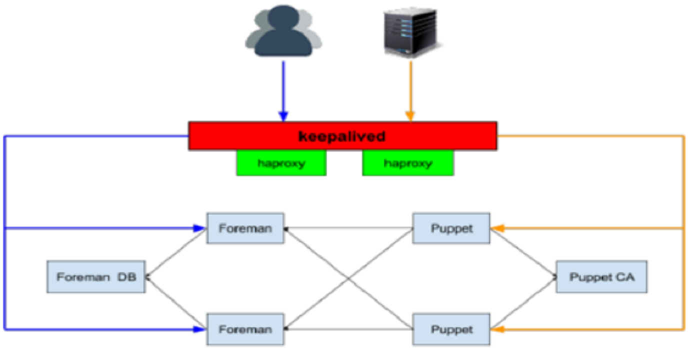


Figure 2. Architecture schema

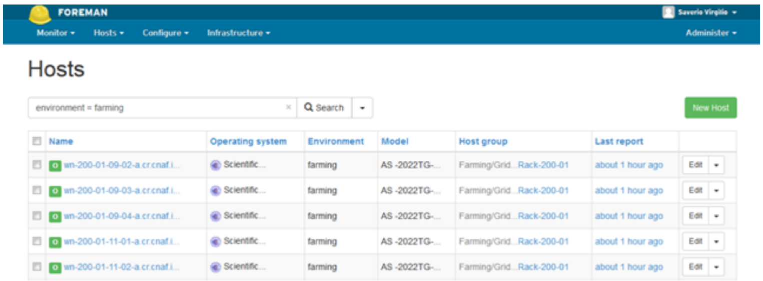
3. Puppet and Foreman

Puppet is "a configuration management tool written in Ruby with a client-server model that uses a declarative language to configure clients". Puppet is open source and supported by a large community of developers and users who contribute to providing documentation and support. Puppet's defining characteristic is that it speaks the local language of your target hosts. This allows Puppet to define systems administration and configuration tasks with generic instructions on the Puppet server. At the heart of how Puppet works is a language that allows to articulate and express the configuration. Configuration components are organized into entities called resources. You describe your configuration in terms of resources such as packages and files. This description is called a manifest. When Puppet runs on a computer, it compares the current configuration to the manifest. It will take whatever actions are needed to change the machine so that it matches the manifest.

Foreman is an open source project that helps system administrators manage servers throughout their lifecycle, from provisioning and configuration to orchestration and monitoring. Using Puppet and Foreman's smart proxy architecture, we can easily automate repetitive tasks, quickly deploy applications and pro-actively manage change, on bare-metal or with VMs. Foreman provides comprehensive interaction facilities including a web frontend (shown in figure 1, CLI and RESTful API which enables the possibility to build higher level business logic on top of a solid foundation. Foreman is deployed in many organizations, managing from 10s to 10.000s of servers.

4. Architecture and system configuration

The infrastructure implemented at CNAF (see figure 2) consists of two Puppet and two Foreman servers. To perform load balancing between the two, a couple of HAProxy [5] have been installed, both configured with virtual IPs (one for Foreman and one for Puppet server). Virtual addresses are managed via a keepalived service, which associates the virtual IP to a different server in case the running one goes down or becomes unavailable for any reason. A single database server



Hosts

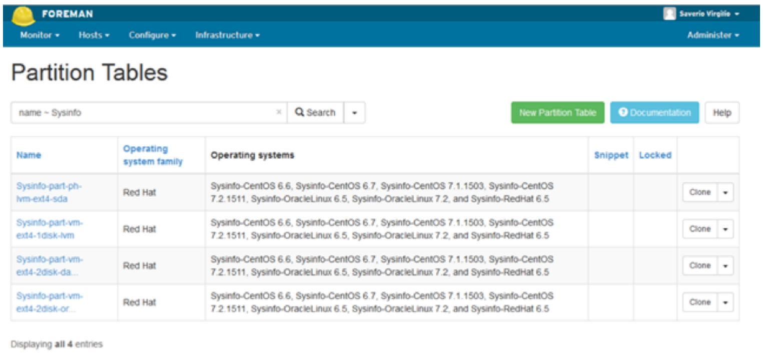
environment = farming

□	Name	Operating system	Environment	Model	Host group	Last report
□	wn-200-01-09-02-a.cr.cnaf.i...	Scientific...	farming	AS -2022TG-...	Farming/Grid...Rack-200-01	about 1 hour ago
□	wn-200-01-09-03-a.cr.cnaf.i...	Scientific...	farming	AS -2022TG-...	Farming/Grid...Rack-200-01	about 1 hour ago
□	wn-200-01-09-04-a.cr.cnaf.i...	Scientific...	farming	AS -2022TG-...	Farming/Grid...Rack-200-01	about 1 hour ago
□	wn-200-01-11-01-a.cr.cnaf.i...	Scientific...	farming	AS -2022TG-...	Farming/Grid...Rack-200-01	about 1 hour ago
□	wn-200-01-11-02-a.cr.cnaf.i...	Scientific...	farming	AS -2022TG-...	Farming/Grid...Rack-200-01	about 1 hour ago

Figure 3. Host list of farming environment

for foreman has been installed, as well as a Puppet Certification Authority (CA), that manages certificates used during the communication between the Puppet server and the installed nodes that need to be managed.

The configurations instruction have been implemented through Puppet modules (classes, files, templates and other resources) in such a general way in order to allow for the reuse of the same modules among different CNAF departments (this was very difficult with the past configuration tools). To achieve this goal, resources available on the PuppetForge public repository were used as much as possible: this repository contains modules developed by the Puppet community worldwide. As an example modules were taken to configure generic system services like cron, logrotate, yum, sudo, httpd as well as modules to setup the installation infrastructure itself (Puppet, Foreman, HAProxy, etc.).



Partition Tables

name ~ Sysinfo

Name	Operating system family	Operating systems	Snippet	Locked
Sysinfo-part-ph-lvm-ext4-sda	Red Hat	Sysinfo-CentOS 6.6, Sysinfo-CentOS 6.7, Sysinfo-CentOS 7.1.1503, Sysinfo-CentOS 7.2.1511, Sysinfo-OracleLinux 6.5, Sysinfo-OracleLinux 7.2, and Sysinfo-RedHat 6.5		
Sysinfo-part-vm-ext4-1disk-lvm	Red Hat	Sysinfo-CentOS 6.6, Sysinfo-CentOS 6.7, Sysinfo-CentOS 7.1.1503, Sysinfo-CentOS 7.2.1511, Sysinfo-OracleLinux 6.5, Sysinfo-OracleLinux 7.2, and Sysinfo-RedHat 6.5		
Sysinfo-part-vm-ext4-2disk-da...	Red Hat	Sysinfo-CentOS 6.6, Sysinfo-CentOS 6.7, Sysinfo-CentOS 7.1.1503, Sysinfo-CentOS 7.2.1511, Sysinfo-OracleLinux 6.5, Sysinfo-OracleLinux 7.2, and Sysinfo-RedHat 6.5		
Sysinfo-part-vm-ext4-2disk-or...	Red Hat	Sysinfo-CentOS 6.6, Sysinfo-CentOS 6.7, Sysinfo-CentOS 7.1.1503, Sysinfo-CentOS 7.2.1511, Sysinfo-OracleLinux 6.5, Sysinfo-OracleLinux 7.2, and Sysinfo-RedHat 6.5		

Displaying **all 4** entries

Figure 4. Partition table list of sysinfo enviroment

As the time of writing, this infrastructure is adopted by Farming and Storage departments of the Tier1 center, the Software Development and Distributed Systems (SDDS) division and the Information System, managing a total of roughly 1600 nodes. In order to isolate the different departments the “Environment” functionality of Puppet has been used: this allows to create subsets of nodes and to apply a different version of the same module to one or more nodes. This allows the creation of a test environment for code.

Internally developed modules are stored on a Git server and are organized in projects. Here is a subset of the main ones:

Modules: repository for commonly used classes

Farming: repository for farming specific configurations

Storage: repository for storage specific configurations

SDDS: repository for SDDS specific configurations

Sysinfo: repository for information system specific configurations

Production: repository for classes used to setup the core infrastructure

Each project has two branches: development and master. Master branch contains the set of Puppet modules used in production and is kept synchronized with the corresponding environment on Puppet server through a specific git tool (webhook).

The user authentication on Foreman web page is done through the INFN LDAP credentials, while the access to Foreman resources (templates, OSes, hosts) is done through the implementation of different roles. Right now a role for every department has been defined, in this way a user can manage only the resources belonging to his department (unless the user is also an administrator). Each role is defined by a set of actions that a particular user can make on the resources managed by Foreman. Roles have been created by filtering actions based on the resource environment they belong to (e.g. see hosts in figure 3) or, in case environment is not available, by their name (see as an example the partition tables as depicted in figure 4). To ease the process it was decided to add a prefix with the name of the department involved to every resource name present in the repository.

An help-desk activity is provided by staff people with weekly shifts: users requests as well as infrastructure problems can be reported. Every help-desk activity is tracked on Jira under the "Bebop CNAF" project.

5. References

[1] Foreman webpage: http://theforeman.org

[2] Puppet webpage: http://puppetlabs.com

[3] quattor webpage: http://www.quattor.org

[4] EGI project: http://www.egi.eu

[5] HAProxy webpage: http://www.haproxy.org/

CNAF Monitoring system

Stefano Bovina[1], **Diego Michelotto**[1], **Giuseppe Misurelli**[1]

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

E-mail: stefano.bovina@cnaf.infn.it, diego.michelotto@cnaf.infn.it, giuseppe.misurelli@cnaf.infn.it

1. Abstract

Over the past two years, the operations at INFN-CNAF have undergone significant changes. The adoption of configuration management tools like Puppet [5] and the constant increase of dynamic and cloud infrastructures, led us to investigate a new monitoring approach. We are going to describe our monitoring infrastructure, based on Sensu [1] as monitoring router, InfluxDB [4] as time series database to store data gathered from sensors and Grafana [3] to create dashboards and to visualize time series metrics.

2. The monitoring issue

Monitoring systems that are currently mainly used at CNAF are Nagios and Lemon. Lemon is a metric gatherer and viewing tool developed at CERN, but it is no longer maintained. Nagios is a monitoring tool currently widely used at CNAF but it's not well suited for a modern computing center like our, since it does not scale properly, has an old-style UI, is not designed to manage a dynamic infrastructure and is difficult to be managed via configuration management tools.

Our aim is the centralization of the monitoring service at CNAF through a scalable and highly configurable monitoring infrastructure.

The selection of tools has been made taking into account the following requirements given by our users:

- cloud oriented monitoring system
- horizontally scalable monitoring system
- management of monitoring through CM tools (Puppet)
- capability to provide more flexibility
- re-usability of Nagios scripts
- re-usability and ease of access to information and data
- interaction with API
- modern UI and dashboard composer system
- separation of contexts among CNAF departments

The chosen tools to achieve these goals are Sensu, Uchiwa, InfluxDB and Grafana

3. Sensu

Sensu is an infrastructure and application monitoring and telemetry solution that allows to reuse monitoring checks and plugins from legacy monitoring tools like Nagios.

Sensu was designed from day one as a replacement for an aging Nagios installation and was specifically designed to solve monitoring challenges introduced by modern infrastructure platforms with a mix of static, dynamic, and ephemeral infrastructure at scale (i.e. public, private, and hybrid clouds).

We are currently using Sensu for:

- Check systems status
- Send alerts and notifications
- Gather metrics from systems that have to be sent to InfluxDB

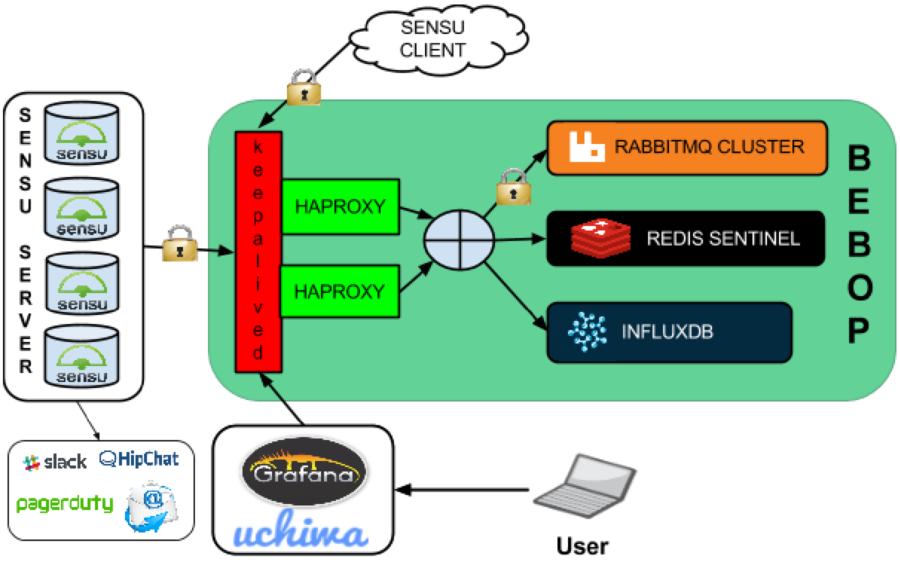


Figure 1. Monitoring system Architecture

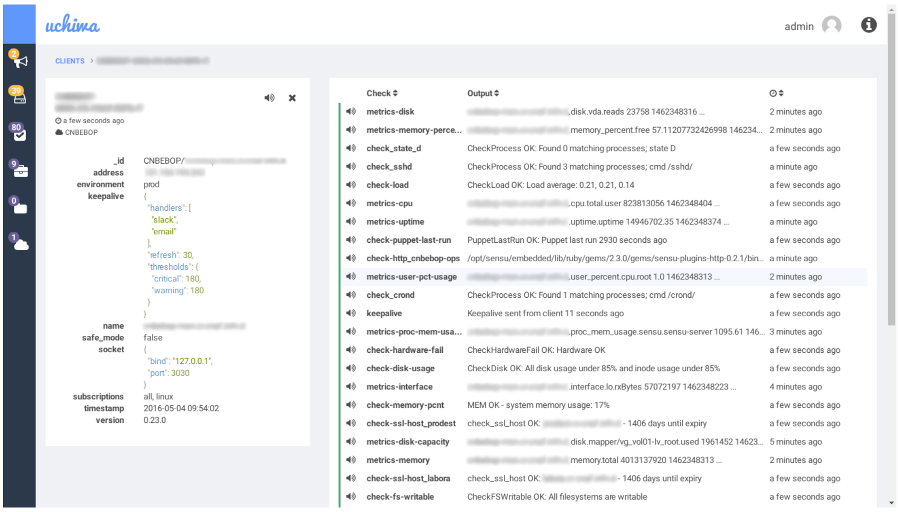
Sensu provides its own plugins, but is 100% compatible with Nagios plugins: this makes the migration from Nagios (but also Lemon) to Sensu very easy with just a little additional work.

Currently 500 nodes are monitored and many use cases have been dealt with:

- custom extension to optimize the writing of data in InfluxDB
- aggregation of check results and daily reports
- notification based on a percentage of failing host on a specific check
- notification routing and escalation

4. Uchiwa

Uchiwa is an open source dashboard for Sensu. It provides a modern UI in AngularJS to visualize the state of your systems and an easy way to interact with Sensu server through API provided by Sensu api service.



[redacted]

[redacted]

a few seconds ago

CNBEBOP

```
          _id   CNBEBOP/[redacted]
      address   [redacted]
  environment   prod
    keepalive   {
                  "handlers": [
                    "slack",
                    "email"
                  ],
                  "refresh": 30,
                  "thresholds": {
                    "critical": 180,
                    "warning": 180
                  }
                }
         name   [redacted]
    safe_mode   false
       socket   {
                  "bind": "127.0.0.1",
                  "port": 3030
                }
subscriptions   all, linux
    timestamp   2016-05-04 09:54:02
      version   0.23.0
```

Check	Output	
metrics-disk	[redacted].disk.vda.reads 23758 1462348316 ...	2 minutes ago
metrics-memory-perce...	[redacted].memory_percent.free 57.11207732426998 146234...	2 minutes ago
check_state_d	CheckProcess OK: Found 0 matching processes; state D	a few seconds ago
check_sshd	CheckProcess OK: Found 3 matching processes; cmd /sshd/	a minute ago
check-load	CheckLoad OK: Load average: 0.21, 0.21, 0.14	a few seconds ago
metrics-cpu	[redacted].cpu.total.user 823813056 1462348404 ...	a few seconds ago
metrics-uptime	[redacted].uptime.uptime 14946702.35 1462348374 ...	a minute ago
check-puppet-last-run	PuppetLastRun OK: Puppet last run 2930 seconds ago	a few seconds ago
check-http_cnbebop-ops	/opt/sensu/embedded/lib/ruby/gems/2.3.0/gems/sensu-plugins-http-0.2.1/bin...	a minute ago
metrics-user-pct-usage	[redacted].user_percent.cpu.root 1.0 1462348313 ...	2 minutes ago
check_crond	CheckProcess OK: Found 1 matching processes; cmd /crond/	a few seconds ago
keepalive	Keepalive sent from client 11 seconds ago	a few seconds ago
metrics-proc-mem-usa...	[redacted].proc_mem_usage.sensu.sensu-server 1095.61 146...	3 minutes ago
check-hardware-fail	CheckHardwareFail OK: Hardware OK	a few seconds ago
check-disk-usage	CheckDisk OK: All disk usage under 85% and inode usage under 85%	a few seconds ago
metrics-interface	[redacted].interface.lo.rxBytes 57072197 1462348223 ...	4 minutes ago
check-memory-pcnt	MEM OK - system memory usage: 17%	a few seconds ago
check-ssl-host_prodest	check_ssl_host OK: [redacted] - 1406 days until expiry	a few seconds ago
metrics-disk-capacity	[redacted] disk.mapper/vg_vol01-lv_root.used 1961452 14623...	5 minutes ago
metrics-memory	[redacted] memory.total 4013137920 1462348313 ...	2 minutes ago
check-ssl-host_labora	check_ssl_host OK: [redacted] - 1406 days until expiry	a few seconds ago
check-fs-writable	CheckFSWritable OK: All filesystems are writable	a few seconds ago

Figure 2. Uchiwa dashboard

5. InfluxDB

InfluxDB is an open source database optimized to handle time series data. It has no external dependencies, that means once you install it theres nothing else to manage.

InfluxDB is targeted at use cases for DevOps, metrics, sensor data, and real-time analytics. Some of the key features that InfluxDB currently supports are:

- SQL like query language
- HTTP(S) API
- Store billions of data points
- Database managed retention policies for data

6. Grafana

Grafana is an open source dashboard composer and is most commonly used for visualizing time series data for Internet infrastructure and application analytics. Some of the key features are:

- Support for several datasources: InfluxDB, Graphite [6], Elasticsearch [7] etc.
- Fully-interactive
- Editable graphs
- Create variables that are automatically filled with values from your DB
- Variables in metric queries and panel titles can be reused
- Automatically repeat rows or panels for each selected variable value
- Multi tenancy and LDAP integration

7. Puppet automation

Our environment is mainly composed of dynamic infrastructures: for this reason we need to be able to automatically manage the installation and configuration of alarm and monitoring software.

The Puppet community has developed a wide set of templates to configure Sensu: since we already use Puppet to provision and configure our servers, we tought that adapting these templates to our needs could have helped in saving time and effort.

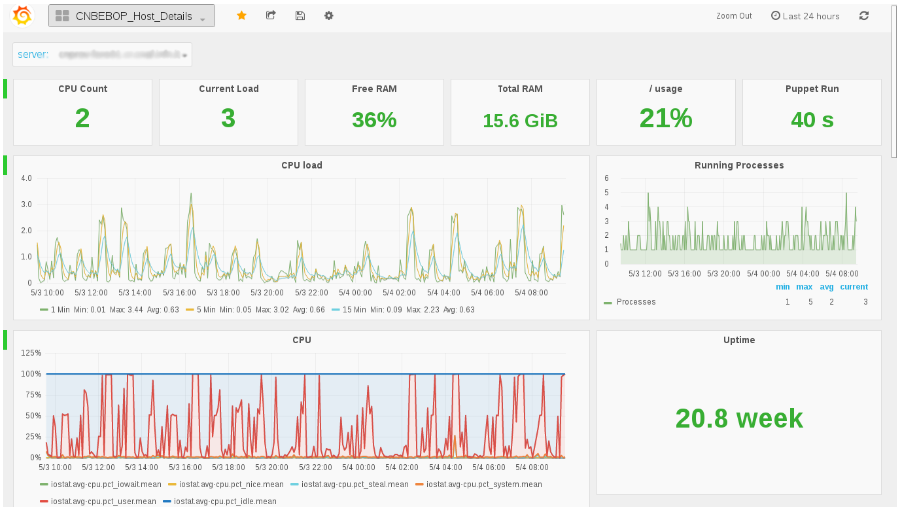


Figure 3. Grafana dashboard

Today, through our provisioning system, we can easily update and reconfigure our Sensu clients and server as well as modify the general configuration.

8. Future works

The monitoring environment we have now is only a first step of a bigger plan. The set of nodes is growing daily and with this we may hit possible scalability problems: we already applied best practices to achieve maximum scalability and reliability, that may be put under stress in the forthcoming months. Anyway further tuning may be required, to possibly achieve better performance and to improve data consumption. The next step is to study InfluxDB server tuning and downsampling of data series.

9. References

[1] Sensu webpage: https://sensuapp.org/
[2] Uchiwa webpage: https://uchiwa.io/
[3] Grafana webpage: http://grafana.org/
[4] InfluxDB: https://influxdata.com/
[5] Puppet webpage: http://puppetlabs.com
[6] Graphite webpage: http://graphite.wikidot.com/
[7] Elasticsearch webpage: https://www.elastic.co/

An initial study on software metrics thresholds

Elisabetta Ronchieri, Marco Canaparo

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

Abstract. The issue of defining thresholds of software metrics has always been of paramount importance in the field of software quality over the years. In papers, software engineers and researches have striven to propose the most general approach to tackle this problem over time. The result is a plethora of different techniques to calculate thresholds which **(that)** range from personal experience, to statistics and, in the last decade, to methods based on machine learning, which are applied to different contexts. For these reasons, it may be difficult to understand when and how to use them in a efficient way. In this report, we provide some background information about software quality, metrics and thresholds; furthermore, we describe the methodology we followed to analyse existing papers that have faced the issue of determining thresholds of software metrics.

1. Introduction

Different organizations have defined software quality over the years. The IEEE defines quality as the degree to which a system, component, or process meets specified requirements or customer or user needs or expectations [1]. The International Organization for Standardization [2] defines quality as the degree to which a set of inherent characteristics fulfils requirements. Other experts define quality based on conformance to requirements and fitness for use. However, a good definition must lead us to measure quality meaningfully. According to Fenton and Bieman [3], measurement is the process by which numbers or symbols are assigned to attributes of entities in the real world in such a way as to describe them according to clearly defined rules.

From the dawn of software engineering, software metrics have been used to measure code features. The IEEE defines software metrics as "the quantitative measure of the degree to which a system, component or process possesses a given software attribute" [1] related to quality characteristics. Measurement enable developers to have knowledge about the quality of software during (or at the end) of its development. Moreover, software metrics can contribute to get information about the future maintenance needs of software.

In order to change the target of metrics employment from simple measurement, to source that indicates the level of quality or maintenance needs of a software project, it is essential to meaningfully defining threshold values. One of the most widespread metrics is McCabe complexity metrics [4]. Its author specified a threshold of 10: as a consequence, a subroutine with a higher number was expected to be unmaintainable and untestable. However, given a certain software project and its metrics, the issue of determining its main feature (in terms of, for example, maintainability) from them is particularly challenging, because this would imply to fix a certain threshold for each metric: every piece of software whose metrics outdid it, it would be considered unmaintainable.

Table 1. Search query

$Search_{Id}$	**Goal**	**Search String**
$Search_1$	papers 1970-2014	(TITLE-ABS-KEY (software) AND TITLE-ABS-KEY (metrics) AND TITLE-ABS-KEY (thresholds)) AND SUBJAREA (mult OR ceng OR CHEM OR comp OR eart OR ener OR engi OR envi OR mate OR math OR phys) AND PUBYEAR > 1969 AND PUBYEAR < 2015

2. Search Process and Data Extraction

In our study, we decided to consider all the papers published from 1970 to 2014; we have used the SCOPUS tool to search for relevant papers. SCOPUS is a very useful system, because it indexes IEEE, ACM, Elsevier and Springer Lecture Notes publications. This makes SCOPUS potentially a very powerful tool for research. To the best of our knowledge, until now, there are no surveys in the thresholds' field.

As a first step, we determined the string to use in the search engine which is outlined in Table 1; in addition to this, we specified the range 1970-2014 and we refined the research by excluding fields not related to software engineering. Search 1 found 253 articles many of which were irrelevant for example papers that reported biodiversity or photon in the title. As a second step we filtered papers according to their features. We removed all the papers with the words above in the title and other 2 with no title at all; as a consequence, there were 95 records left. Then, amongst these, we found 78 papers that looked relevant by leveraging a brief review of both the abstract and the text. We also removed some papers that were off topic or not found on internet (even the abstract) or not written in English. As a final step, we sorted the remaining papers in terms of relevance, removing the ones without any citations. There were 55 records left.

In our research, we looked for papers that describe an approach to the use of software metrics thresholds. We kept papers among the valid ones even though they did not include experimental results. We did not exclude conference papers because conference proceedings publish experience reports, also including a source of information about the industry's experience.

As regards the extraction phase, we collected some standard information about all the papers we analyzed. Our aim is to classify all of them according to the following features:

(i) the main topic
(ii) whether the paper was empirical, theoretical or both
(iii) whether the paper applied metrics in the context of open source software;
(iv) whether the paper use public data-sets of metrics;
(v) which languages the analyzed projects are written;
(vi) whether the techniques discussed were statistical or artificial intelligence based;
(vii) whether the context of the paper was maintenance;
(viii) what sort of metrics were being evaluated;
(ix) the summarized goal of the study.

3. Future work

In 2016, this work will be improved by the specification of other search queries in order to include as many papers as possible; furthermore, we will collect all the information about all the studies in the software metrics thresholds field according to the criteria listed above. In the long term, we also foresee the adoption of other tools (such as ACM, IEEE and CiteSeer digital libraries) in order to have a comparison with the results obtained by SCOPUS.

References

[1] IEEE 1990 Ieee standard glossary of software engineering terminology. ieee std 610.12-1990. Tech. rep.

[2] ISO 1987 Iso 9000 - quality management Tech. rep. URL $http : //www.iso.org/iso/home/standards/management - standards/iso_9000.htm$

[3] Fenton N and Bieman J 2014 *Software Metrics: A Rigorous and Practical Approach, Third Edition* (CRC Press)

[4] McCabe T J 1976 *IEEE Transactions on Software Engineering* **SE-2** 308–320

Dynfarm: a dynamic farm extension

V Ciaschini[1]**, D De Girolamo**[2]

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

E-mail: [1]`vincenzo.ciaschini@cnaf.infn.it`

E-mail: [2]`donato.degirolamo@cnaf.infn.it`

Abstract. Requests for computing resources for physics experiments are constantly increasing and so are peak usages. New resources are continuously required, and procurement strategies different than simple acquisitions are being considered. One of them is the opportunistic utilization of resources from commercial clouds.

This report describes a software product designed to configure remote resources and make them part of the computing farm, allowing full use of all management tools as if the resources were local, modulo, obviously, an increased network latency.

1. Introduction

Dynfarm is a new software product developed in 2015. Its aim is to allow seamless integration into the existing Tier1 farm of computing resources physically located outside of the computing centre. It thus allows to dynamically increase the number of resources available to better copy with times of peak usage, while putting only minimal requirements on the remote nodes themselves.

2. Architecture

The main architecture is a client-server one, and can be observed in figure 1:

The remote host needs to have a client installed on it that at startup will connect to a configuration server present at CNAF. After an authentication phase, the server will send to the client three things:

- a VPN client configuration and corresponding credentials that will be used to connect to a VPN server at CNAF and create an end-to-end tunnel, which by itself does not provide access to anything else.
- a set of commands to execute
- a ssh public key to allow remote control of the machine

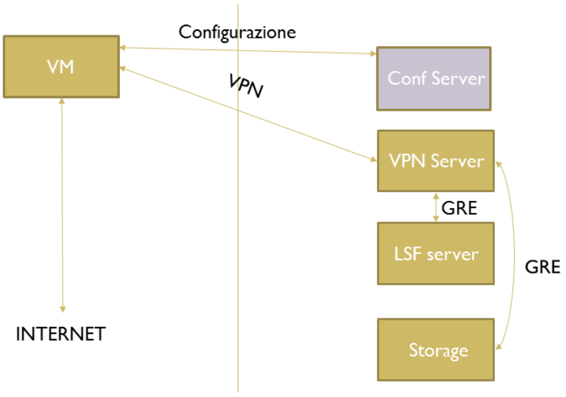


Figure 1. dynfarm architecture

In order to allow the remote machine to act as a member of the Tier1 farm, some local resources need to be reachable. This is obtained via GRE tunnels between the VPN server and the those resources at CNAF, along with an appropriate route setup.

This system puts only minimal requirements on the remote machines. Namely, they must have outgoing connectivity, must be able to receive UDP packets (for the VPN establishment) and must allow the installation of the client program. Direct addressing through an IP address is not required, and indeed we have a working setup where all remote machines are behind a NAT that masks all of their IP addresses.

After the initial setup, no more connections need occur between the configuration server occur and the remote host. This however, does not mean that no connection may occur. If the remote host is running a SSH server, the SSH public key allow the server to run arbitrary commands on it, and to send and receive files. Note that this is not an hard requirement. If the SSH server is not running these additional capabilities cannot be used, but this does in no way interfere with the rest of the system.

3. Use in production

Thanks to the collaboration of the CMS experiment and of all the CNAF Tier1, this system has been setup in production using resources hosted in Aruba's cloud centre (the leading Italian cloud provider), and has proved to be both working and viable.

4. Thanks

The authors wish to thank all the members of the CNAF Tier1 and members of the Bologna T3 and CMS experiment for their collaboration in creating both the prototype and the production instance.

References

[1] V. Ciaschini, S .Dal Pra, D. de Girolamo, L. dell'Agnello, A. Chierici, V. Sapunenko, T. Boccali, A. Italiano, Elastic CNAF DataCenter extension via opportunistic resources, forthcoming in proceedings of ISGC 2016

An InfiniBand-based Event Builder software for the LHCb experiment

A Falabella[1], F Giacomini[1], M Manzali[3,1] and U Marconi[2]

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

E-mail: matteo.manzali@cnaf.infn.it

Abstract. The Data Acquisition (DAQ) of the LHCb experiment will be upgraded in 2020 to a high-bandwidth trigger-less readout system. In the new DAQ event fragments will be forwarded to the to the Event Builder (EB) computing farm at 40 MHz. Therefore the front-end boards will be connected directly to the EB farm through optical links and PCI Express based interface cards. The EB is requested to provide a total network capacity of 32Tb/s, exploiting about 500 nodes. In order to get the required network capacity we developed the Large Scale Event Builder (LSEB) software, an Event Builder implementation designed for an InfiniBand interconnect infrastructure. We present the results of the measurements performed to evaluate throughput and scalability measurements on HPC scale facilities.

1. Introduction

The LHCb experiment [1] is one of the four main experiments performed with LHC at Cern. It aims at the study of CP violation and precision measurements of rare decays of b and c quark hadrons. A major upgrade of the detector is foreseen during the second long shutdown of the LHC (2018-2019). A new trigger-less [2] readout system will operate at the bunch crossing frequency of 40MHz. The Event Builder (EB) is the part of the readout system responsible for the collection of the event fragments from the sub-detector readout boards. These fragments are then combined to form the full event. To operate at the aforementioned frequency the EB will be redesigned. The idea is to implement it with an high throughput LAN using off-the-shelf hardware.

2. Upgrade DAQ implementation

The main feature of the LHCb upgraded detector is the trigger-less readout system, without any hardware trigger. The logic scheme of the readout system is shown in Figure 1, with the main components: The Event Builder, the Timing and Fast Control (TFC) system, The Experiment Control System (ECS) for monitoring and configuration, and the trigger Event Filter Farm (EFF). The aggregated bandwidth of the EB network can be estimated given the foreseen nominal event size of 100 KBytes, assuming the maximum event rate of 40 MHz, to be of the order of 32 Tbit/s. The size of the EB farm can be estimated of the of the order 500 PCs, this by assuming the input rate of a server, through a custom readout board [3], is about 100 Gbit/s.

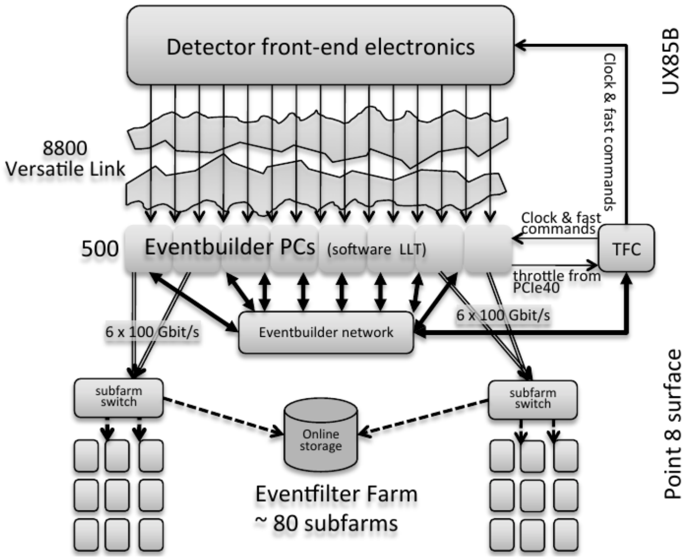


Figure 1. The architecture of the upgraded LHCb readout-system.

The EB network can be effectively implemented by using commercial local area network technologies such as Ethernet, OmniPath or InfiniBand. In the following we present the results of scalability tests of Large Scale Event Builder (LSEB) software [4], an Event Builder implementation designed for an InfiniBand interconnect infrastructure. The InfiniBand standard is widely used in HPC clusters, since it provides cost-effective high-speed performance.

3. The Large Scale Event Builder design

In the EB design foreseen for the upgrade, each EB node includes two distinct logical components: the Readout Unit (RU) and the Builder Unit (BU). A RU receives event fragments from the detector and ships them to a receiving BU in a many-to-one pattern. Each BU gathers the event fragments and assembles them in full events, which are then sent out to the EFF for processing. The LSEB software implementation reflects this design and its main blocks are represented in the schematic view of Figure 2.

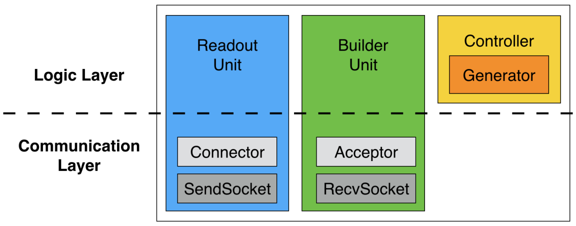


Figure 2. Schematic view of the main blocks of LSEB.

In order to keep the communication management separated from the logic of the event-building, LSEB can be nominally split into two distinct layers, namely the Communication Layer and the Logic Layer. The Communication Layer includes primitives for data communication between nodes and relies on the InfiniBand *verbs* interface [5], a library that offers a user-space API to access the Remote Direct Memory Access (RDMA) capabilities of the network device. On top of the Communication Layer sits the Logic Layer, a set of software components performing the actual event-building under realistic conditions.

4. Scalability tests

We had the opportunity to perform some scalability tests with LSEB on Galileo [6], one of the main supercomputers hosted at the CINECA Consortium [7]. Galileo is the Tier-1 system of CINECA, introduced in January 2015. It is composed by 516 compute nodes, each containing two Intel Xeon Haswell 8-core E5-2630 v3 processors, with a clock of 2.40 GHz. All the compute nodes have 128 GB of memory. Each node is connected to an Infiniband network through a QLogic Single-Port QDR InfiniBand HCA [8]. Table 1 summarizes the system specification of Galileo.

Table 1. Galileo System Specification

Model	IBM NeXtScale
Architecture	Linux Infiniband Cluster
Nodes	516
Processors	2 x 8-cores Intel Haswell 2.40 GHz
Cores	16 cores/node, 8256 cores in total
Accelerators	2 Intel Phi 7120p per node on 344 nodes
RAM	128 GB/node, 8 GB/core
Internal Network	Infiniband with 4x QDR switches
HCA	QLogic Single-Port QDR InfiniBand

Due to the policies of the cluster it was not possible to apply fine-tuning operations on the kernel in order to achieve the best performance. Indeed for reasons related to power consumption, in Galileo each node has the CPU Frequency governor (CPUFreq) set to ”ondemand” and the *c-states* [9] enabled.

4.1. Standard benchmark

Generically speaking, before testing an application on one or more nodes, it is recommended to execute a standard benchmark in order to establish a performance baseline. In case of applications that make use of RDMA, there is a set of micro benchmarks provided by the OpenFabrics Enterprise Distribution (OFED) package [10], that allows to verify the effective point-to-point network capacity. One of these micro benchmarks, the so-called *ib_write_bw*, was chosen and used to identify the real maximum bandwidth attainable between two random nodes belonging to the cluster. The tests performed with *ib_write_bw* foresee the execution of 5000 bidirectional RDMA write transactions for each different buffer size from 2 to 2^{22} bytes.

In Figure 3 the average bandwidth obtained running 10 times *ib_write_bw* on two Galileo nodes is plotted as a function of the buffer size. It is important to note that the average bandwidth is affected by a significant statistical error. Moreover, the average bandwidth and the peak bandwidth differ of about 15%. These behaviours are mostly caused by the constant presence of external jobs running on the cluster and by the missing optimization settings on the nodes. Considering the average bandwidth, the benchmark reaches 21.4 Gb/s, that is the 78.8% of the maximum theoretical bandwidth expected with QLogic QDR InfiniBand cards (27.2 Gb/s).

4.2. Results with the Event Builder software

As it is typical for clusters offering concurrent access to shared resources, Galileo computing nodes are accessible exclusively through a job scheduler. A job is described by providing a list of computational requirements on a limited set of resources, such as the number of needed cores to run on. On Galileo the maximum limit for the usable number of cores is unfortunately set to 1024; thus, although more than 500 nodes are available on the system, not all of them can

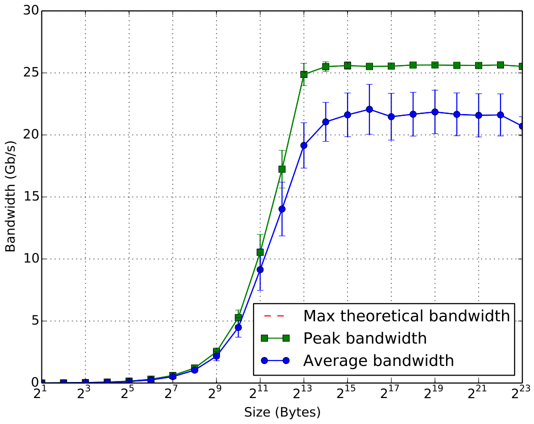


Figure 3. Benchmark with *ib_write_bw* on Galileo: the blue line represents the resulting average bandwidth and the green line represents the peak bandwidth.

be used concurrently by the same application. Even asking for eight cores on each node instead of the 16 available, the maximum number of allocable nodes is 128. Reducing the number of required cores per node may increase the number of allocable nodes but also increase the possibility to have external jobs running on the same nodes at the same time, which would disturb significantly our measurements.

Starting from a 4-node setup, several tests were performed, doubling the number of nodes at every test, up to 128. On each node LSEB could run on just two cores, one for the RU thread and one for the BU thread. However, in order to avoid the concurrent execution of external processes, as discussed above, it is desirable to reserve all the available cores on each node. This was indeed done for the tests up to 32 nodes. For the tests with 64 and 128 nodes, instead, only half of the 16 cores available on each node were allocated.

The bandwidths measured running LSEB on Galileo in different node configurations are reported in Figure 4 (left) as a function of time, presenting good stability over time. The scalability plot is shown in Figure 4 (right), where the average bandwidth is plotted as a function of the number of nodes: the bandwidth slowly decreases as the number of nodes increases, reaching about the 58% of the benchmarked bandwidth for 128 nodes.

5. Conclusions

The LHCb will under a major upgrade during the second long shutdown. Apart from detector upgrades also the readout system will redesigned in order to allow for a trigger-less data acquisition. The EB can be realized using commercial network technologies such as InfiniBand. We realized a performance evaluator in order to exploit the design possibilities of the EB. The scalability tests performed with the Galileo cluster show that the implementation we realized is viable. The slight decrease of the performance can be accounted to the fact that the cluster is a production cluster, so the network usage is not exclusive.

6. References

[1] LHCb Collaboration, *The LHCb detector at the LHC*, JINST,3,2008,S08005

[2] LHCb Collaboration, *LHCb Trigger and Online Upgrade Technical Design Report*, CERN,LHCC,2014,C10026

[3] P. Durante, N. Neufeld, R. Schwemmer, G. Balbi, and U. Marconi, *100 Gbps PCI-Express readout for the LHCb upgrade*, DOI 10.1088/1748-0221/10/04/C04018, Journal of Instrumentation (2015).

[4] M. Manzali, *lseb 2.0*, DOI 10.5281/zenodo.46935.

[5] RDMA Protocol Verbs Specification (Version 1.0), http://www.rdmaconsortium.org/home/draft-hilland-iwarp-verbs-v1.0-RDMAC.pdf.

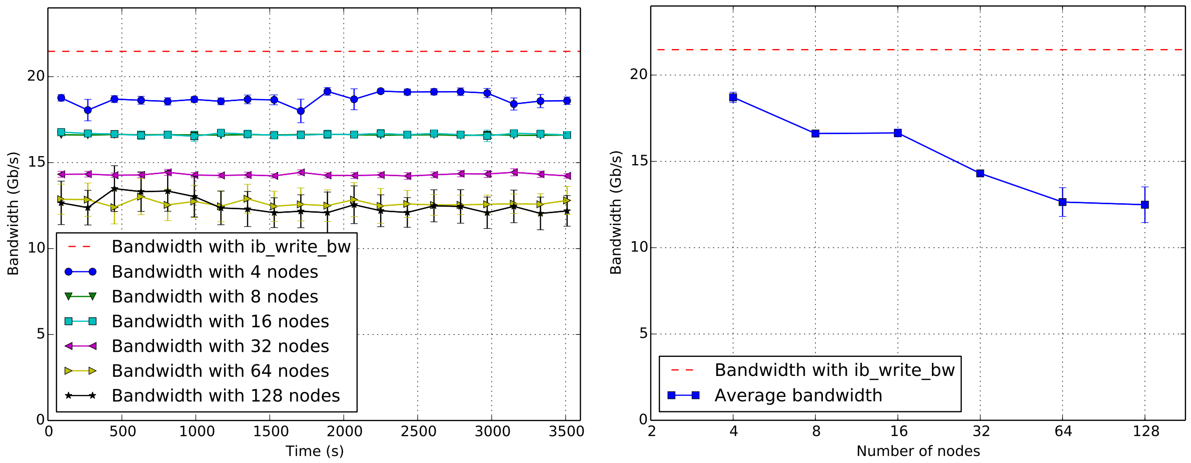


Figure 4. Bandwidth measurements (left) and scalability on an increasing number of nodes (right) for LSEB on Galileo.

[6] Galileo Cluster, http://www.hpc.cineca.it/hardware/galileo.

[7] CINECA, http://www.cineca.it/en.

[8] QLogic QLE7340 datasheet, http://www.mullet.se/dokument/QLE7340_datasheet.pdf.

[9] J. Kukunas, *Power and Performance: Software Analysis and Optimization*, ISBN 978-0-12800-814-0, Morgan Kaufmann, (2015).

[10] OpenFabrics Alliance, https://www.openfabrics.org/index.php/openfabrics-software.html.

Development of the KM3NeT-Italy acquisition and control system

T Chiarusi[1], **M Favaro**[1,2], **F Giacomini**[2], **M Manzali**[2,3], **A Margiotta**[2] **and C Pellegrino**[1,4]

[1] INFN Sezione di Bologna, Bologna, Italy
[2] INFN CNAF, Bologna, Italy
[3] Università degli Studi di Ferrara, Ferrara, Italy
[4] Università degli Studi di Bologna, Bologna, Italy

E-mail: `matteo.manzali@cnaf.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 this contribution the technology behind the implementation of the TriDAS infrastructure is reviewed, focusing on the relationship between the various components.

1. Introduction

The INFN's project KM3NeT-Italy [1], supported with Italian PON fundings, consists of 8 vertical structures, called Towers, instrumented with a total number of 700 Optical Modules (OMs) and will be deployed 3500 m deep in the Ionian Sea, at 80 km from the Sicilian coast [2][3]. A Tower is made of 14 horizontal bars, piled up one by one with 90°heading difference. Each bar hosts 6 OMs. Each OM contains a 10" PMT and the readout electronics. The detection principle exploits the measurement of the Cherenkov light from relativistic particles outgoing high-energy neutrino interaction within a fiducial volume around the telescope. In order to reduce the complexity of the underwater detector, the *all data to shore* approach is assumed, demanding for a Trigger and Data Acquisition System (TriDAS) [4] software running at the shore station. The collected data stream from all the Towers is largely affected by the optical background in the sea [5], mainly due to the ^{40}K decays and bioluminescence bursts. Ranging up to 30 Gbps, such a large throughput puts strong constraints on the required TriDAS performances and the related networking architecture. In the following sections we describe the final implementation of the physics-data handling (TriDAS Core), the user and management interfaces (TriDAS Control) and the large-band network infrastructure. Finally, we present results of scalability tests in order to demonstrate the system capabilities.

2. The TriDAS software

In this section we describe the final implementation of each component of the TriDAS Core and the control and user interfaces.

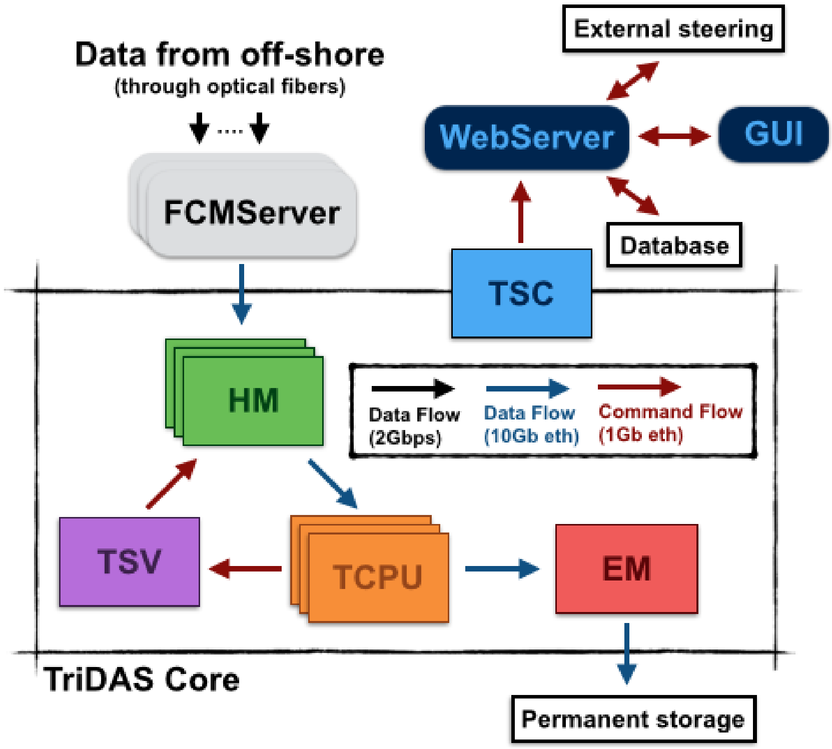


Figure 1. Scheme of TriDAS components and their interactions with external services.

Floor Control Module Server (FCMServer): the FCMServer represents the interface of TriDAS with the data from the off-shore detector. It performs the read-out of data coming from a number (up to 4) of floors through a dedicated ASIC [6]. After having performed a consistency check, it sends the data to the connected HitManager.

HitManager (HM): The HMs represent the first aggregation stage for the incoming datastream. Each HM handles a number of floors, which is called "Sector", slicing data in subsequent TimeSlices (TS) of the same fixed duration and referred to a common time origin. Each HM organizes its own sliced data in special structures called SectorTimeSlices (STSs) and sends them to the TriggerCPUs.

TriggerCPU (TCPU): The TCPUs are responsabile for the last step of data aggregation and online analysis. Each TCPU receives the STSs from HMs creating a TelescopeTimeSlice (TTS), then it applies triggers to this new object and finally sends it to the Event Manager.

Event Manager (EM): The EM is the software component of TriDAS dedicated to the storage of triggered data. A single EM process collects triggered data from the whole TCPU set and performs data writing on local storage.

TriDAS SuperVisor (TSV): The TSV supervises the data exchange between HM and TCPU, taking note of the processed TSs. When a TCPU is ready to handle new data, it sends a token to the TSV. The TSV selects a TS ID among those not yet processed and communicates to all HMs to send the STS for the selected TS ID to that TCPU.

TriDAS Controller (TSC): The TSC is the software interface that permits to control the entire TriDAS environment. Its purpose is to organize and control the launch of each software, allowing a correct acquisition and real time analysis of the data. In order to achieve this functionality the TSC implements a State Machine.

WebServer and GUI: The WebServer is the unique entry point for the TSC, which can be steered either via the GUI or other external application. The GUI is a web application that graphically represent information provided by the WebServer and acts as control interface for the user. The WebServer is also used to interface the TriDAS with the Database for retrieving the running configurations.

3. Scalability tests

In order to prove the scalability of TriDAS we performed simulation tests at the Bologna Common Infrastructure (BCI). The aim of these tests is to observe how the system behave increasing the data load (i.e. the number of Towers). Due to the limited number of available nodes at the BCI we were able to simulate up to 4 Towers. For each test we measured the time needed to analyze a TTS of 200 ms. The design of TriDAS allows to execute concurrently several TCPU processes on different nodes: the setup adopted for these tests foreseen to use 4 TCPU nodes, each one able to elaborate 24 TTS in parallel, for a total of 96 parallel TTS. This means that the system is able to run as long as the time needed to analyze a single TTS is less than the duration time of a TTS multiplied by the number of TTS parallely processed by the system; in the described setup every TCPU node has at most $0.2\ s * 96 = 19.2\ s$ for completing a TTS. Figure 2 shows the measurements performed scaling from half to 4 Towers. In the worst case the mean time required to analyze a TTS is 1.5 s; this value is well below 19.2 s that is the maximum allowed by the setup.

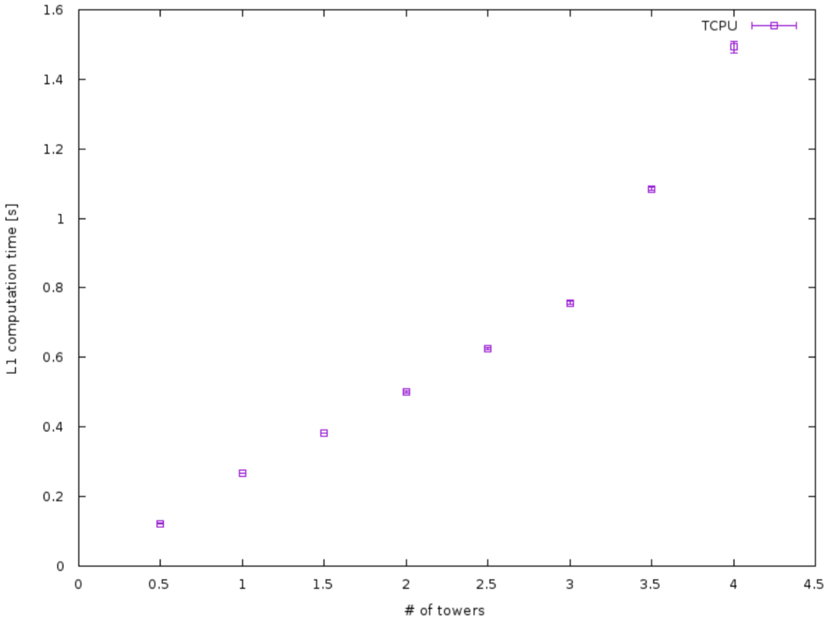


Figure 2. TimeSlice computation time as function of the number of Towers.

4. Conclusions

In order to the requirements of the KM3Net-Italy detector, an on-line Trigger and Data Acquisition System (TriDAS) has been design and implemented. The preliminary test phase demonstrates that TriDAS is stable and it is able to scale up to 4 towers. Moreover, in November 2015 the installation and functionality test of the farm at Portopalo has been completed. Extended tests of TriDAS will be also realized at the Portopalo infrastructure, in advance with respect to the first deployment of the Towers.

5. References

[1] KM3NeT Web Site, http://www.km3net.org/home.php.

[2] S. Aiello et al., *Measurement of the atmospheric muon depth intensity relation with the NEMO Phase-2 tower*, DOI 10.1016/j.astropartphys.2014.12.010, Astroparticle Physics (2015).

[3] T. Chiarusi, M. Spurio, *High-energy astrophysics with neutrino telescopes*, DOI 10.1140/epjc/s10052-009-1230-9, The European Physical Journal C (2010).

[4] C. Pellegrino, et al., *The trigger and data acquisition for the NEMO-Phase 2 tower*, DOI 10.1063/1.4902796, AIP Conference Proceedings (2014).

[5] M. Pellegriti et al., *Long-term optical background measurements in the Capo Passero deep-sea site*, DOI 10.1063/1.4902780, AIP Conference Proceedings (2014).

[6] A. Lonardo et al., *NaNet: a configurable NIC bridging the gap between HPC and real-time HEP GPU computing*, DOI 10.1088/1748-0221/10/04/C04011, JINST (2015).

A semi-automatic IaaS configuration tool for the Open City Platform project

C. Aiftimiei[1,6], A. Costantini[1], R. Bucchi[1], A. Italiano[2], D. Michelotto[1], M. Panella[1], M. Pergolesi[3], M. Saletta[4], S. Traldi[5], D. Salomoni[1], C. Vistoli[1] and G. Zizzi[1]

[1]INFN CNAF, Bologna, Italy
[2]IFIN - "Horia Hulubei", Bucharest - Magurele, Romania
[3]INFN Bari, Bari, Italy
[4]INFN Perugia, Perugia, Italy
[5]INFN Torino, Torino, Italy
[6]INFN Padova, Padova, Italy

E-mail: `cristina.aiftimiei@cnaf.infn.it, alessandro.costantini@cnaf.infn.it`

Abstract. **Open City Platform (OCP)** is an industrial research project [1] funded by the Italian Ministry of University and Research (MIUR), started in 2014. It intends to research, develop and test new technological solutions open, interoperable and usable on-demand in the field of Cloud Computing, along with new sustainable organizational models for the public administration, to innovate, with scientific results, with new standards and technological solutions, the provision of services by the Local Public Administration (PAL) and Regional Administration to citizens, companies and other public administrations. OCP inherits the experience of other projects of Cloud Computing applied to the Public Administration and research like PRISMA[2], Marche Cloud[3], INFN Cloud. The IaaS layer of OCP is based on OpenStack [4], the open source cloud solution most widespread in the world to manage physical resources, virtual machines and containers. In this paper we will present the progress of the research activity done at CNAF [5] aimed at designing and developing a set of automatic procedures in order to simplify the installation and configuration of the IaaS layer of the OCP-platform not only for the Regional testbeds part of the project, but also other other that will be interested in using the project results.

1. Introduction

As it is well known from the literature [6], Cloud Computing is a way of providing and making use of distributed computing, storage resources and mainly services that has been vigorously developed and more extensively adopted by industry, science and government. For these actors, however, the acceptance of the Cloud paradigm, even if it represents a strong opportunity, is often limited by two factors: (a) the difficulties to manage and operate the infrastructure and (b) the rules and regulations that impose specific behaviours to protect providers and consumers of public services (citizen or company).

On such premises, the Open City Platform (OCP) project [1] intends to research, develop and test new technology solutions that are open, interoperable and usable on-demand on the Cloud, as well as innovative organizational models that will be sustainable over time. The aim of the project is to innovate, with scientific results and new standards, the delivery of services by Local

Government Administrations (LGA) and Regional Administrations to citizens, Companies and other Public Administrations (PA).

In this paper we present the progress of the research activity started at CNAF last year [5] aimed at designing and developing a set of automatic procedures able to simplifying the installation and configuration of the Regional testbeds who have an active participation to the OCP project.

The paper is organized as follows: in Section 2 the new AutomaticOCP deploying tool is presented and the developed features are focused; in section 3 the developed Puppet Roles and Profiles are discussed and in Section 4 the Foreman Web-GUI supporting the AutomaticOCP tool is shown. Finally, Section 5 concludes the paper and presents directions of the future work.

2. The AutomaticOCP deployment tool

The investigated methods already described in [5] present some missing features that are here recap:

- The manual installation and configuration method can be time-consuming for the infrastructure maintainers due to the repetitive operations they have to perform to keep configurations correctly aligned among different servers and nodes.
- The automatic instalaltion method based on Fuel tool [7], even if it solves many of the disadvantages identified in the manual method, does not permit a full tuning of the infrastructure being not flexible enough to cope with the architectural needs of the PA.

For the above mentioned reasons, we designed a semi-automatic installation method able to take the advantages of the methods presented in [5], including the use of a Web-GUI, and, at the same time, flexible enough to meet the architectural requirements of the OCP project (advertised to use OpenStack version Juno [8] on Ubuntu 14.04 LTS [9]) as well as the ones of the Data Center where the OpenStack Infrastructure will be deployed.

The solution proposed is leveraging two of the most popular open source automation tools, namely Foreman [10] and Puppet [11], making use as much as possible of the official OpenStack Puppet modules [12], as well as of other community supported Puppet modules for services like MySQL/Percona [13], Ceph [14], and others.

3. Roles and profiles using Puppet

Following the Puppet documentation [15], we choose to use the roles and profiles method as it is one of the most reliable ways to build reusable, configurable, and refactorable system configurations. In the present work the Puppet modules used for the configuration of the different components and services are those available in the official Puppet repository and provided by the community members. Some of the Puppet modules, like the ones for Ceph and Percona/MySQL, have been slightly modified (due to some missing features), and have been provided as internal OCP modules, made available via the INFN Gitlab public repository [16].

An overview of the infrastructure architecture can be seen in Figure 1 where all the main components and services are present and here briefly described:

- **Master Node**: it hosts the configuration management services, like Foreman and Puppet, used to install and configure the whole OCP-IaaS. At present, the configuration management tools have to be manually configured. The same node can eventually host a Zabbix [17] monitoring server for resources and application monitoring. The Zabbix server has been designed as a Puppet Role named **monit_server**. In the same way,the Zabbix agents on the OCP-IaaS nodes have been designed as a Puppet Role named **monit_agent**.
- **RHMK Nodes**: they provide external OpenStack services such as: (i) database cluster services (Percona/Mysql [13] and MongoDB [18]); (ii) a messaging system implementing

the AMQP protocol to let the Openstack services to connect and horizontally scale in case of an increase of demands (RabbitMq [19]) and (iii) a set of services for High Availability and Load Balancing (HAproxy [20], Zookeeper [21] and Keepalived [22]). All the RHMK services have been identified in Puppet with a Role named **rhmk**, while each individual mentioned service is represented by a Profile.

- **Storage Nodes**: Ceph, version Hammer [23], has been chosen as a block distributed storage platform whereby a minimum of three nodes are configured to ensure data replication. Ceph has been also used as a distributed object storage by configuring the Ceph Object Gateway [24], an interface compatible with OpenStack Swift [25] and Amazon S3 [26] services. The Ceph storage service has been designed as a Puppet Role named **Storage**.
- **Controller and Network Nodes**: they contain the common OpenStack services which are defined as Puppet Profiles. The configuration variables are hosted in the Puppet Role named **Controller&Network**. The service is designed to run in High Availability setup but it can be deployed on a single node also. **Controller** and **Network** services can be split in different nodes if required by architectural needs and each service is designed to run as single server as well as part of a HA cluster. OCP Network currently support the configuration of five different networks as from the OpenStack documentation [27].
- **Compute Node**: it contains the Nova services which are defined as Puppet Profiles. The configuration variables are hosted in the Puppet Role named **Compute** that permits to deploy the node and add it to the OpenStack infrastructure at any time.

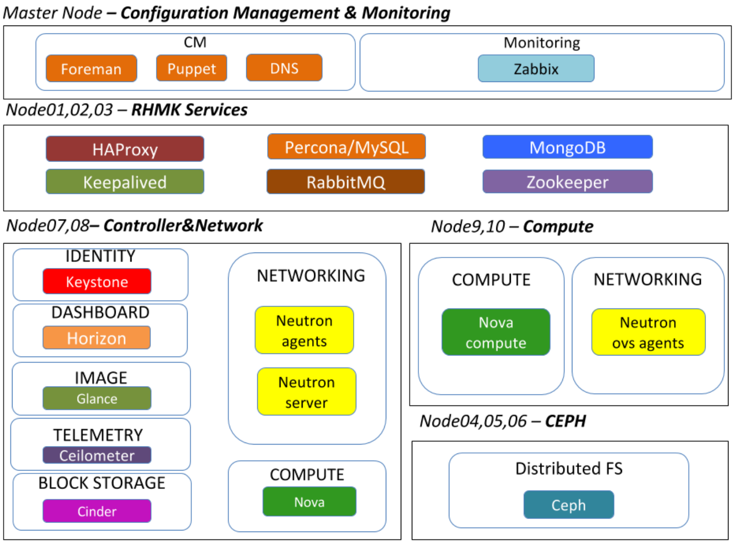


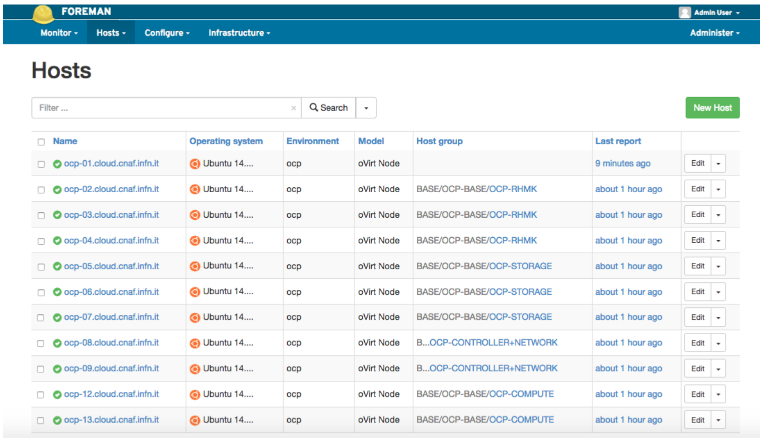
Figure 1. Overview of the OCP Infrastructure architecture and related Services.

4. Infrastructure deployment using the Foreman GUI

As already mentioned, the proposed solution is leveraging also the Foreman [10], an open source life cycle systems management tool that provides a Web-GUI for the IaaS installation and configuration precess.

The implementation of Foreman available and suitable for our project purposes was the 1.10.1. Foreman has been used as it is, leveraging on its native features such as the **Host Group**, a collection of user-selected classes and parameters through which each Puppet Role (and the related class parameters) has been defined.

After the OCP-IaaS installation and configuration process is terminated and all the services are up and running on the related hosts, Foreman can be used also to control and maintain the configuration status of the nodes or reconfigure them in case of errors due to misconfiguration, including therein the worst case where a complete reinstallation is needed (see Figure 2 for details).



Hosts

□ Name	Operating system	Environment	Model	Host group	Last report	
□ ocp-01.cloud.cnaf.infn.it	Ubuntu 14....	ocp	oVirt Node		9 minutes ago	Edit
□ ocp-02.cloud.cnaf.infn.it	Ubuntu 14....	ocp	oVirt Node	BASE/OCP-BASE/OCP-RHMK	about 1 hour ago	Edit
□ ocp-03.cloud.cnaf.infn.it	Ubuntu 14....	ocp	oVirt Node	BASE/OCP-BASE/OCP-RHMK	about 1 hour ago	Edit
□ ocp-04.cloud.cnaf.infn.it	Ubuntu 14....	ocp	oVirt Node	BASE/OCP-BASE/OCP-RHMK	about 1 hour ago	Edit
□ ocp-05.cloud.cnaf.infn.it	Ubuntu 14....	ocp	oVirt Node	BASE/OCP-BASE/OCP-STORAGE	about 1 hour ago	Edit
□ ocp-06.cloud.cnaf.infn.it	Ubuntu 14....	ocp	oVirt Node	BASE/OCP-BASE/OCP-STORAGE	about 1 hour ago	Edit
□ ocp-07.cloud.cnaf.infn.it	Ubuntu 14....	ocp	oVirt Node	BASE/OCP-BASE/OCP-STORAGE	about 1 hour ago	Edit
□ ocp-08.cloud.cnaf.infn.it	Ubuntu 14....	ocp	oVirt Node	B...OCP-CONTROLLER+NETWORK	about 1 hour ago	Edit
□ ocp-09.cloud.cnaf.infn.it	Ubuntu 14....	ocp	oVirt Node	B...OCP-CONTROLLER+NETWORK	about 1 hour ago	Edit
□ ocp-12.cloud.cnaf.infn.it	Ubuntu 14....	ocp	oVirt Node	BASE/OCP-BASE/OCP-COMPUTE	about 1 hour ago	Edit
□ ocp-13.cloud.cnaf.infn.it	Ubuntu 14....	ocp	oVirt Node	BASE/OCP-BASE/OCP-COMPUTE	about 1 hour ago	Edit

Figure 2. Foreman Web-GUI: A way to control hosts configuration status.

5. Future work

The positive results obtained and the experience gained during the testing phase, led us to investigate new semi-automatic procedures able to install and configure a complete OCP-IaaS providing the full support to the OpenStack Identity API v3 [28]. As a challenge, we are designing the semi-automatic tool to be used for performing the upgrade of the OCP-IaaS layer currently in use by a Data Center, to a new OpenStack version. Moreover, automated tools devoted to the installation and configuration of the PaaS layer (the so called CloudFormation as a Service) are under investigation and will be part of future enhancements.

References

[1] OpenCityPlatform Project, *http://www.opencityplatform.eu/*

[2] PRISMA project,*http://www.ponsmartcities-prisma.it/*

[3] MCloud project,*http://www.ecommunity.marche.it/AgendaDigitale/MCLoud/Obiettivi/tabid/206/Default.aspx*

[4] OpenStack, *http://www.OpenStack.org/*

[5] C. Aiftimiei, M. Antonacci, G. Donvito, E. Fattibene, A. Italiano, D. Michelotto, D. Salomoni, S. Traldi, C. Vistoli and G. Zizzi: Provisioning IaaS for the Open City Platform project. INFN-CNAF Annual Report 2014, pp 149.155 (2014) ISSN 2283-5490

[6] The NIST definition of Cloud Computing, Special Publication 800-145, *http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf*

[7] Fuel framework, *https://www.mirantis.com/products/mirantis-openstack-software/*

[8] OpenStack Juno, *https://www.OpenStack.org/software/juno/*

[9] Canonical, *https://wiki.ubuntu.com/TrustyTahr/ReleaseNotes/14.04*

[10] Foreman framework, *http://theforeman.org/*

[11] Puppet, *https://puppetlabs.com/*

[12] OpenStack Puppet, *https://wiki.openstack.org/wiki/Puppet*

[13] Percona cluster, *https://www.percona.com/software/mysql-database/percona-server*

[14] CEPH storage platform, *http://ceph.com/*

[15] Puppet Documentation, *https://docs.puppet.com/pe/2016.2/r_n_p_intro.html*
[16] OCP Internal Repository, *https://baltig.infn.it/groups/ocp-tools*
[17] Zabbix monitoring service, *http://www.zabbix.com/*
[18] MongoDB no SQL database, *https://www.mongodb.com/*
[19] RabbitMQ messaging service, *https://www.rabbitmq.com*
[20] Haproxy TCP/HTTP Load Balancer, *www.haproxy.org*
[21] Apache ZooKeeper, *https://zookeeper.apache.org/*
[22] Keepalived, *www.keepalived.org/*
[23] CEPH V0.94 (Hammer), *http://ceph.com/releases/v0-94-hammer-released/*
[24] CEPH object storage, *http://docs.ceph.com/docs/master/radosgw/*
[25] OpenStack Swift, *https://wiki.openstack.org/wiki/ReleaseNotes/Juno#OpenStack_Object_Storage_.28Swift.29*
[26] Amazon Simple Storage Service, *https://aws.amazon.com/it/s3/*
[27] OpenStack Networking documentation, *http://docs.openstack.org/security-guide/networking/architecture.html*
[28] OpenStack API v3, *http://developer.openstack.org/api-ref/identity/v3/*

Cloud@CNAF — the road to Juno

C. Aiftimiei[1,2], R. Bucchi[1], A. Costantini[1], D. Michelotto[1], M. Panella[1], D. Salomoni[1] and G. Zizzi[1]

[1]INFN CNAF, Viale Berti Pichat 6/2, 40126 Bologna, Italy
[2]IFIN - "Horia Hulubei", Bucharest - Magurele, Romania

E-mail: `cristina.aiftimiei@cnaf.infn.it alessandro.costantini@cnaf.infn.it`

Abstract.
Cloud@CNAF is a project aiming to offer a production quality Cloud Infrastructure, based on open source solutions to serve the different CNAF use cases. The project is the result of the collaboration of a transverse group of people from all CNAF departments: network, storage, farming, national services, distributed systems. The Cloud@CNAF IaaS (Infrastructure as a Service) based on OpenStack [1], a free and open-source cloud-computing software platform, has undergone in 2015 an upgrade not only on the version deployed, now Juno[2], but also in the quality of the services offered. As the infrastructure is used by many projects, like EEE [3], !CHAOS [4], OpenCityPlatform [5], INDIGO-DataCloud [6], and also by many local users and collaborations, there was a need to ensure a higly available setup in order to guarantee production quality services, minimizing the downtimes in case of problems. This paper presents the activity carried out at CNAF to upgrade the infrastructure together with a perspective of its evolution.

1. Introduction

The main goal of Cloud@CNAF project is to provide a production quality Cloud Infrastructure for CNAF internal activities as well as national and international projects hosted at CNAF:

- Internal activities
  - Provisioning VM for CNAF departments
    - Backend for provisioning of VMs to other services/processes like Jenkins[7], LBaaS[8]
    - Provisioning of VM for the User Support service (VirtualBox-like)
  - Provisioning of VM for CNAF staff members
- National and international projects
  - Providing VMs for experiments hosted at CNAF, like CMS and ATLAS
  - testbeds for the OpenCityPlatform project
    - testing of the components of particular interest, developed through the project, like the Automatic Installation and Configuration tool, based on Puppet [9] and Foreman [10] and tests of CEPH[11] installation and configuration
  - VMs for testing the solutions provided by the INDIGO - DataCloud project:
    - Data Center solutions, like TOSCA-in-HEAT [12], OpenStack clients with support for INDIGO solutions [13], Nova-Docker [14] or Partition Director [15]
    - Data Solutions, like OneData [16]

    * Automated Solutions, like CloudInfoProvider [17] or the Orchestrator [18]
    * Common Solutions, like INDIGO IAM [19]

The infrastructure made available is based on OpenStack, an open source product that can be deployed on open source platforms and has strong support from the industry. During 2015 an important work was done in order to upgrade this infrastructure not only from the point of view of the version deployed, upgrading to Juno version the latest available at the moment, but also for what regards the quality of the service offered, providing a highly available infrastructure, deploying all the services using a High-Availability (HA) setup or in a clustered manner (for ex. for the DBs used).

2. IaaS upgrade

During 2014 the first "version" of the infrastructure was made available to users, but several drawbacks were seen, like:

- Missing a High-Availability setup, in order to decrease the downtimes seen by the users in case of problems
- Problems with OpenVSwitch [20], like initial malfunctioning of the security-groups and periodic loss of instances connectivity
- No SSL/TLS on API endpoint
- Strong structural limitations of Heat [21]
- Missing service: Cinder

Taking into account the fast (6-month) release cycle [22] of Openstack it was decided to upgrade from Havana version [23] to the latest version available at the moment, Juno taking also into consideration the high number of new features, 342, and bug fixes, 3219, between which of our interest were:

- Nova:
    - easier upgrades with less impact on the applications users are running
    - scheduling updates to support scheduling services and extensibility
- Neutron:
    - High Availability for the L3 Agent
    - L3 agent performance improvement and new plugins, like Brocade L3 routing plugin
- Keystone:
    - Federated authentication improvements
    - support for multiple identity backends
- Dashboard (Horizon):
    - ability to deploy Apache Hadoop [24] clusters in seconds
    - extending the RBAC system to support OpenStack projects Compute, Networking, and Orchestration
- Orchestration (Heat):
    - easier roll-back of a failed deployment and thorough cleanup, improved scalability
    - administrators can delegate resource creation privileges to non-administrative users

As a consequence of the fact that still the upgrade procedure is very complex, and it was foreseen to skip one release (the Icehouse), the deployment of a new infrastructure with redundant components (Controller and Network) and an increased number of hypervisor resources was planned. For stability reasons it was decided to base the Neutron services on Linux Bridge [25] instead of Open vSwitch. At the same time, the compute nodes were increased from

4 to 22 nodes, in this way making available a total computing power of about 350 CPUs and 1,3TB of RAM.

The GPFS [26] filesystem was used as backend storage for all OpenStack services, Nova, Cinder and Glance. Data persistence and the messaging service, AMQP [27], was guaranteed by a three nodes cluster hosting MySQL Percona [28] and RabbitMQ [29] software tools.

In the following paragraphs the caractheristics of both old and new infrastructures are presented, highlithing the HA setup for the different services:

Havana version:

- **1 Controller Node** providing the core services:
  - Keystone, Glance (LVM), Heat, Horizon, Ceilometer, MySQL, QPID, hosted on a physical machine with:
  - 2x8 HT (32) Intel(R) Xeon(R) CPU E5-2450 @ 2.10GHz, 64 GB RAM
- **1 Network Node** providing:
  - Neutron configured to use 100 VLANs with Open vSwitch virtual switch, hosted on a physical machine with:
  - 2x8 HT (32) Intel(R) Xeon(R) CPU E5-2450 @ 2.10GHz, 64 GB RAM
- **4x Compute Node** (for a total of 64 CPU and 256 GB of RAM) providing:
  - Nova services based on KVM/QEMU hypervisors, hosted on:
  - 2x8 core AMD, 64 GB RAM
- **Shared Storage** (Dell PowerVault StorageSystem + 3 GPFS servers) providing:
  - 16 TB on GPFS for Nova backend
- **1 Web Proxy** per la dashboard

Juno version:

- **2x Controller Node** HA active/active:
  - Keystone, Heat, Horizon, Ceilometer, Neutron server
  - HAProxy & Keepalived x API
  - Glance & Cinder, hosted on physical nodes with:
  - 2x8 HT (32) Intel(R) Xeon(R) CPU E5-2450 @ 2.10GHz, 64 GB RAM
- **2x Network Node** HA active/active partial (DHCP agent in active/active, L3 agent in hot standby)
  - Neutron with LinuxBridge + VLAN, hosted on physical nodes with:
  - 2x6 HT (24) Intel(R) Xeon(R) CPU E5-2450 0 @ 2.10GHz, 64 GB di RAM
- **22x Compute Node** = 352 CPU + 1,3 TB RAM
  - Nova per KVM/QEMU, LinuxBridge agent, hosted on physical nodes with:
  - 2x8 core AMD Opteron 6320 @ 2.8GHz con 64 GB RAM
- **Shared Storage** (PowerVault + 3 GPFS servers), providing:
  - 16TB on GPFS for Nova (instances), Glance (images), Cinder (volumes) backends
- **1 Web Proxy** for the Dashboard
- **3x Percona XtraDB, RabbitMQ, MongoDB, ZooKeeper**
- **2x HAProxy** for Percona (failover, no roundrobin!)

A cloud User Interface (UI) node is also available in order to guarantee the access to the infrastructure using the powerful OpenStack APIs [30] for an advanced use of OpenStack.

Currently, the vesion of the OpenStack APIs available on the cloud UI node is still Havana, as there are incompatibilities between the two versions. The availability of a new cloud UI providing the new APIs is expected for 2016.

In early 2015 more than 50 users were registered in the infrastructure, providing for them 50 tenants with 48 instances (VMs) that used about 90 VCPUs and 180GB of RAM. With the increase of resources made available, by 2016 the number of the VMs doubled (130 VM) using 644 GB of RAM and 316 VCPU.

The new cloud-infrastructure architecture is represented bellow, in Fig. 1

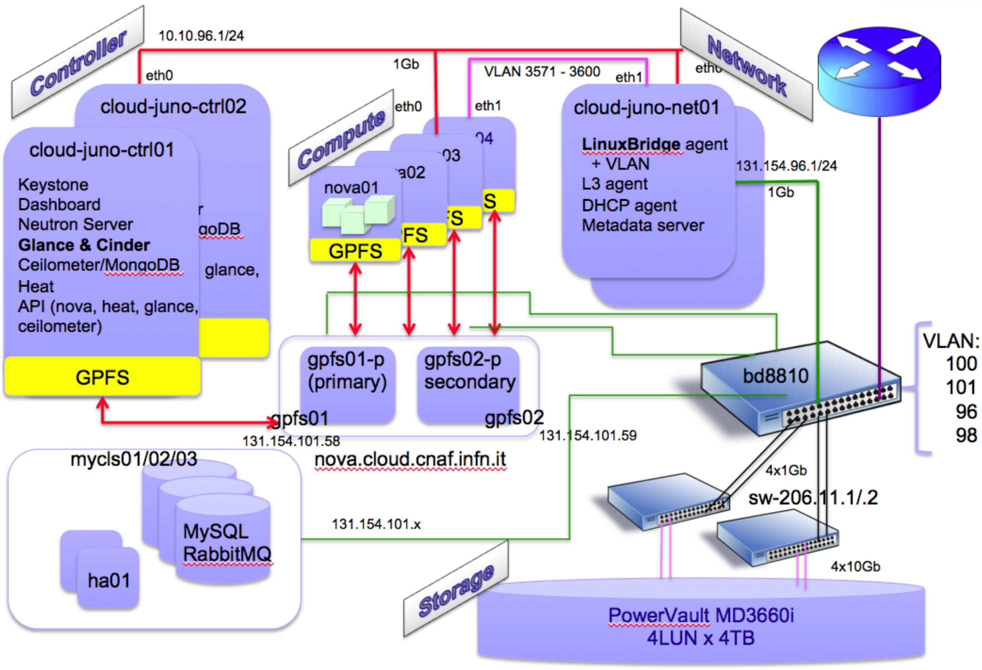


Figure 1. Cloud@CNAF v. Juno

3. Use cases

The active use cases working at CNAF can be separated in two sets, the first represents the internal CNAF use cases, and the second collects the external collaborations and projects in which CNAF is involved. In 2015 we saw an increase on the number of the projects that were served by the Cloud@CNAF infrastructure was registered, with the addition of projects like OCP and INDIGO - DataCloud, described bellow.

During the year all users and previous existing projects/tenants saw their instances migrated to the new infrastructure with a very small downtime required by the fact that running instances needed to be stopped, snapshot taken and only afterwards the resulted images were migrated between the two distinct infrastructures.

3.1. CNAF Internal use cases

This set of use cases collect internal work groups and single user's projects.

- The **Middleware Developers** team - composed by people active on the development of new software like EMI Middleware [31] for the Grid distributed computing resources, VOMS [32], StoRM [33], Argus [34] and INDIGO DataCloud IAM. Middleware developers activity on Cloud@CNAF is splitted between three tenants (projects):

  - *MW-CI - tenant for the Continuous Integration* : the instances of this project are coordinated by an external Jenkins master. The instances are used for building the software and deploying it (Installation and Upgrade). Moreover, some instances based on CoreOS [35] are implementing Docker [36] nodes to speed up test deployment.
  - *MW-TEST and MW-DEVEL - tenants hosting testing instances*: these projects are used to test various deployment scenarious of the developed components.
- **CNAF staff**: Every employee at CNAF can rely on a personal project that can be used for their R&D needs. Each personal project has a limited amount of resources: 3 instances, 6 VCPU, 12GB RAM, 4 Floating IPs and 10 security groups. At the time of writing there are about 36 active users having personal projects.

3.2. Collaborations & Projects

CNAF personnel is involved in several collaborations and projects, both national and international. Resources dedicated to tenants representing collaborations, experiments or projects are made available as requested at the creation time and increased later if the use is justified by the work to be performed.

- The **Extreme Energy Event** (EEE) experiment is devoted to the search of high energy cosmic rays through a network of telescopes installed in about fifty high-grade schools distributed throughout the Italian territory.
  - the same amount of the resources were dedicated also during 2015. An increase of the storage space is forseen for 2016.
- The **!CHAOS** activity was originally developed within the context of High Energy Physics (HEP) as a candidate of Distributed Control Systems (DCS) and Data Acquisition (DAQ) for the SuperB experiment. Since 2014 it evolved into the project *!CHAOS: a cloud of controls* supported by MIUR and developed by INFN through its four sites: Laboratori Nazionali di Frascati (LNF), Laboratori Nazionali del Sud (LNS), Padova Section and CNAF Centre.
- The **OpenCityPlatform (OCP)** - an industrial research project funded by the Italian Ministry of University and Research (MIUR), started in 2014, intending to research, develop and test new technological solutions open, interoperable and usable on-demand in the field of Cloud Computing.
  - The CNAF-team was responsible of the design and development of an automatic tool based on Puppet and Foreman able to simplify the installation and configuration of the OCP Cloud-platform, starting from the IaaS level. The Cloud@CNAF infrastructure become the primary testbed of this automated solution, and it is also part of the OCP project-wide testbed aimed in testing also other OCP-solutions, like the one offered for the management of Big Data, through the deployment of some of the most widespread and promising open source technologies, in particular Hadoop [24] & Apache Mesos [37]
- The **INDIGO - DataCloud** - a project started in April 2015, funded under the EC Horizon 2020 framework program, addressesing the challenge of developing open source software, deployable in the form of a data/computing platform, aimed to scientific communities and designed to be deployed on public or private Clouds and integrated with existing resources or e-infrastructures. Cloud@CNAF infrastructure was used for:
  - the first tests of Apache Mesos & Marathon, and Calico project [38] for the networking
  - preparation of Docker containers for building HTCondor clusters [39]

4. Future Work

During the next year effort will be put in the definition and test of an upgrade procedure, on a parallel, dedicated testbed, through which fast and "painless" upgrades, at least once per year, can be performed and low downtimes for the users can be garanteed. A hardware upgrade for the new inftastructure is also planned.

Next year will see the first releases of the INDIGO-DataCloud components and many of them, like the Identity and Access Management Service, the OneData solution to manage, share and access remote data, the geographically distributed Batch System as a Service, exploiting HTCondor and Docker container technology, will be tested and deployed on the Cloud@CNAF infrastructure, as part of the INDIGO Pilot Services testbed.

References

[1] OpenStack, *http://www.OpenStack.org/*
[2] OpenStack Juno, *https://www.openstack.org/software/juno/*
[3] Extreme Energy Events (EEE) project, *http://eee.centrofermi.it/*
[4] !CHAOS: a cloud of controls - MIUR project proposal, *https://www.lnf.infn.it/sis/preprint/detail.php?id=5349*
[5] OpenCityPlatform (OCP), *http://opencityplatform.eu/*
[6] INDIGO - DataCloud, *https://www.indigo-datacloud.eu/*
[7] Jenkins - Continuous Integration System, *https://jenkins-ci.org/*
[8] Load Balancer as a Service (LBaaS) *http://docs.openstack.org/liberty/networking-guide/adv-config-lbaas.html*
[9] Puppet, *https://puppet.com/*
[10] Foreman, *https://theforeman.org/*
[11] CEPH, *http://ceph.com/*
[12] TOSCA-in-HEAT, *https://indigo-dc.gitbooks.io/indigo-datacloud-releases/content/indigo1/heat-translator1.html*
[13] OpenStack Clients supporting INDIGO solutions, *https://indigo-dc.gitbooks.io/indigo-datacloud-releases/content/indigo1/datacenteri_solutions.html*
[14] Nova-Docker, *https://indigo-dc.gitbooks.io/indigo-datacloud-releases/content/indigo1/nova-docker1.html*
[15] Partition Director, *https://indigo-dc.gitbooks.io/indigo-datacloud-releases/content/indigo1/dynpart1.html*
[16] OneData, *https://indigo-dc.gitbooks.io/indigo-datacloud-releases/content/indigo1/onedata1.html*
[17] CloudInfoProvider, *https://indigo-dc.gitbooks.io/indigo-datacloud-releases/content/indigo1/cip1.html*
[18] Orchestrator, *https://indigo-dc.gitbooks.io/indigo-datacloud-releases/content/indigo1/orchestrator1.html*
[19] IAM, *https://indigo-dc.gitbooks.io/indigo-datacloud-releases/content/indigo1/iam1.html*
[20] Open vSwitch, *http://openvswitch.org/*
[21] Heat, *https://wiki.openstack.org/wiki/Heat*
[22] OpenStack Release, *https://releases.openstack.org/*
[23] OpenStack Havana, *https://www.OpenStack.org/software/havana/*
[24] Apache Hadoop, *http://hadoop.apache.org/*
[25] Linux Bridge, *https://wiki.linuxfoundation.org/networking/bridge*
[26] GPFS aka Spectrum Scale, *http://www-03.ibm.com/systems/uk/storage/spectrum/scale/*
[27] AMQP, *https://www.amqp.org/*
[28] Percona, *https://www.percona.com/*
[29] RabbitMQ, *https://www.rabbitmq.com/*
[30] OpenStack APIs, *http://developer.openstack.org/api-guide/quick-start/api-quick-start.html*
[31] EMI, *http://www.eu-emi.eu*
[32] VOMS, *http://italiangrid.github.io/voms/*
[33] StoRM, *http://italiangrid.github.io/storm/*
[34] Argus, *http://argus-authz.github.io/*
[35] CoreOS, *https://coreos.com/*
[36] Docker, *https://www.docker.com/*
[37] Apache Mesosphere (Mesos & Marathon), *https://www.digitalocean.com/community/tutorials/an-introduction-to-mesosphere*
[38] Calico, *https://www.projectcalico.org/*
[39] HTCondor, *https://research.cs.wisc.edu/htcondor/*

The !CHAOS Project at CNAF

C Aiftimiei[1], E Fattibene[1], M Panella[1], D Salomoni[1]
in collaboration with:
S Angius[2], C Bisegni[2], P Buzzi[3], L Catani[4], S R Cavallaro[5], B Checcucci[3], P Ciuffetti[2], B F Diana[5], C Di Giulio[4], G Di Pirro[2], F Enrico[5], L G Foggetta[2], F Galletti[2], R Gargana[2], E Gioscio[2], P Lubrano[3], D Maselli[2], G Mazzitelli[2], A Michelotti[2], M Michelotto[6], R Orrù[2], M Piccini[3], M Pistoni[2], S Pulvirenti[5], G Salina[4], F Spagnoli[2], D Spigone[2], A Stecchi[2], T Tonto[2], M A Tota[2]

[1] INFN CNAF (Centro Nazionale Tecnologie Informatiche)
[2] INFN-LNF (Laboratori Nazionali di Frascati)
[3] INFN-PG (Sezione di Perugia)
[4] INFN-TV (Sezione di Tor Vergata)
[5] INFN-LNS (Laboratori Nazionali del Sud)
[6] INFN-PD (Sezione di Padova)

E-mail: `enrico.fattibene@cnaf.infn.it matteo.panella@cnaf.infn.it`

Abstract. The “!CHAOS: a cloud of controls” project financed by MIUR (Italian Ministry of Research and Education) aimed to develop a scalable and distributed system for process control and data acquisition based on a dynamic cloud-like infrastructure. The final achievement of the project, ended in December 2015, is a prototype of a “Control-as-a-Service” open platform for scientific and industrial applications, deployable on OpenStack, a free and open-source software platform for cloud computing, with strong support from the industry. The CNAF team’s work concentrated on the development of the infrastructure orchestration system. It had also a big contribution on the testing activities by running full-scale tests of !CHAOS software on the Cloud@CNAF OpenStack infrastructure.

1. Introduction

The !CHAOS project started as a candidate Distributed Control System (DCS) and Data Acquisition (DAQ) system for the SuperB INFN Flagship Project and subsequently evolved into a fully autonomous project “!CHAOS: a cloud of controls” [1] supported by MIUR and developed by INFN through its four sites Laboratori Nazionali di Frascati (LNF), Laboratori Nazionali del Sud (LNS), Sezione di Padova and CNAF. National Instruments and ESCO also supported the project for the implementation of several case studies. The project goal was to deliver an open and scalable platform based on a cloud infrastructure suitable for scientific and industrial use cases.

CNAF team has been involved in the Work Package (WP) 5 “IT infrastructure development and implementation”, that worked on the design and implementation of the cloud infrastructure and services needed in order to offer the !CHAOS framework as a service.

2. The !CHAOS Cloud infrastructure

2.1. Software components

The !CHAOS software framework can be deployed in a wide range of setups, ranging from bare-metal setups for high-throughput DAQ (tests have been successfully performed on the Beam Test Facility of the INFN LNF DAFNE accelerator complex) to cloud-based deployments for highly distributed systems like environmental monitoring systems.

The framework itself requires a number of basic IT services in order to run:

- a non-relational database for runtime configuration data
- a fast key-value store for keeping track of live process state
- a shared filesystem with POSIX semantics for long-term data archival
- a VPN concentrator for remote access (optional)

The entire software stack for basic IT services uses exclusively open-source software.

On these backend services rely the !CHAOS [2] main services (!CHAOS Data Service - CDS and MetaData Service - MDS). The !CHAOS team studied a solution for each of these services (both the common and the !CHAOS specific ones), in order to integrate them in a generic IaaS based on OpenStack. The solution is based on Heat [3], the OpenStack component that implements the function of infrastructure orchestration, i.e. a programmatic approach to create and deploy full stack configurations.

2.2. Orchestration

The main focus of the CNAF contribution has been developing an automated orchestration system capable of starting up a complete !CHAOS deployment (see figure 1) on an OpenStack cloud environment and optionally allow the system to scale as needed with little to no human intervention.

The orchestration is performed by OpenStack Heat [3] using a template which describes all the resources needed by the !CHAOS deployment and the software deployment steps of the configuration management system (SaltStack [4]). Once all virtual resources have been deployed, SaltStack takes over and performs the deployment of basic IT services and the !CHAOS application components.

The complexity of setting up application clusters for the basic IT services is hidden from the deployer and handled entirely by the configuration management system. As a side-effect, the setup time is significantly shorter than manual setup performed by a trained system administrator (on average, a manual setup would require a week) while the automated setup takes 30 minutes.

Furthermore, the orchestration system offers the capability to scale up some IT services as needed, allowing an already existing !CHAOS cloud deployment to cope with higher throughput and storage needs without having to perform complex manual operations on the deployed clusters.

The main Heat template is composed of a series of sub-templates for automated deployment of major services of !CHAOS in a cloud environment. Two templates describe the deployment of database services required by !CHAOS: a MongoDB cluster for runtime configuration management and a Couchbase cluster for live data caching. The template for MongoDB was derived from an open-source implementation by Rackspace and is capable of performing cluster bring-up in a completely unattended way. The Couchbase cluster template was partially derived from the MongoDB template, but the software configuration was entirely developed in-house. The setup is again fully unattended and can handle horizontal scaling of the Couchbase cluster via HTTP POST requests to an endpoint provided by the OpenStack Orchestration Service, without the need for human intervention on the Couchbase cluster itself.

Network access services are implemented as a virtualized OpenVPN server and a Heat template that configures the OpenStack component to manage virtual network services (Neutron) to route

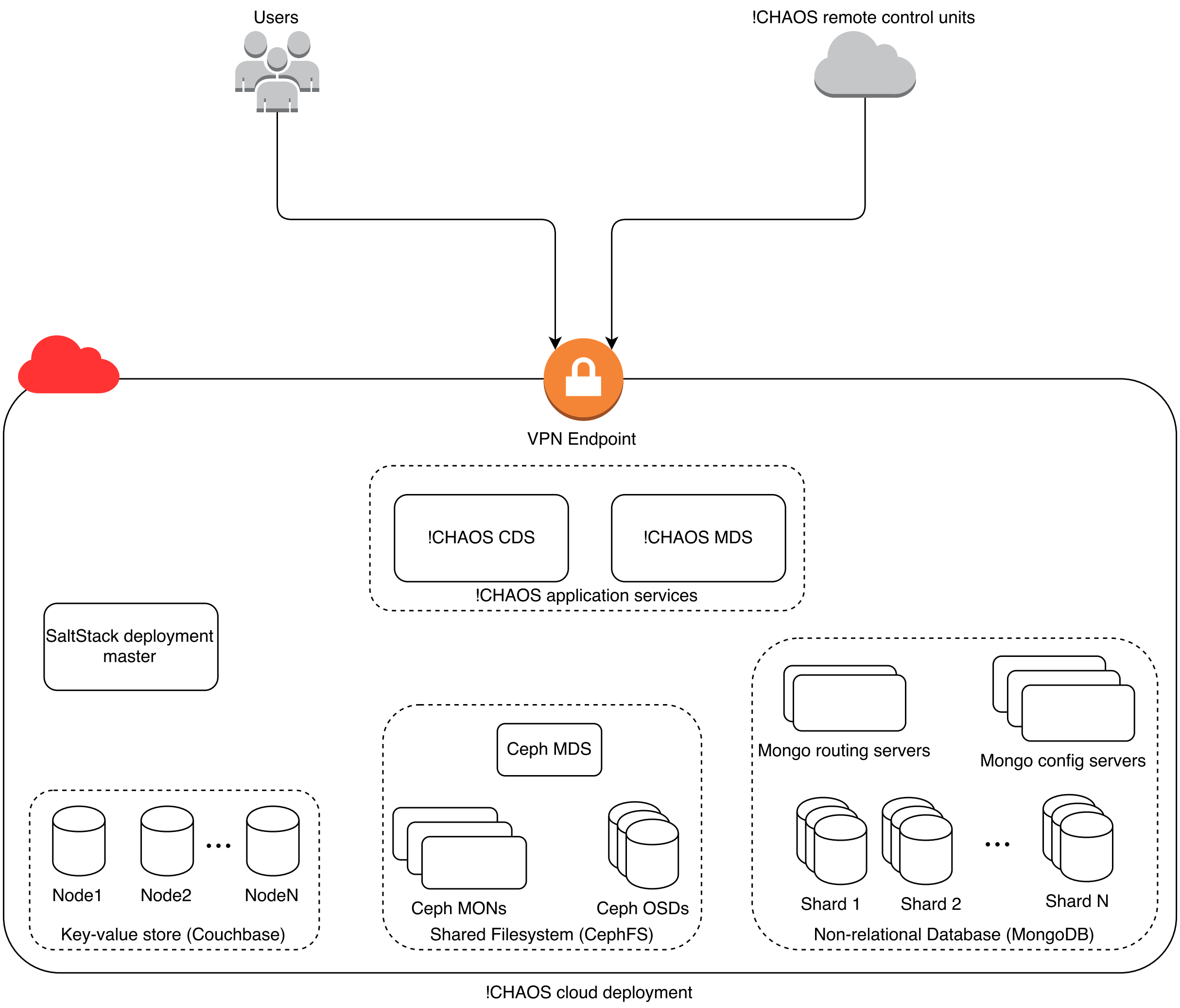


Figure 1. !CHAOS cloud deployment architecture

traffic to/from the VPN. Additional security can be guaranteed by Neutron services (like Firewall-as-a-Service) if supported by the target cloud.

The shared POSIX filesystem is a virtualized Ceph cluster. It provides a 3-way replicated POSIX-like filesystem to !CHAOS services for archiving purposes. The same cluster may also act as an object storage if so configured. The setup is fully unattended and the filesystem size is only limited by the available quota on the target cloud. The !CHAOS application components are deployed using a specialized template that populates the configuration files automatically with the major backend services addresses. Moreover, the shared POSIX filesystem is mounted automatically on all Data Service nodes.

The final Heat template can run on OpenStack Juno with Cinder component enabled and requires a heat-softwareconfig-enabled Ubuntu 14.04 image.

In order to simplify the usage of the specialized Heat template, a prototype of a web interface has been developed in PHP language, exploiting the Heat APIs. Through this interface the user can instantiate a complete !CHAOS infrastructure, choosing the filesystem size and the number of MongoDB shards and Couchbase instances. As soon as the infrastructure is running, the web interface offers the possibility to scale the infrastructure, adding an instance to the Couchbase

Service	Subsystem	Instance count
Couchbase	Master	1
	Slaves	2
Ceph	MON	3
	MDS	1
	OSD	3
MongoDB	Routing servers	2
	Configuration servers	3
	Shard 1 replica set members	3
	Shard 2 replica set members	3
	Shard 3 replica set members	3
!CHAOS	!CHAOS Data Service (CDS)	1
	MetaData Service (MDS)	1
VPN	VPN Server	1
SaltStack	Deployment master	1
	TOTAL	**28**

Table 1. Resource requirements for a full-scale deployment

cluster.

2.3. System tests on Cloud@CNAF

Due to the large amount of resources needed by a full-scale deployment, the !CHAOS project has also been instrumental in providing feedback and validation of the Cloud@CNAF OpenStack cloud infrastructure. The resources needed by such a deployment are summed up in table 1.

The tests have been performed using an OpenStack tenant in a completely empty state, so every resource has been created from scratch using the Heat orchestration template. In such conditions the average startup time is 30 minutes, mainly spent downloading and installing the required software components on each virtual machine.

3. Conclusions

The work performed at CNAF during 2015 contributed to the result of a "!CHAOS as a service", deployable on open-source cloud infrastructures. This solution reduces the time of framework deployment and permits to instantiate all the !CHAOS frontend and backend services with minimal human intervention.

Possible future improvements could be the deployment of services on different regions (i.e. IaaS running on different data centres) to provide geographical redundancy for services and a solution to auto-scale deployed infrastructures on the basis of monitoring metrics calculated on the running !CHAOS application instance.

Moreover, the approach of abstracting the complexity of application deployment through an orchestration service and an overhead web interface can be applied to other use cases for datacenter users and administrators, i.e. deployment of recurring infrastructures or testbeds.

4. References

[1] F. Antonucci et al, “!CHAOS: a cloud of controls - MIUR project proposal”, INFN-14-15/LNF; https://www.lnf.infn.it/sis/preprint/

[2] L. Catani et al, “Introducing a new paradigm for accelerators and large experimental apparatus control systems”, Phys. Rev ST Accel. Beams 15, 112804 - Published 29 November 2012

[3] OpenStack Heat: http://docs.openstack.org/developer/heat/

[4] SaltStack: https://saltstack.com/community/

Middleware support, maintenance and development

A Ceccanti, E Vianello, M Caberletti and F Giacomini

INFN-CNAF, Bologna, Italy

E-mail: andrea.ceccanti@cnaf.infn.it, enrico.vianello@cnaf.infn.it, marco.caberletti@cnaf.infn.it, francesco.giacomini@cnaf.infn.it

Abstract.
INFN-CNAF plays a major role in the support, maintenance and development activities of key middleware components (VOMS, StoRM and Argus) widely used in the WLCG [2] and EGI [1] computing infrastructures. In this report, we discuss the main activities performed in 2015 by the CNAF middleware development team.

1. Introduction

The CNAF middleware development team has focused, in 2015, on 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 [5]: the lightweight storage element in production at the CNAF Tier1 and in several other WLCG sites;
- Argus [4]: the Argus authorization service, which provides a centralized management and enforcement point for authorization policies on the Grid.

The main activities for the year centered around support and maintenance of the software and on the improvement of the continuous integration and testing processes.

2. The Continuous Integration and Testing Infrastructure

All the code for VOMS, StoRM and Argus is hosted on Github [7]. The software is continuously built and tested by our Jenkins [6] server, which is configured with build nodes for the main supported platforms (Scientific Linux 5 and 6 and CentOS 7).

The Github Webhooks [16] mechanism provides efficient integration between the code repository and our CI server, so that whenever a change is pushed to one of our software repositories a new build job is started on our CI infrastructure.

During 2015, significant effort was devoted on introducing Docker [31] as a central component in our continuous integration infrastructure.

We focused on the following tasks:

- Run the build and packaging for our software inside a Docker container;
- Run deployment tests inside a Docker container.

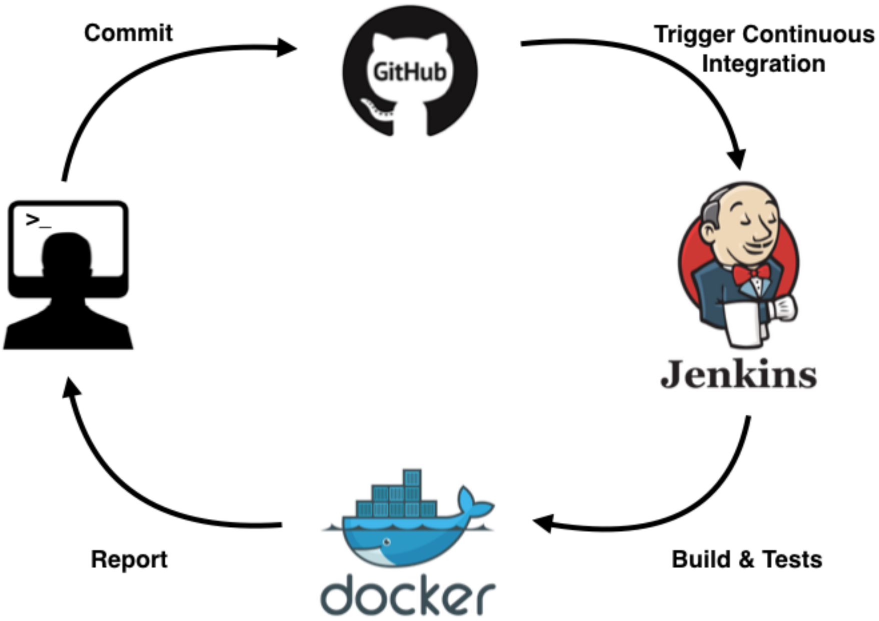


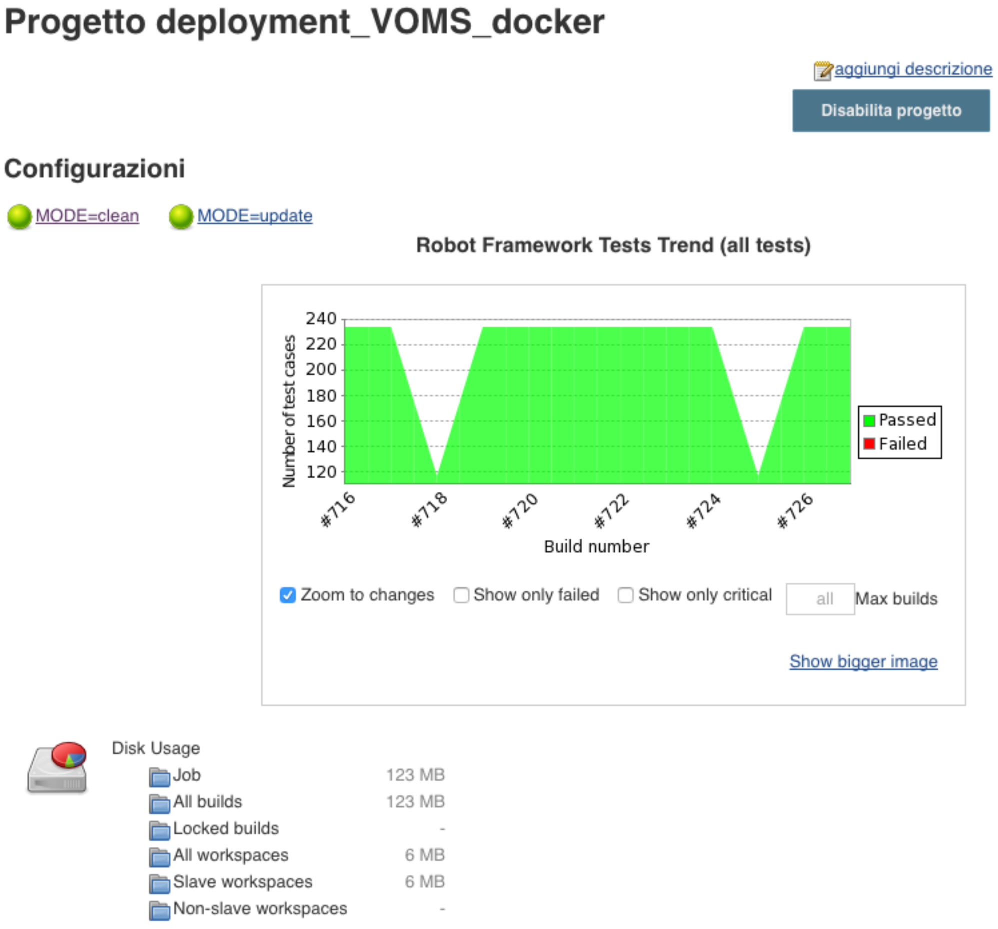
Figure 1. The Continuous Integration feedback loop

2.1. Dockerized build and packaging
Relying on Docker containers for the build and packaging of software gives the following advantages:

- a clear definition of the build environment, via the Dockerfile;
- build isolation: builds run in a clean environment which is instantiated from "scratch" at each build;
- the ability to support different platforms without needing a different VM for each platform.

As a first step, we started investigating which Linux distribution would best support the execution of Docker containers. After observing several issues in Docker support on CentOS, we moved to CoreOS [33], a Linux distribution optimized for the execution of Linux containers. The peculiarity of CoreOS is that it does not provide a package manager like Red-Hat or Debian-based distributions, and that all user applications run inside containers. A CoreOS VM can be provisioned via a cloud-config [34] configuration, which provides the tools to bootstrap the system as desired (e.g., defining users, active services, etc..). We defined a cloud-config configuration [35] for our dockerized Jenkins slave nodes, and defined a Jenkins job that can automatically bootstrap a dockerized Jenkins slave on-demand, via a script [37]. This way we can easily scale or migrate our CI infrastructure to a different platform[1].

As a second step, we defined the basic build and package containers for our software [36]. The first product that was migrated was Argus and, learning from that experience, StoRM and VOMS followed. Another advantage is that developers can now easily build and package software on their workstation and expect the same results and output produced by the continuous integration server.



Progetto deployment_VOMS_docker

Configurazioni

● MODE=clean ● MODE=update

Robot Framework Tests Trend (all tests)

☑ Zoom to changes ☐ Show only failed ☐ Show only critical all Max builds

Disk Usage

Job	123 MB
All builds	123 MB
Locked builds	-
All workspaces	6 MB
Slave workspaces	6 MB
Non-slave workspaces	-

Figure 2. The Jenkins page for the VOMS deployment test job

2.2. Dockerized deployment tests

Deployments tests were also migrated to Docker, first by moving the execution of integration and load testsuites and then the execution of the services.

By using Docker, instead of VMs provisioned by the Cloud@CNAF OpenStack infrastructure as was done previously, we reduced the deployment test bootstrap time from minutes to seconds, and gained the ability to run several deployment tests in parallel and more insight into the deployment test outcome.

The other advantage is that now developers can execute deployment test directly on their workstation, for instance to check that everything behaves as expected before pushing code changes to an upstream product repository.

3. VOMS

During 2015, 61 new issues were opened in the VOMS issue tracker [10] to track maintenance, development and release activities. In the same period, 46 issues were resolved.

Development work on VOMS in 2015 focused on the evolution of VOMS admin server, leading to three official releases (3.3.2 [18], 3.3.3 [19], 3.4.0 [20]) providing several bug fixes

[1] This approach helped a lot in migrating the CI infrastructure from OpenStack Havana to Juno without incurring downtimes.

and introducing new features, mainly in support of the CERN VOMS deployment, which serves the main LHC experiments VOs.

Significant effort was also put in providing a self-contained development environment running as a set of Docker containers orchestrated via Fig [30] (later renamed to docker-compose[32]).

4. StoRM

During 2015, 52 new issues were opened in the StoRM issue tracker [21] to track maintenance, development and release activities. In the same period, 64 issues were resolved [22].

The main highlights for StoRM are:

- The releases of StoRM from v1.11.6 to v1.11.10 [23, 24, 25, 26, 27] providing fixes for several issues found in production and during development;
- The first release of the StoRM WebDAV service, within StoRM 1.11.7, which replaces the StoRM GridHTTPs service improving performance and stability of the service;
- Various optimizations of SQL queries to keep the load on the StoRM MySQL database under control;
- Evolution and dockerization of the StoRM functional and regression testsuite [14];
- Evolution and dockerization of the StoRM load testsuite [15];
- Evolution and dockerization of our deployment tests infrastructure [29].

In November, the workshop “Storage evolution for physics experiments: the role of StoRM” was organised to gather together StoRM developers, WLCG and LHC experiments representatives and CNAF Tier-1 staff in order to discuss the current status of StoRM and future development plans [28].

5. Argus

During 2015, work on Argus focused on the porting of the software on RedHat7-based operating systems (CentOS 7) and the move to Java 8.

The main activities were:

- Upgrade the core dependencies to the latest versions, such as Jetty 9, VOMS 3.1, CANL 2.2.0 and OpenSAML 2.6.4;
- Support to run Argus services under systemd on CentOS 7;
- The development of a functional and regression testsuite based on RobotFramework [39];
- The development of a load testsuite based on Grinder [40].

In the same period, maintenance work focused on resolving outstanding issues in the core Argus components. In particular, the main issues solved were:

- (PAP) Removed the hardcoded SSLv3 default value in `pap-admin` tool;
- (PAP) Add a further validation to avoid the loading of policy with an empty string as value for user subject;
- (PEPd) Fix a race condition in the mapping code that could lead to corrupt mappings.

Significant effort was put in the transition of the build and packaging for the software to Docker and the integration in our CI infrastructure [38].

6. Future work

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

- Evolution and refactoring of the StoRM services, to reduce code-base size and maintenance costs, to provide horizontal scalability and simplifiy the services management;
- Evolution of the VOMS attribute authority for better integration with SAML federations;
- Support for container-based execution for all our services.

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] The Argus authorization service website http://argus-authz.github.io
[5] The StoRM website http://italiangrid.github.io/storm
[6] Jenkins https://jenkins-ci.org/
[7] GitHub https://github.com/
[8] Openstack http://www.openstack.org
[9] INFN issue tracker https://issues.infn.it
[10] VOMS on INFN JIRA https://issues.infn.it/jira/browse/VOMS
[11] Argus on INFN JIRA https://issues.infn.it/jira/browse/ARGUS
[12] The European Middleware Initiative http://www.eu-emi.eu
[13] The VOMS clients testsuite https://github.com/italiangrid/voms-testsuite
[14] The StoRM testsuite https://github.com/italiangrid/storm-testsuite
[15] The StoRM load testsuite https://github.com/italiangrid/grinder-load-testsuite
[16] Github webhooks https://help.github.com/articles/about-webhooks
[17] Github pages https://pages.github.com
[18] VOMS Admin 3.3.2 http://italiangrid.github.io/voms/release-notes/voms-admin-server/3.3.2/
[19] VOMS Admin 3.3.3 http://italiangrid.github.io/voms/release-notes/voms-admin-server/3.3.3/
[20] VOMS Admin 3.4.0 http://italiangrid.github.io/voms/release-notes/voms-admin-server/3.4.0/
[21] StoRM issues created in 2015 https://issues.infn.it/jira/issues/?filter=14710
[22] StoRM issues resolved in 2015 https://issues.infn.it/jira/issues/?filter=14712
[23] StoRM v. 1.11.6 http://italiangrid.github.io/storm/2015/01/20/storm-v.1.11.6-released.html
[24] StoRM v. 1.11.7 http://italiangrid.github.io/storm/2015/02/09/storm-v.1.11.7-released.html
[25] StoRM v. 1.11.8 http://italiangrid.github.io/storm/2015/03/13/storm-v1.11.8-released.html
[26] StoRM v. 1.11.9 http://italiangrid.github.io/storm/2015/05/29/storm-v1.11.9-released.html
[27] StoRM v. 1.11.10 http://italiangrid.github.io/storm/2016/01/22/storm-v1.11.10-released.html
[28] StoRM Workshop https://agenda.cnaf.infn.it/conferenceDisplay.py?confId=750
[29] StoRM deployment tests https://github.com/italiangrid/docker-scripts/tree/master/storm-deployment-test
[30] Fig http://www.fig.sh/
[31] Docker https://www.docker.com/
[32] Docker compose https://docs.docker.com/compose/
[33] CoreOS https://coreos.com
[34] CoreOS cloud-config https://coreos.com/os/docs/latest/cloud-config.html
[35] Our Jenkins Slave cloud-config https://github.com/italiangrid/ci-scripts/commits/master/openstack/coreos-cloudinit
[36] The Docker images repository https://github.com/italiangrid/docker-scripts
[37] Jenkins slave bootstrap script https://github.com/italiangrid/ci-scripts/blob/master/openstack/start-docker-slave.sh
[38] Argus dockerized packaging code https://github.com/argus-authz/pkg.argus
[39] The Argus regression testsuite https://github.com/argus-authz/argus-robot-testsuite
[40] The Argus Load testsuite https://github.com/argus-authz/load-testsuite

Continuous assessment of software characteristics: a proof of concept

E Ronchieri and F Giacomini

E-mail: elisabetta.ronchieri@cnaf.infn.it, francesco.giacomini@cnaf.infn.it

Abstract. Software characteristics detail various aspects of software, such as sustainability, maintainability and usability. Continuous integration allows software development team members to automate their assessment process, orchestrating the measurement of software product metrics, the static analysis of software and the statistical analysis of data collected. By adopting this technique a team can detect and identify errors quickly and more easily.

In this report we describe a successful integration of a subset of software metrics and static code analysis tools in a continuous system, called Jenkins, as planned in the INFN CCR Uncertainty Quantification project. We have used various Geant4 releases as guinea pig code to be assessed.

1. Introduction

Existing standard, called ISO/IEC 25010:2011 [1], has defined software characteristics, such as functionality, efficiency, compatibility, usability, reliability, security, maintainability and portability (relevant to all software), for evaluating and measuring software quality. For each characteristic this standard has identified a set of product metrics, whose measurements can be performed by commercial and free tools. According to literature [2] and our experience [3], these tools implement metrics differently: therefore, it is essential software team assess them before their usage [4]. In addition, there are other tools, belonging to the static analysis category, that allow developers to detect errors in the code (i.e., to have zero false positives).

In this report, we present the results we have obtained by integrating software metrics and static code analysis tools in continuous integration by combining Jenkins and Docker together [5]: Jenkins enables team members to focus on their work by automating the build, artifact management and deployment processes; Docker accelerates and optimizes the continuous integration pipeline. We have created docker images to envelop these tools, such as Cppcheck [6], Clang [7], cloc [8] and SLOCCount [9], in order to be used at build time by Jenkins' jobs and publish tools results on the correspondent project dashboard. In the following we provide a short summary of the tools aforementioned.

Cppcheck is a static analysis tool for C/C++ code.

Clang Static Analyzer is a source code analysis tool that finds bugs in C, C++ and Objective-C programs.

cloc counts blank lines, comment lines, and physical lines of source code in many programming languages.

SLOCCount is a set of tools for counting physical Source Lines of Code (SLOC) in a large number of languages of a potentially large set of programs.

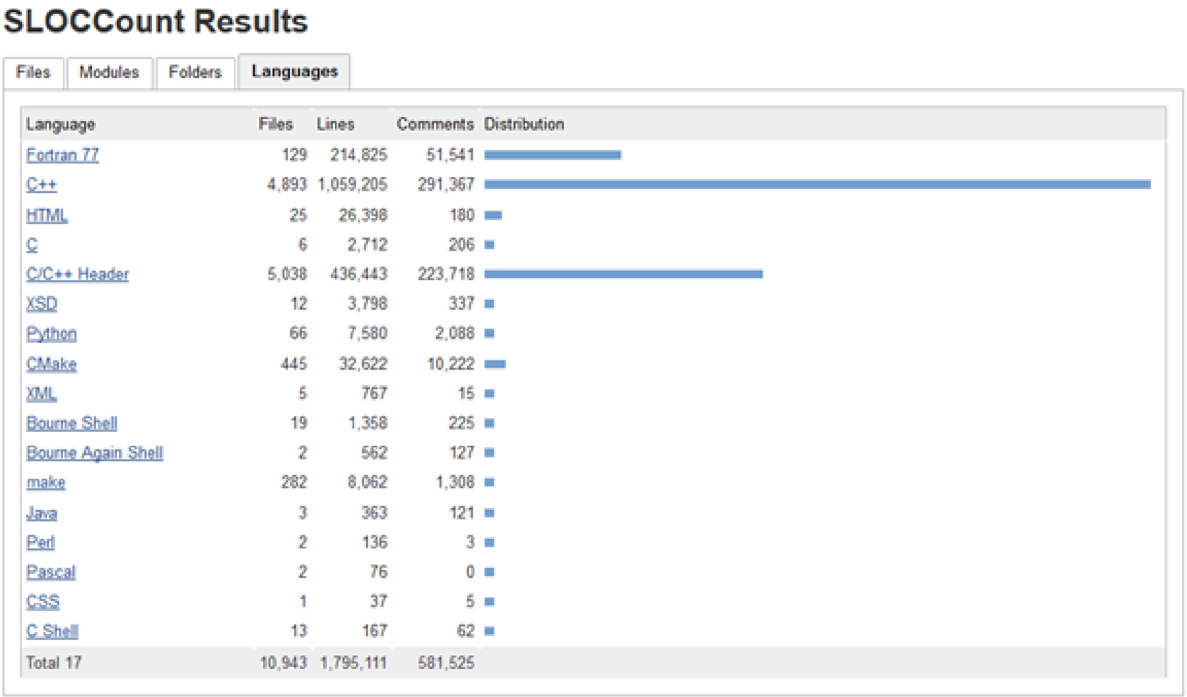
Referring to the code, we have used Geant4 10 major release to assess Geant4 quality over time as planned in the INFN CCR Uncertainty Quantification (UQ) project.

2. Proof of concept

To setup a continuous assessment of software characteristics we have:

- considered a subset of Geant4 releases, such as 10.0.4 to 10.1.2, (whose release naming follows the convention: major.minor.patch), neglecting the previous ones since our solution is easily applicable to the others;
- created a customized Docker image containing software metrics and static analysis tools;
- created a set of bash scripts to run software metrics and static analysis tools inside the Docker container;
- defined Jenkins jobs;
- installed and enabled in the Jenkins CI Cppcheck, Clang and SLOCCount plugins in order to publish analysis results on the job dashboard.

Figures 1 and 2 show some results for the 10.1.2 and 10.0.4 releases respectively.



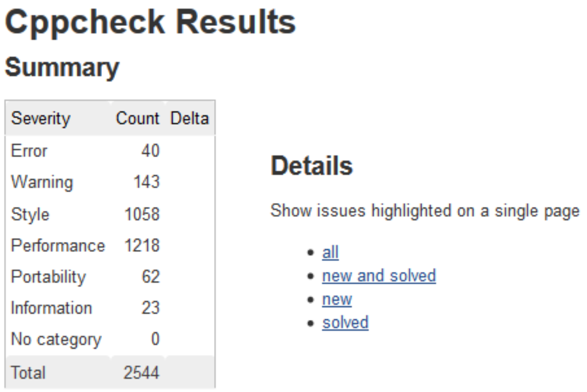
SLOCCount Results

Language	Files	Lines	Comments	Distribution
Fortran 77	129	214,825	51,541	
C++	4,893	1,059,205	291,367	
HTML	25	26,398	180	
C	6	2,712	206	
C/C++ Header	5,038	436,443	223,718	
XSD	12	3,798	337	
Python	66	7,580	2,088	
CMake	445	32,622	10,222	
XML	5	767	15	
Bourne Shell	19	1,358	225	
Bourne Again Shell	2	562	127	
make	282	8,062	1,308	
Java	3	363	121	
Perl	2	136	3	
Pascal	2	76	0	
CSS	1	37	5	
C Shell	13	167	62	
Total 17	10,943	1,795,111	581,525	

Figure 1. SLOCCount results for Geant4 10.0.4 release.

3. Future Work

The next steps of this activity consist of: including statistical analysis tools in the docker image, such as R [10]; integrating other software metrics and static analysis tools, such as Imagix 4D [11] and Parasoft C/C++/Java Test [12, 13], in the continuous integration; specifying software engineering references for the metrics measured whenever possible. In the following we provide a short summary of the tools aforementioned.



Cppcheck Results

Summary

Severity	Count	Delta
Error	40	
Warning	143	
Style	1058	
Performance	1218	
Portability	62	
Information	23	
No category	0	
Total	2544	

Details

Show issues highlighted on a single page

- all
- new and solved
- new
- solved

Figure 2. Cppcheck results for Geant4 10.1.2 release.

R is the leading programming language in statistics and data science.

Imagix 4D is a tool to reverse engineer and analyze your software.

Parasoft C/C++/Java Tests are integrated Development Testing solutions for automating a broad range of testing best practices proven to improve development team productivity and software quality.

References

[1] ISO/IEC 2011 Iso/iec 25010:2011 - systems and software engineering - systems and software quality requirements and evaluation (square) - system and software quality models Tech. rep. ISO URL *http : //www.iso.org/iso/home/standards/management − standards/iso*$_9$000.*htm*

[2] Lincke R, Lundberg J and Löwe W 2008 *Proceedings of the 2008 International Symposium on Software Testing and Analysis* ISSTA '08 (ACM) pp 131–142 URL `http://dx.doi.org/10.1145/1390630.1390648`

[3] Ronchieri E and Giacomini F 2015 An assessment of software metrics tools Tech. rep. INFN

[4] Ronchieri E, Pia M G and Giacomini F 2015 *Journal of Physics: Conference Series*

[5] Jenkins and Docker https://jenkins.io/solutions/docker/

[6] Cppcheck A tool for static C/C++ code analysis http://cppcheck.sourceforge.net/

[7] Clang Static analyzer http://clang-analyzer.llvm.org

[8] cloc https://github.com/AlDanial/cloc

[9] SLOCCount http://www.dwheeler.com/sloccount/

[10] R What is R? https://www.r-project.org/about.html

[11] Imagix https://www.imagix.com/support/docs.html

[12] Parasoft C/C++test https://www.parasoft.com/product/cpptest/

[13] Parasoft Jtest https://www.parasoft.com/product/jtest/

Software Maintenance and Release Management in INDIGO-DataCloud

C Aiftimiei[1,2], A Ceccanti[1], D Michelotto[1], E Ronchieri[1] and D Salomoni[1]

[1]INFN CNAF, Viale Berti Pichat 6/2, 40126 Bologna, Italy
[2]on leave from IFIN - ”Horia Hulubei”, Bucharest - Magurele, Romania

E-mail: cristina.aiftimiei@cnaf.infn.it

Abstract.
INDIGO-DataCloud (INtegrating Distributed data Infrastructures for Global ExplOitation, short INDIGO) [1], is a project started in April 2015, funded under the EC Horizon 2020 framework program. It includes 26 European partners located in 11 countries and addresses the challenge of developing open source software, deployable in the form of a data/computing platform, aimed to scientific communities and designed to be deployed on public or private Clouds and integrated with existing resources or e-infrastructures. In this paper we will present the activities done in order to support the software lifecycle by providing services to project software developers, user communities, and infrastructure providers. In particular we'll describe the initial plans to provide a continuous software improvement process that includes software quality assurance, software release management, maintenance, support services, and the pilot infrastructures needed for software integration and deployment testing. As part of software requirements verification, plans were defined to interface with projects' user communities in order to enable them to preview, test and evaluate the software under adequate conditions thus gathering early feedback and promoting exploitation.

1. Introduction

INDIGO-DataCloud project aims is to develop and deliver software to simplify the execution of applications on Cloud and Grid based infrastructures, as well as on HTC and HPC clusters.

It will target three software layers:

- The lower layer, developed by the WP4 (Resource Virtualization) workpackage, will deliver cloud IaaS, HTC and HPC enhancements enabling seamless access to a wide range of existing infrastructures using standard APIs.
- The middle layer, developed by WP5 (PaaS Platform Development) workpackage, will deliver PaaS capabilities that will leverage the WP4 functionalities to federateheterogeneous computing, storage and network resources providing elasticity and orchestration.
- The upper layer, provided by WP6 (Portal Workflows and User Interfaces), will make use of the PaaS layer, through data access and application deployment APIs provided by WP5, to provide user friendly frontends such as Scientific Gateways, desktop and mobile interfaces, to enable complex application workflows aimed at big data analytics.

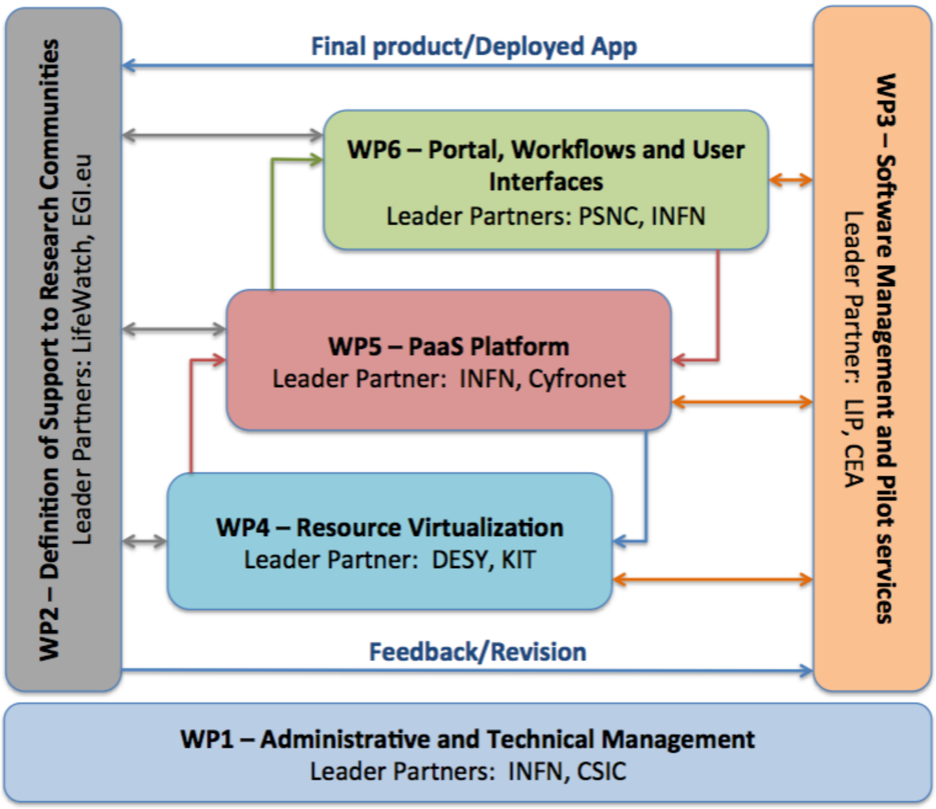

Final product/Deployed App
WP2 – Definition of Support to Research Communities
Leader Partners: LifeWatch, EGI.eu
WP6 – Portal, Workflows and User Interfaces
Leader Partners: PSNC, INFN
WP5 – PaaS Platform
Leader Partner: INFN, Cyfronet
WP4 – Resource Virtualization
Leader Partner: DESY, KIT
WP3 – Software Management and Pilot services
Leader Partner: LIP, CEA
Feedback/Revision
WP1 – Administrative and Technical Management
Leader Partners: INFN, CSIC

Figure 1. INDIGO - DataCloud WPs

This article details the Software Maintenance and Release procedures, which are part of the WP3 activities. The software improvement cycle is outlined in Figure 2. It starts with user and infrastructure provider requirements and technical plans which guide the software development and related test plans. From the WP3 point of view these plans, as well as any other software characteristic relevant for quality assurance and release preparation, are described in software specifications (blueprints) provided by the projects Product Teams for each product going through the process. Software is developed and continuously tested according to the blueprints and released according to a predefined schedule. Preview releases will be made available for evaluation by user communities and resource providers. Release candidates are subjected to integration testing, which may include the participation of user communities. Once the required release quality is attained the software will be made available to the general public through INDIGO-DataCloud repositories. WP3 will then provide support for the released software and manage bug fixes in close coordination with the developers. Finally, WP3 will contribute directly to software exploitation by liaising with resource providers and large infrastructures, ensuring that the project developments match their requirements and collecting new requirements and feedback that are reinserted in the process

During 2015 an initial plan of the support activities was defined in order to guide the management and deployment of the INDIGO-DataCloud software lifecycle processes in this complex and diversified environment. This plan is described in Figure 3 through a simplified view centred in the process flows across the initial set of tools and services that will be deployed and managed by WP3.

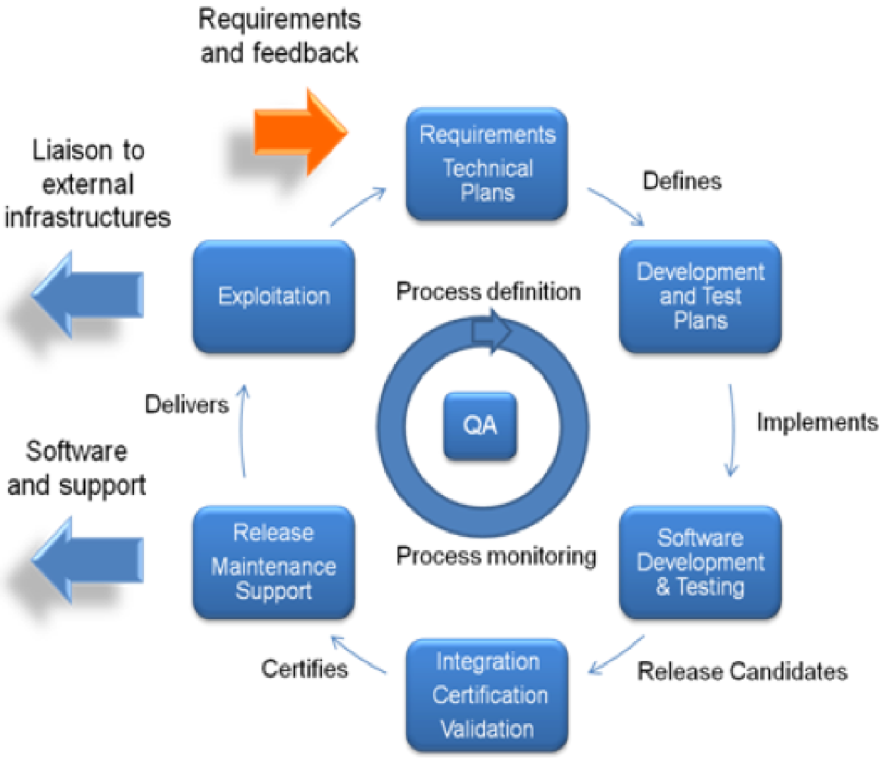

Requirements and feedback
Liaison to external infrastructures
Requirements Technical Plans
Defines
Exploitation
Process definition
Development and Test Plans
QA
Delivers
Implements
Software and support
Release Maintenance Support
Process monitoring
Software Development & Testing
Certifies
Integration Certification Validation
Release Candidates

Figure 2. Software Improvement Cycle

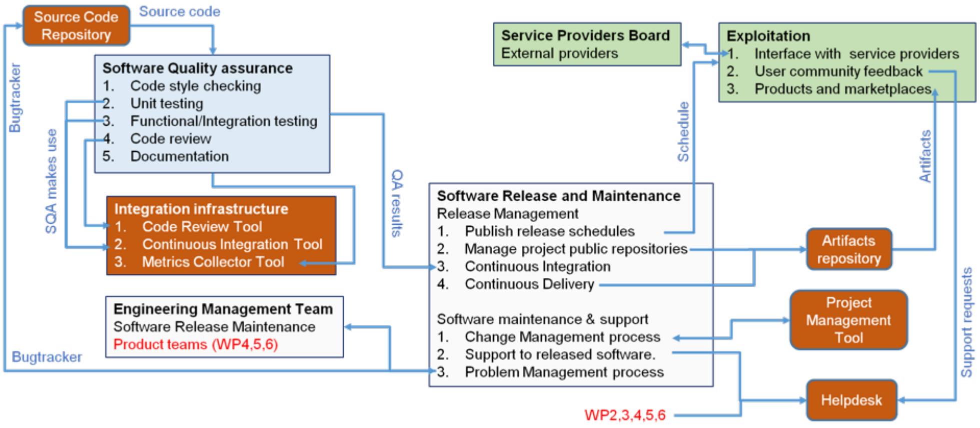

Source Code Repository
Source code
Bugtracker
SQA makes use
Software Quality assurance
1. Code style checking
2. Unit testing
3. Functional/Integration testing
4. Code review
5. Documentation
Integration infrastructure
1. Code Review Tool
2. Continuous Integration Tool
3. Metrics Collector Tool
Engineering Management Team
Software Release Maintenance
Product teams (WP4,5,6)
Bugtracker
QA results
Service Providers Board
External providers
Exploitation
1. Interface with service providers
2. User community feedback
3. Products and marketplaces
Schedule
Artifacts
Software Release and Maintenance
Release Management
1. Publish release schedules
2. Manage project public repositories
3. Continuous Integration
4. Continuous Delivery
Software maintenance & support
1. Change Management process
2. Support to released software.
3. Problem Management process
Artifacts repository
Project Management Tool
Support requests
Helpdesk
WP2,3,4,5,6

Figure 3. Overview of the process defined by the WP3 initial plan

2. Software Quality Assurance

INDIGO-DataCloud software components are expected to enable integration and interoperability between existing frameworks instead of providing a unique and monolithic solution. Code contributions to external entities as well as forking of existing products will be common scenarios. Hence new products developed from scratch will not be the only case to be addressed. Based on this heterogeneity in the project's definition, the SQA task proposed a well-defined spectrum of design principles and testing methodologies, flexible enough to cope with the projects inherent diversity. These SQA principles will pursue the ultimate goal of achieving continuous delivery (CI) of quality software products within INDIGO-DataCloud project. With the help of the WP3.1 (SQA) team INDIGO-DataCloud software developments MUST comply with the quality criteria principles described in the initial plan:

- Code Style - Code style enforcement promotes good practices that seek the maintainability of the code, emphasizing its readability through a set of conventions. More details can be found in [2]
  - Style guidelines MUST be defined by the Product Teams (PTs)
  - Code style compliance MUST be checked by automated means and for any given change in the code
  - Style guidelines MUST be maintained and SHOULD evolve over time
- Unit testing - Unit testing aims to early detect failures in functions and areas of the code of incoming changes that could potentially break any working functionality.
  - Unit tests MUST be provided to WP3 by the PTs
  - Unit testing MUST be checked by automated means and, for any given change in the code
  - Minimum acceptable code coverage SHOULD be 70% for the code developed under INDIGO-DataCloud project
  - Code coverage SHOULD increase over time
- Functional and integration testing - Functional testing involves the verification of the software components identified functionalities, based on requested requirements and agreed design specifications
  - Functional & integration testing definitions MUST be provided to WP3 by the PTs
  - Functional testing MUST be checked by automated means and for any given change in the code
  - Integration testing MAY be checked automatically
  - Regression testing SHOULD be covered at this stage by executing the complete set of functional tests for any given change
  - Changes that incorporate new functionalities MUST be well documented and tested by one or more associated black-box tests
  - Integration testing MAY be triggered for any given change in the code
  - All the non-relevant components that are needed for the functionality check/s to work MAY be simulated by using Mock objects
  - Integration testing outcome MUST guarantee the overall operation of the product whenever new functionalities are involved
- Code review - Software analysis is completed within the code review phase. This term applies to the informal non-automated (peer) human-based review of the candidate code, being already verified by the 3-level automated testing style, unit and functional.
  - Code review SHOULD be done by the PTs for any given change in the code
  - Code review SHOULD use the INDIGO-DataCloud projects agreed peer review tool
  - Code review MUST include an assessment of the inherent security risk of changes and a validation that the security model has not been downgraded by the changes
  - For those components meant to contribute to third-parties which already provide mechanisms and/or tools for code review, this task MAY be taken over externally from INDIGO-DataCloud project as long as it does not interfere with the projects deadlines
  - Changes that incorporate new functionalities MUST be well documented
- Documentation - PTs MUST provide documentation about the products being released under the INDIGO-DataCloud project scope

- Documentation MUST be treated like code, making it available through the projects VCS
- Documentation MUST be produced according to the target audience and MAY vary by component: Developer documentation, Deployment and Administration documentation, Command Line Interface (CLI) and Application Program Interface (API) documentation, whenever the component is exposing any of these interfaces
- Documentation produced MUST contain a reference to the standard license agreed by the project
- Documentation MUST be versioned however no numbering scheme is imposed
- The frequency of releasing new documentation is OPTIONAL whenever no significant changes are involved but it MUST be provided if releasing new functionalities or substantial changes that affect the products' usage

3. Software Release & Maintenance

The main objective of this activity is to make the certified software components available as a set of coherent high quality releases, supported by an efficient maintenance process. Figure 4 shows the position of the release and maintenance activities in the continuous process of handover of software from the developers, passing through the quality checks, testing on the integration and testing infrastructures, acceptance testing from the user communities, until the deployment on external production infrastructures.

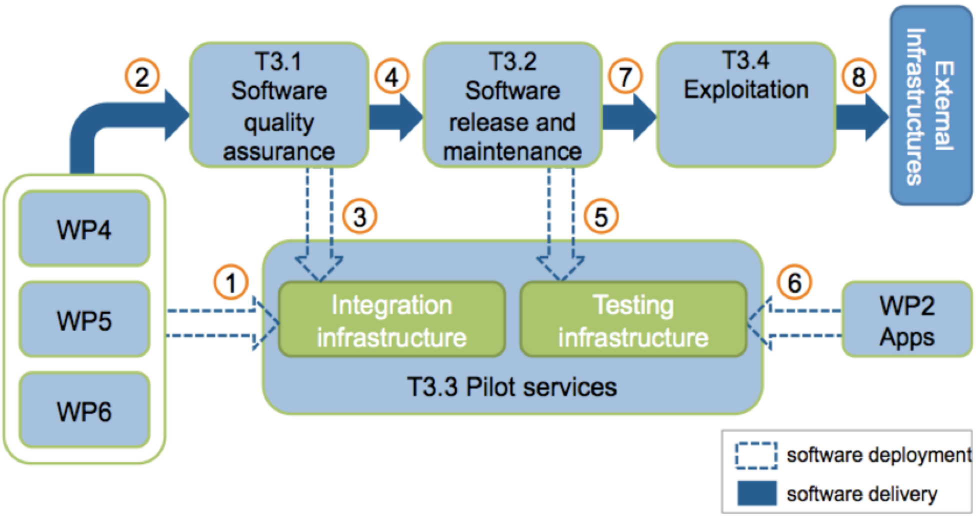


Figure 4. Software handover

The Software Maintenance task is responsible to coordinate the continuous maintenance of the software components developed within the project and included in a consistent distribution, preserving at the same time their stability in terms of interface and behaviour, so that higher-level frameworks and applications can rely on them. The INDIGO-DataCloud Maintenance organization follows the guidelines of the ISO/IEC 14764:2006 standard [R5], and includes a set of organizational roles, mapped onto the Work Packages and administrative roles, to handle maintenance implementation, change management and validation, s oftware release, migration and retirement, support and helpdesk activities. All of these aspects are described in the Deliverable D3.1[2]. The INDIGO-DataCloud distribution is organized in time-based major releases tentatively delivered in M14 and M24 of the project (Figure 5).

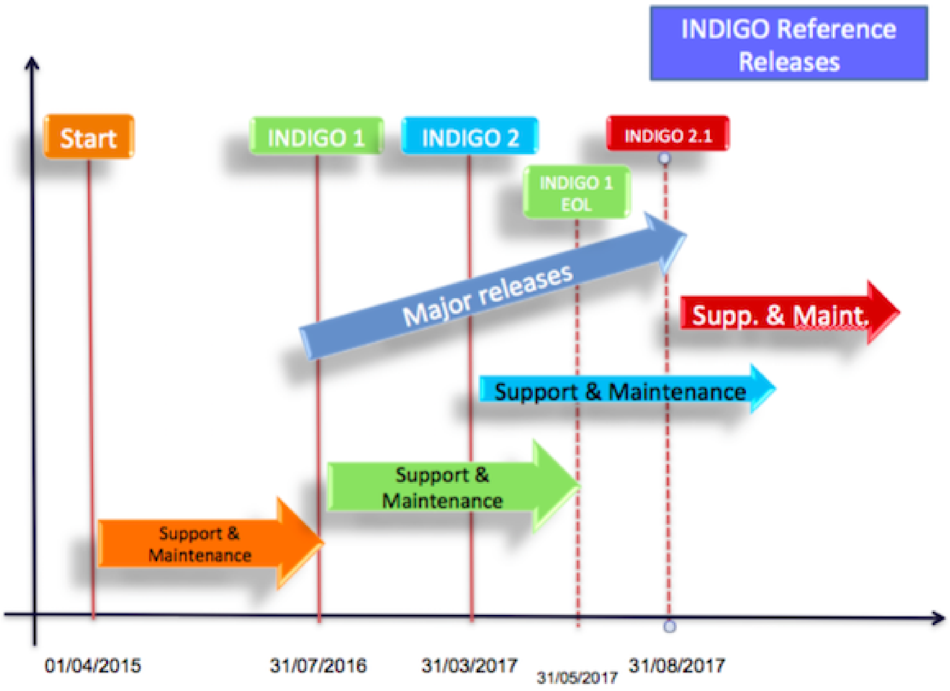


Figure 5. INDIGO-DataCloud Release Timeline

Taking into account the rapidly changing environment, especially some of the upstream release cycles (e.g. OpenStack), the INDIGO-DataCloud maintenance schedule is organised in the following periods:

- Full maintenance period: during this period updates are released to address issues in the code and provide new features for each supported INDIGO-DataCloud major release (6 months)
- Standard maintenance period: during this period updates are released to address issues in the code, but no new feature is introduced, for each supported INDIGO-DataCloud major release (2 months)
- Security updates period: during this period only updates targeting security vulnerabilities are provided for each supported INDIGO-DataCloud major release (2 months)
- End-of-life period: in this period no updates or support are provided. The end-of-life period starts after the end of the security updates period.

Component releases are classified in major, minor, revision and emergency, based on the impact of the changes on the component interface and behaviour. Requests for Change (RfC) are managed adopting a priority-driven approach, so that the risk of compromising the stability of the software deployed in a production environment is minimized. RfCs will also be properly monitored across the different trackers adopted by Product Teams. The Change management and release certification process, that leads from the submission of an RfC to a certified quality software release, is described in detail in the Deliverable D3.1[2].

The User Support activity deals with the coordination of the support, together with EGI, to users of the software components developed within the project and included in an INDIGO-DataCloud distribution. User Support is organized in three levels, of which the second and the third are within the INDIGO-DataCloud project and provide the most specialized knowledge needed to investigate a reported incident. Many Support Units, corresponding approximately to the products delivered by INDIGO-DataCloud, will be established and registered on the reference support portal (GGUS). The support operations are described in detail in the Deliverable D3.1[2].

The User Support activity in WP3.2 depicted in Figure 6 is responsible for coordinating the support, together with EGI, of the users of the software components developed within the project and included in an INDIGO-DataCloud distribution. Service Level Agreements will regulate support activities with customers.

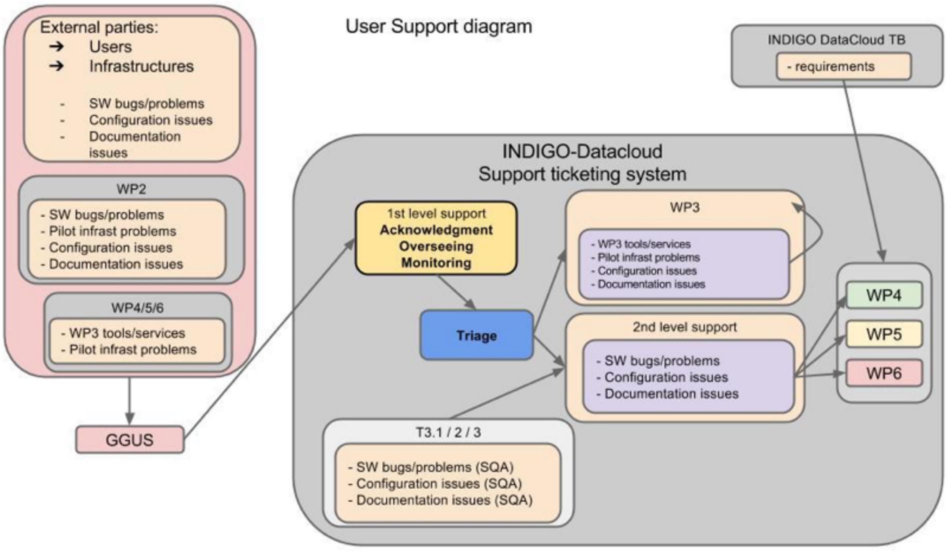
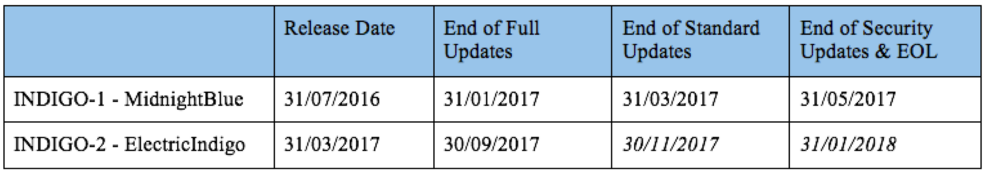


Figure 6. INDIGO-DataCloud Support Diagram

During 2015 were defined all the processes regarding maintenance, release, support, the workflow that will be follwed in order to ensure a high quality software will be delivered to the project user communities. The forseen release schedule is presented in Table 1, bellow.



	Release Date	End of Full Updates	End of Standard Updates	End of Security Updates & EOL
INDIGO-1 - MidnightBlue	31/07/2016	31/01/2017	31/03/2017	31/05/2017
INDIGO-2 - ElectricIndigo	31/03/2017	30/09/2017	*30/11/2017*	*31/01/2018*

Figure 7. INDIGO-DataCloud Major Releases Schedule

Regarding the Scheduled Release Procedure, major, minor and revision releases will follow a scheduled release procedure, consisting of:

- Release Planning:
  - identify requests for new features to be developed and bug fixes to be implemented;
  - prioritize, plan and schedule the development and maintenance activities;
  - document and track release planning activities in the Development and Test Plans and by creating Request for Changes (RfCs) in the PTs trackers with approved priorities;

  - define the Release Schedule by creating the Components Releases items in the corresponding tracker;
- Release Building
  - develop new features, implement required bug fixes;
  - test and certify developed components;
  - guidelines are provided by WP3 regarding:
    - Build Integration and Configuration, defining how to configure the selected build/integration tool to properly build each software component;
    - Packaging INDIGO-DataCloud software components will follow the Fedora and Debian packaging guidelines where applicable;
    - Certification and Testing, containing details on what kind of tests to perform and when; how to perform certification on a Component Release; how to write a software verification and validation report templates to be used in the definition of automatic tests jobs in the CI system;
- Certification and Validation (acceptance testing)
  - Component Releases are validated against the set of acceptance criteria defined by the Customers and the INDIGO-DataCloud Production Release criteria [R13];
  - components are made available for technical-previews;
  - components are deployed on the INDIGO-DataCloud integration-testbed for a (6 days) grace period under an automatic monitoring tool;
- Release Preparation and Deployment
  - final packaging and signing of components;
  - components uploaded in the official INDIGO-DataCloud Software/Container/Image Repositories;
  - prepare and publish release documentation: Release Notes, Known Issues;
  - announce the Release to user communities/production infrastructures/customers

The public availability of a new INDIGO-DataCloud release will be done through:

- general announcements:
- INDIGO-DataCloud RSS Feed;
- indigo-announce mailing-list [R14];
- general user communities mailing-lists.
- specific announcements: - follow EGI Staged-Rollout procedures [R15]: new releases are announced to the EGI UMD Release team mailing-list [R16]

4. Pilot Services

The task WP3.3 will provide two pilot infrastructures and the services needed for the integration and testing of software components provided by the PTs in WP4, 5 and 6. The pilot infrastructures will use resources operated by the project partners.

4.1. Services for Continuous Integration and Software Release

A set of tools and services are needed to support the PTs, the Software Quality Assurance, the Continuous Integration and the software release and maintenance.

The overall architecture is shown in Figure 7. It has two main blocks, on the left hand side (light coloured block) are public cloud services, which are not deployed or operated by INDIGO-DataCloud WP3, and on the right hand side (dark coloured block) are services that are deployed and operated by WP3.

The list of services needed are given below with a small description for each service or tool and depicted in Figure 8:

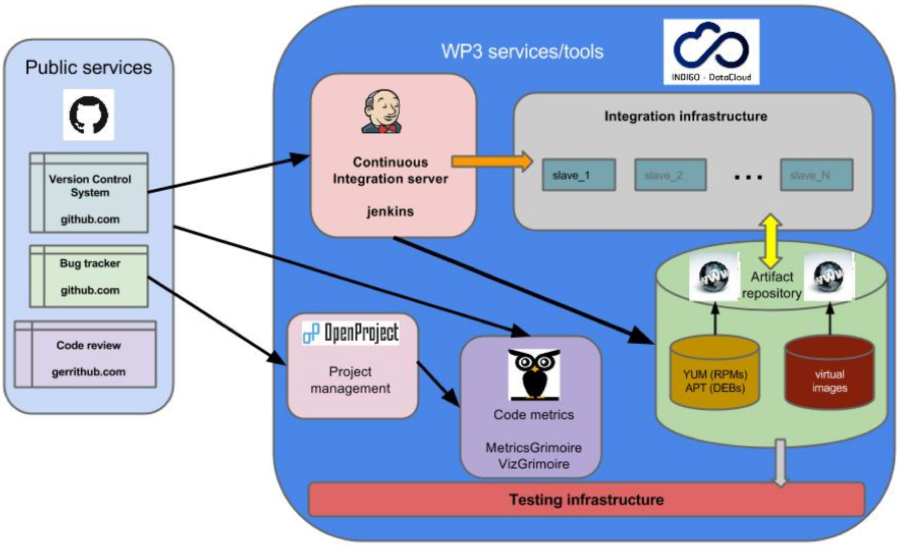


Figure 8. Overall architecture of the services provided by WP3.3 for the SQA and Continuous Integration process

- Project management service: https://www.openproject.org/ - It provides tools such as an issue tracker, wiki, a placeholder for documents and a project management timeline. This service is deployed and operated by WP3.
- Version Control System: https://github.com/indigo-dc - Service for configuration management and revision control for software source code. This service is based on a public cloud service.
- Continuous Integration service: https://jenkins.indigo-datacloud.eu:8080/ - Service to automate the building, packaging (where applicable) and execution of unit and functional tests of software components. Automatic source code style checking is also foreseen. This SHOULD cover most of the SQA automation (sections 3.3.1, 3.3.2 and 3.3.3). The service MUST include a dashboard for the visualization of the test results. This service is deployed and operated by WP3.
- Artifacts repository: http://repo.indigo-datacloud.eu/ - In INDIGO-DataCloud there are two types of artifacts, packaged software and virtual images. Both types of artifacts will be stored in a storage resource and accessed through a frontend service that MAY be accomplished via web servers. The repository will be used both for INDIGO-DataCloud purposes as well as to expose releases to the public. This service is deployed and operated by WP3. For the released software is RECOMMENDED, when possible, to be also registered or published in public upstream repositories, like OpenStack and Docker Hub [R18].
- Code review service: http://gerrithub.io/ - Code review of software source code is one integral part of the SQA process [section 3.3.4]. This service facilitates the code review process. It records the comments and allows the reviewer to approve the software modification. This service is based on a public cloud service.
- Code metrics services:
  - MetricsGrimoire: https://metricsgrimoire.github.io
  - VizGrimoire: http://vizgrimoire.bitergia.org
  - To collect and visualize several metrics about the software components. Some of these metrics are part of the overall WP3 metrics. These services are deployed and operated

by WP3.

- Bug trackers:
  - GitHub: https://github.com/
  - OpenProject: https://www.openproject.org/
  - Services to track issues and bugs of INDIGO-DataCloud software components as well as services provided by WP3.
- Integration infrastructure: this infrastructure is composed of computing resources to support directly the Continuous Integration service. Its the place where building and packaging of software occurs as well as the execution of unit and functional tests. These resources are provided by WP3 partners
- Testing infrastructure: this infrastructure aims to provide several types of environment. A stable environment for users where they can preview the software and services developed by INDIGO-DataCloud, prior to its public release. There will be also environments for the PTs that need external software or services for their work. Examples are complete OpenStack and OpenNebula deployments. These environments will be deployed by the PTs with support from WP3 on infrastructures made available by WP3-partners.

The INFN CNAF team was involved in the deployment and management of the Continuous Integration system and is actively contributing to both integrationg and testing infrastructures.

5. Exploitation

The INDIGO-DataCloud exploitation activities aim to maximize the options available to the users to access the INDIGO-DataCloud software products. INDIGO-DataCloud is producing software, integrated solutions, appliances and services, which will eventually be deployed in external production infrastructures used by the user communities. It is critical for the success of the project, that the outputs of INDIGO-DataCloud are available in the infrastructures that are relevant to the users. Therefore the activities within this task aim to maximize (from a technical point of view) the uptake of the outputs of the project by the e-infrastructures. In other words, to remove constraints that would prevent the users from accessing the INDIGO-DataCloud software at the infrastructure level.

5.1. Service Providers Board

The Service Providers Board (SPB) is one of the advisory bodies of INDIGO-DataCloud, the purpose of the board is to have a communication channel with the service providers, who provide the resources to the user communities, to make sure that their technical requirements are fulfilled and that the software and platforms developed in INDIGO-DataCloud can run on their resources. The main output of the work of this board will be a set of requirements, both policy and technical, that will then be fed via the same channels used by WP2 to communicate to the JRA work packages the user communities requirements.

INFN-CNAF is involved in the SPB activities with teh Release Manager, that is presenting the schedules of new releases and follows up eventual support requests.

5.2. Staged Rollout

Another important activity of T3.4 is to coordinate the Staged Rollout (SR) activities of the projects outputs.

WP3 other tasks, together with the JRAs work packages, already foresee in their work plan the testing of the software produced by INDIGO-DataCloud in the WP3 integration and testing infrastructures using unit testing and other testing methods implemented by the developers.

Once products reach production their features are used in real use case scenarios by real users. Some bugs and issues in a new release of software not identified during the certification of the products done by the developers only appear when used in production.

Staged Rollout is the best practice implemented in many production environments to mitigate this issue. It consists in the deployment in production of a new release of software in a subset of selected service providers called Early Adopters (EAs) at first, and give the green light for a wider deployment only when after a period of testing with the users the software is considered stable. This best practice is widely used in commercial environments where new releases are exposed only to a percentage of the users at first, percentage that is then increased over time to reach 100

This activity is mostly a coordination action, which triggers the deployment of the new releases in the selected set of EAs when new updates are ready for production.

6. Future Work

All the activities described will be followed during the next period in order ensure the provisionig of a production quality software. The SQA criterias will be completely defined and implemented in an automatic way. Various testbeds for the integration, dedicated to developers, and the preview one for the user communities to test heir use cases. Repositories of the artefacts, testing, preview and production will be populated during each phase of the relese process. Also a signing procedure will be implemented in order to guarantee the provenance both of the packages (rpms and debs) and containers images, leveraging the The Update Framework (TUF)[3]

References

[1] INDIGO-DataCloud, *https://www.indigo-datacloud.eu/*

[2] Initial Plan for WP3 – D3.1 *https://www.indigo-datacloud.eu/documents/initial-plan-wp3-d31*

[3] The Update Framework (TUF), *https://theupdateframework.github.io/*

INDIGO-DataCloud – Filling the gaps

C Aiftimiei[1,2], **M Caberletti**[1], **A Ceccanti**[1], **S Dal Pra**[1], **E Fattibene**[1], **D Michelotto**[1], **E Ronchieri**[1], **D Salomoni**[1], **V Sapunenko**[1], **S Taneja**[1], **S Zani**[1]

[1]INFN CNAF, Bologna, Italy
[2]on leave from IFIN - "Horia Hulubei", Bucharest - Magurele, Romania

E-mail: `cristina.aiftimiei@cnaf.infn.it`

Abstract.
In Cloud computing, both the public and private sectors are already offering Cloud resources as IaaS (Infrastructure as a Service). However, there are numerous areas of interest to scientific communities where Cloud Computing uptake is currently lacking, especially at the PaaS (Platform as a Service) and SaaS (Software as a Service) levels. In this context, INDIGO-DataCloud (INtegrating Distributed data Infrastructures for Global ExplOitation), a project funded under the Horizon 2020 framework program of the European Union, aims at developing a data & computing platform targeted at scientific communities, deployable on multiple hardware, and provisioned over hybrid e-Infrastructures. In this paper we'll provide a high-level description of the project golas and architecture highlighting the contributions of the various INFN-CNAF teams to the development of many of the solutions of the project

1. Introduction

The **INDIGO-DataCloud** (INDIGO for short, see https://www.indigo-datacloud.eu) approved in January 2015 in the EINFRA-1-2014 call, with a budget of 11.1 M, for a duration of 30 months (from April 2015 to September 2017) sees the collaboration of 26 European partners in 11 European countries, under the coordination of the Italian National Institute for Nuclear Physics (INFN), including developers of distributed software, industrial partners, research institutes, universities, e-infrastructures. It aims to answer to the technological needs of scientists from multi-disciplinary scientific communities, like structural biology, earth science, physics, bioinformatics, cultural heritage,astrophysics, life science, climatology, seeking to easily exploit distributed Cloud/Grid compute and data resources.solutions will be deployable on hybrid (public or private) Cloud infrastructures. Following the raccomandations of the EC Expert Group Report on Cloud Computing presented in their *Advenced in the Cloud*[1] report, there are two main activities that INDIGO explicitly targeted in the EINFRA-1-2014 proposal:

- Large scale **virtualisation of data/compute centre resources** to achieve on- demand compute capacities, improve flexibility for data analysis and avoid unnecessary costly large data transfers.
- **Development and adoption of a standards-based computing platform** (with open software stack) that can be deployed on different hardware and e-infrastructures (such as clouds providing infrastructure-as-a-service (IaaS), HPC, grid infrastructures...) to abstract application development and execution from available (possibly remote) computing systems.

In this environment the INDIGO - DataCloud foundations are:

- **Put users first**: involve from the onset researchers, big resource centers, industry, software developers.
- **Develop open source software** in order to **fill technological gaps** that prevent the exploitation of current European e-infrastructures by many scientific communities, like:
  - Open **interoperation** / federation across (proprietary) CLOUD solutions at IaaS, PaaS, and SaaS levels
  - Managing **multitenancy** at large scale and in heterogeneous environments
  - Dynamic and seamless **elasticity** for both private and public cloud and for complex or infrequent requirements
  - **Data management** in a Cloud environment due to technical as well as to legal problems
- **Exploit 15 years of experience in software development of production-quality distributed infrastructures for science** matured by the project participants.
- Define and validate software components to be developed through **concrete scientific use cases**.
- **Reuse and extend existing components** wherever possible, **develop missing pieces** whenever necessary.
- Be as **multidisciplinary** and as neutral as possible through the adoption of both **de jure and de facto standards** to achieve **interoperability**.

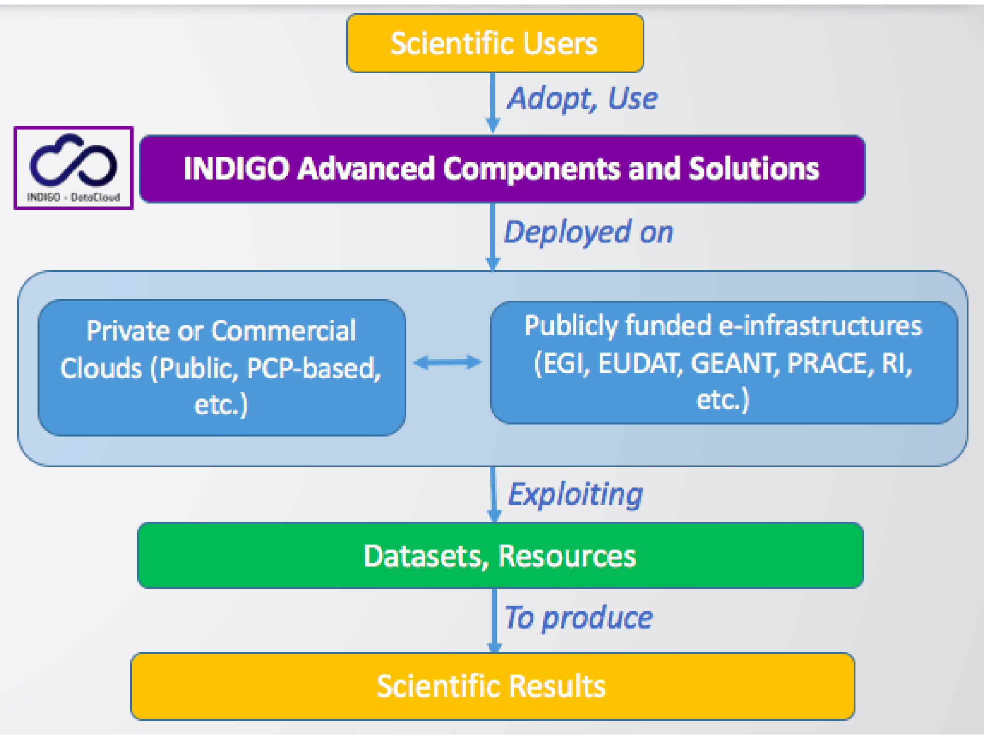

Scientific Users
Adopt, Use
INDIGO - DataCloud
INDIGO Advanced Components and Solutions
Deployed on
Private or Commercial Clouds (Public, PCP-based, etc.)
Publicly funded e-infrastructures (EGI, EUDAT, GEANT, PRACE, RI, etc.)
Exploiting
Datasets, Resources
To produce
Scientific Results

Figure 1. INDIGO - DataCloud Positioning

INDIGO - DataCloud project aims to:

(i) Develop open, interoperable solutions for scientific data.

(ii) Support open science organizing the European data space.

(iii) Enable collaborations across diverse scientific communities worldwide. The project will offers its architecture, analysis, expertise and software components as a concrete step toward the definition and implementation of a European Open Science Cloud and Data Infrastructure.

2. Overall Project Structure

The project is structured into six work packages, covering Network Activities (NA), Service Activities (SA) and Joint Research Activities (JRA), that are shortly described below. The diagram showing the interrelation between work packages is shown in Figure 2.

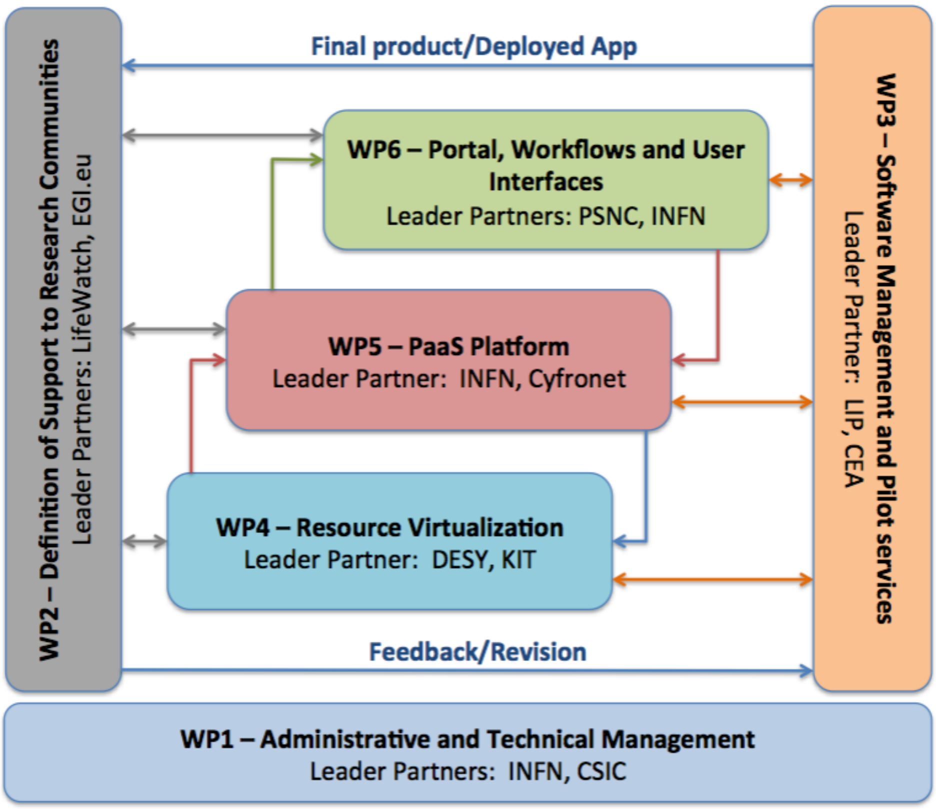

WP2 – Definition of Support to Research Communities
Leader Partners: LifeWatch, EGI.eu
Final product/Deployed App
WP6 – Portal, Workflows and User Interfaces
Leader Partners: PSNC, INFN
WP5 – PaaS Platform
Leader Partner: INFN, Cyfronet
WP4 – Resource Virtualization
Leader Partner: DESY, KIT
Feedback/Revision
WP3 – Software Management and Pilot services
Leader Partner: LIP, CEA
WP1 – Administrative and Technical Management
Leader Partners: INFN, CSIC

Figure 2. Diagram showing the interrelation among the Work Packages

Work package 1 (WP1, NA) - Project Management is dedicated to the overall project administration including definition of the project Quality Assurance plan. The objectives include the overall efficient operation of the consortium, careful monitoring of resource and financial expenditures, fulfillment of contractual obligations, periodic reporting and relationship with the European Commission.

Work package 2 (WP2, NA) - Definition of Support to Research Communities represents the interest of Research Communities to assure that their requirements will be satisfied by the project outcomes, by providing feedback and participating in the revision of the services deployed. WP2 will keep the focus also on big data research use and management through a dedicated task oriented to track the different needs at the data life-cycle, following the reference models used by the different Research Communities. The proposed Dissemination and

Communication activities include both strengthening Research Community Forums and relations with e-infrastructure stakeholders and policy makers. A task in WP2 will then be devoted to sustainability, where the analysis of the relationships between the different stakeholders in an open framework, like the one proposed in INDIGO, will be done. Cooperation mechanisms between the participants and also with external users and providers, will be analyzed. Research Communities covering a wide and significant spectrum of areas and expertise are represented through the participation of relevant institutions associated to ESFRIs or EIROs:

- Biological and Medical Sciences: EuroBioImaging BBMRI (UPV), ELIXIR (CNR), INSTRUCT (U.Utrecht, CIRMMP)
- Social Sciences and Humanities: DARIAH (RBI), DCH-RP (ICCU)
- Environmental and Earth Sciences: LifeWatch (CSIC), EMSO (INGV), ENES (CMCC)
- Physical Sciences: LBT, CTA (INAF) [+conduit to WLCG+HEP from CERN]

Work package 3 (WP3, SA) - Software Management and Pilot Services will provide software lifecycle services and related support to project developers, user communities, and infrastructure providers. In particular WP3 will provide a continuous software improvement process that includes software quality assurance, software release management, maintenance, support services, and the pilot infrastructures needed for software integration and testing. The software requirements and lifecycle processes will be designed to ensure compatibility with a wide range of infrastructures and user communities, through a requirements gathering process conducted in collaboration with partner infrastructures, individual resource providers, and complemented by feedback collected from selected applications through which the final software assessment will be performed. Security will be considered at all stages of design, implementation and deployment.

The core of the technical activities, and in particular the key development/adaptation of software packages, will be handled through the three JRA (Joint Research Activities) work packages WP4, WP5 and WP6.

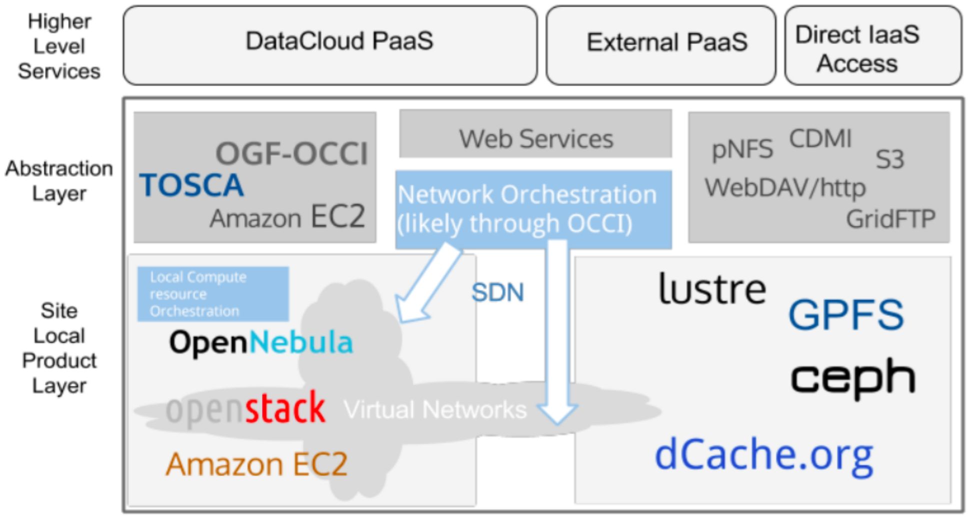


Figure 3. Technical layout of WP4

Work package 4 (WP4, JRA) - Resource Virtualization - activities are the closest to the e-Infrastructure resources, will address the required **improvements in the Virtualization**

of Resources. For what regards virtualization of computing resources, solutions for elastic resource allocation will be developed. Important work on containers will be carried out as well, which will assure a more efficient use of the hardware, and in particular a better exploitation of high performance components such as accelerators (GPU) or HPC network interfaces (IB). The adoption of both elastic resource allocation and containers will also have a positive impact in the efficiency to deploy applications and large software images, such as those required in some analytic packages to process large datasets. The second important change regarding virtualization proposed in WP4 is the adoption of the Storage Interface Concept, providing a single umbrella to the integration of the different data access mechanisms, including both block and object/file models. The third main area in WP4 is related to network virtualization, where solutions for topology on-demand, for enhancing the capabilities of local virtual networks and for advanced network services, aimed at obtaining greater control over network traffic, improve response time and optimize content delivery. The technical layout of WP4 is shown in Figure 3.

Work package 5 (WP5, JRA) - PaaS Platform Development will exploit these improvements to **extend an open PaaS framework** to support the requirements of the Research Communities starting with the detailed analysis of Use Cases. One of the challenges to improve current services is the provision of an authorization layer well integrated with existing federated identity mechanisms and resource providers, offering dynamic fine-grained access control to data and services shared at various levels of the infrastructure. As a result, a set of APIs to manage the feature-driven execution of applications and services will be made available, supporting QoS, dependency among tasks and advance reservation. The PaaS framework will assure that popular user application/software suites such as Octave/MATLAB, R-Studio or ROOT can be executed transparently in e-Infrastructures, in particular integrating already existing features such as parallelism (MATLAB, ROOT) or parameter sweeping. By providing SLA templates for application execution, including transparent default versions prepared by RI managers, the INDIGO PaaS framework will also optimize the selection of resources/sites for execution. This will take into account not only scheduling but also computing and data storage/access performance/transfer needs, as well as the associated costs. In fact, integration of three different options for data access in the resource provisioning scheme (compute where data is, transfer data to computing resources, and compute on any data, anytime, anywhere through exploitation of high capacity networks) is a key point of the INDIGO PaaS platform.

Work package 6 (WP6, JRA) - Science Gateways, Workflows and Toolkits will address the complex challenge of **guaranteeing a simple and effective final user experience**, both for software developers and for researchers running the applications. This objective requires different activities, starting with the development of APIs to access the PaaS framework, so that its features can be used by Portals, Desktop Applications and also Mobile Apps. WP6 will also provide (i) the support for distributed (coarse grained) data driven workflows for e-Science on Grid, Cloud and HPC resources and (ii) a specific support addressing tightly coupled (fine grained) workflows orchestration on big data analytics frameworks.

3. Architectural Overview

The challenge is to design an architecture, which contains all the elements needed to provide users with the capability of using heterogeneous infrastructures in a seamless way. The current technology based on lightweight containers and related virtualization developments make it possible to design such Platforms as a Service in a relatively straightforward way. There are already many examples in the industrial sector, in which open source PaaS solutions (eg. OpenShift or Cloudfoundry) are being deployed to support the work of companies in different sectors. However, the case of supporting scientific users is more complex because of the heterogeneous nature of the infrastructures at the IaaS level (i.e. the resource centers) and of the inherent complexity of the scientific work requirements. The the INDIGO - DataCloud

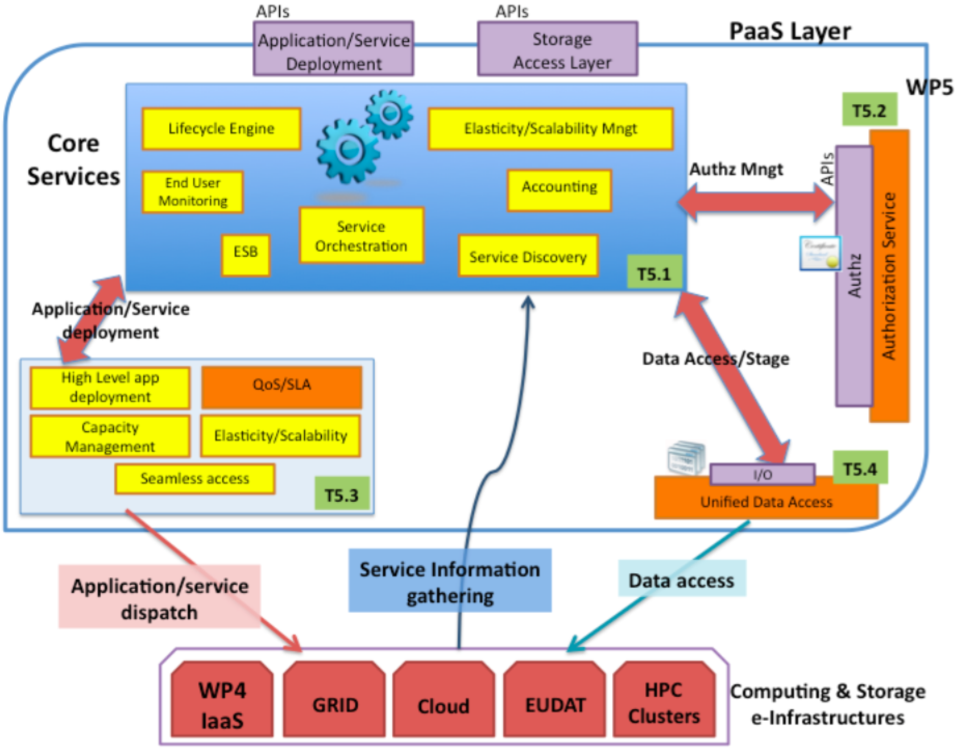


Figure 4. WP5 Architecture

General Architecture is presented in Figure 5 bellow.

3.1. Filling the Gaps

At the **resource provider level**, computing centers offer resources in an Infrastructure as a Service (IaaS) mode. The main challenges to be addressed at this level are:

(i) Improved scheduling for allocation of resources by popular open source Cloud platforms, i.e. OpenStack[2] and OpenNebula[3]. In particular, both better scheduling algorithms and support for spot-instances are currently much needed.

(ii) Improved Quality of Service (QoS) capabilities of storage resources. The challenge here is to develop a better support for quality of services in storage, to enable high-level storage management systems (such as FTS) and make it aware of information about the underlying storage qualities. The impact of such QoS when applied to storage interfaces such as dCache will be obvious for LHC Data analysis and for the long-term support, preservation and access of experiment data.

(iii) Improved capabilities for networking support. This is particularly the case when it comes to deploy tailored network configurations in the framework of OpenNebula and OpenStack.

(iv) Improved and transparent support for Docker containers. Containers provide an easy and efficient way to encapsulate and transport applications. The benefits of using containers, in terms of easiness in the deployment of specialized software, including contextualization features, eg. for phenomenology applications, are clear. They offer obvious advantages in terms of performance when compared with virtual machines, while also opening the door to exploit specialized hardware such as GPGPUs and low-latency interconnection interfaces (InfiniBand).

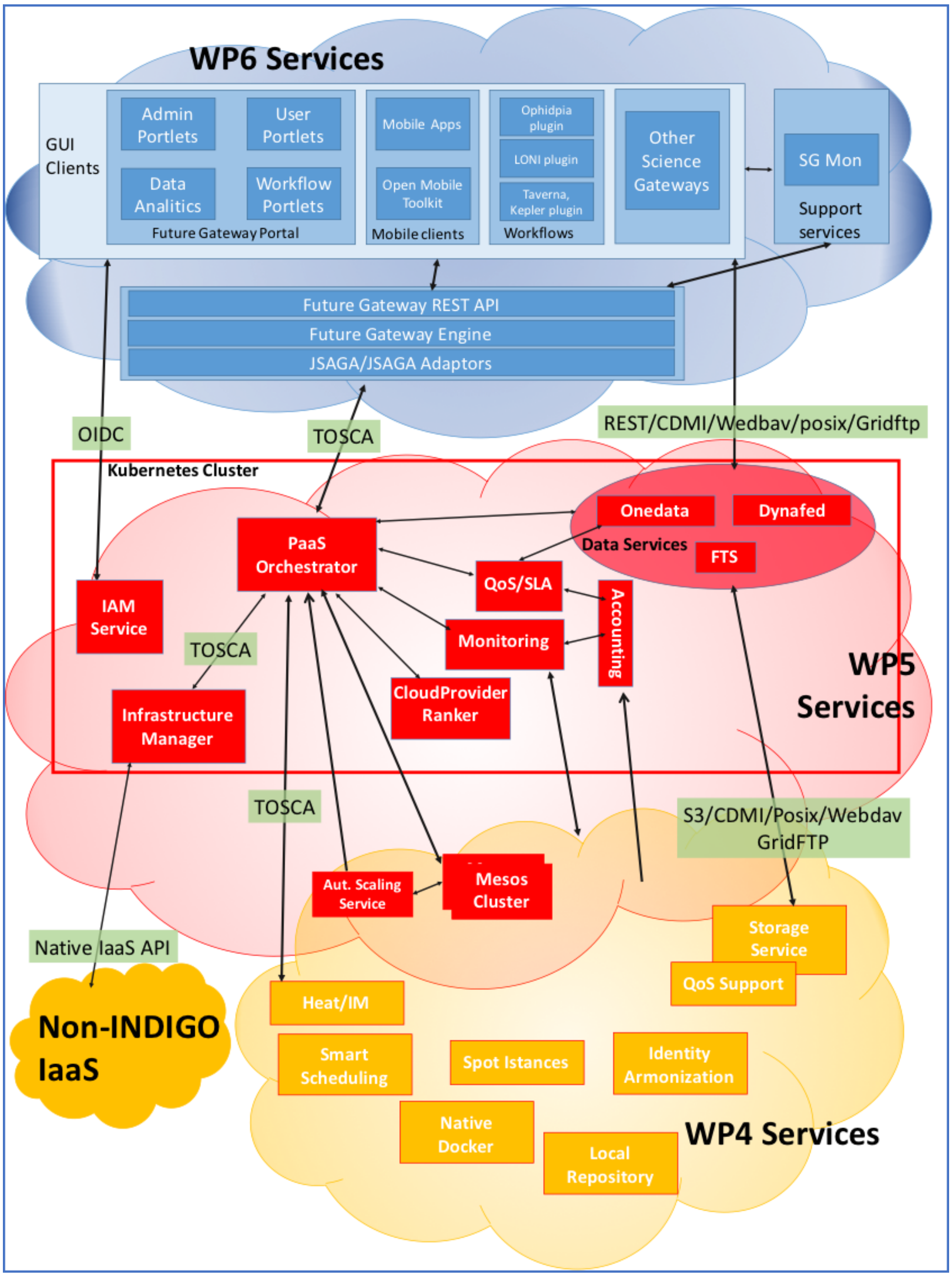


Figure 5. INDIGO - DataCloud General Architecture

The PaaS layer should provide advanced tools for computing and for processing large amounts of data, and to exploit current storage and preservation technologies, with the appropriate mechanisms to ensure security and privacy. The following points describe the most important missing capabilities which today require further developments:

(i) Improved capabilities in the geographical exploitation of Cloud resources.

(ii) Support for data requirements in Cloud resource allocations. Resources can be allocated where data is stored, therefore facilitating interactive processing of data.

(iii) Support for application requirements in Cloud resource allocations. For example, a given user can request to deploy an application on a cluster with Infiniband interfaces, or with access to specialized hardware such as GPGPUs.

(iv) Transparent client-side import/export of distributed Cloud data.

(v) Deployment, monitoring and automatic scalability of existing applications, including batch systems on-demand.

(vi) Integrated support for high-performance Big Data analytics and workfow engines such as Taverna[4] or Ophidia[5].

(vii) Support for dynamic and elastic clusters of computational resources

At the SaaS layer we find the user interface that it should provide ready-to-use tools for such capabilities to be exploited, with the smoothest possible learning curve. Providing such an interface between the user and the infrastructure poses two fundamental challenges:

(i) Enabling infrastructure services to accept state of the art user authentication mechanisms (e.g. OpenID connect, SAML) on top of the already existing X.509 technology. For example, distributed authorization policies are very much needed in scientific cloud computing environments, therefore the Authentication and Authorization Infrastructure (AAI) is a key ingredient to be fed into the architecture.

(ii) Making available the appropriate libraries, servlets and portlets, implementing the different functionalities of the platform (AAI, data access, job processing, etc.) that are the basis to integrate such services with known user tools, portals and mobile applications.

3.2. INDIGO - DataCloud Approach

Supporting scientific users is very complex because of the heterogeneous nature of the infrastructures at the IaaS level (i.e. the resource centers) and of the inherent complexity of the scientific work requirements. The key point is to find the right agreement to unify interfaces between the PaaS and IaaS levels. Providing the means to properly interface with the resource centers is the mission of WP4. This Work Package focuses on virtualizing local compute, storage and networking resources (IaaS) and on providing those resources in a standardized, reliable and performing way to remote customers or to higher level federated services, building virtualized site independent platforms. The IaaS resources are provided by large resource centers, typically engaged in well-established European e-infrastructures. The e-infrastructure management bodies, or the resource centers themselves will select the components they operate, and INDIGO will have limited influence on that process. Therefore, WP4 only concentrates on a selection of the most prominent components and develops the appropriate interfaces to high-level services based on standards.

The PaaS core components, developed by WP5, will be deployed as a suite of small services using the concept of micro-service[6]. This term refers to a software architecture style, in which complex applications are composed of small independent processes communicating with each other via lightweight mechanisms like HTTP resource APIs. The modularity of micro-services makes the approach highly desirable for architectural design of complex systems, where many developers are involved. Kubernetes[7], an open source platform to orchestrate and manage Docker[8] containers, will be used to coordinate the micro-services in the PaaS. Kubernetes is extremely useful for the monitoring and scaling of the services, and will ensure the reliability of all of them. In Figure 6 there are represented the key components of the PaaS, briefly described bellow, and their high-level interrelations:

- the Orchestrator: this is the core component of the PaaS layer. It receives high-level deployment requests from the user interface software layer, and coordinates the deployment process over the IaaS platforms;

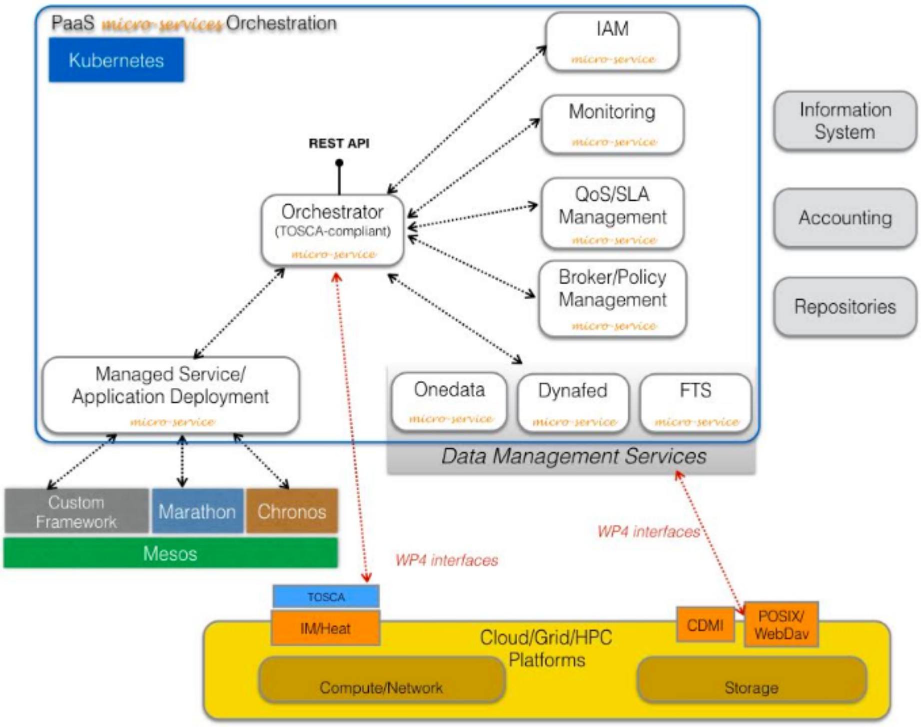


Figure 6. Key components of the PaaS and their high-level interrelations

- the Identity and Access Management (IAM) Service: it provides a layer where identities, enrolment, group membership, attributes and policies to access distributed resources and services can be managed in a homoge- neous and interoperable way;
- the Monitoring Service: this component is in charge of collecting monitor- ing data from the targeted clouds, analysing and transforming them into information to be consumed by the Orchestrator;
- the Brokering/Policy Service: this is a rule-based engine that allows to manage the ranking among the resources that are available to fulfil the requested services. The Orchestrator will provide the list of IaaS instances and their properties to the Rule Engine. The Rule Engine will then be able to use these properties in order to choose the best site that could support the users' requirements. The Rule Engine can be configured with different rules in order to customize the ranking;
- the QoS/SLA Management Service: it allows the handshake between a user and a site on a given SLA; moreover, it describes the QoS that a specific user/group has, both over a given site or generally in the PaaS as a whole. This includes a priority for a given user, i.e. the capability to access different levels of QoS at each site (e.g., Gold, Silver, Bronze services);
- the QoS/SLA Management Service: it allows the handshake between a user and a site on a given SLA; moreover, it describes the QoS that a specific user/group has, both over a given site or generally in the PaaS as a whole. This includes information about the actual service quality of storage spaces and stored files at endpoints plus the possibility to change these service qualities for stored data.
- the Managed Service/Application (MSA) Deployment Service: it is in charge of

scheduling, spawning, executing and monitoring applications and services on a distributed infrastructure; it is implemented as a workflow programmatically created and executed by the Orchestrator, as detailed in the next section.

- the Infrastructure Manager (IM) [38]: it deploys complex and customized virtual infrastructures on IaaS Cloud deployment providing an abstraction layer to define and provision resources in different clouds and virtualization platforms;
- the Data Management Services: this is a collection of services that provide an abstraction layer for accessing data storage in a unified and federated way. These services will also provide the capabilities of importing data, schedule transfers of data, provide a unified view on QoS and distributed Data Life Cycle Management

3.3. Software to interface the Resource Centers with the PaaS

Based on the scientific use cases[9] the project considered, there were identified a set of features that have the potential to impact in a positive way the usability and easy access to the Infrastructure layers. In the computing area, these features are enhanced support for containers, integration of batch systems, including access to hardware specific features like InfiniBand and General Purpose GPUs, support for trusted container reposito- ries, introduction of spot instances and fair-share scheduling for selected Cloud Management Frameworks (CMF), as well as orchestration capabilities common to INDIGO selected CMFs using TOSCA. See Figure 7 for a graphical repre- sentation.

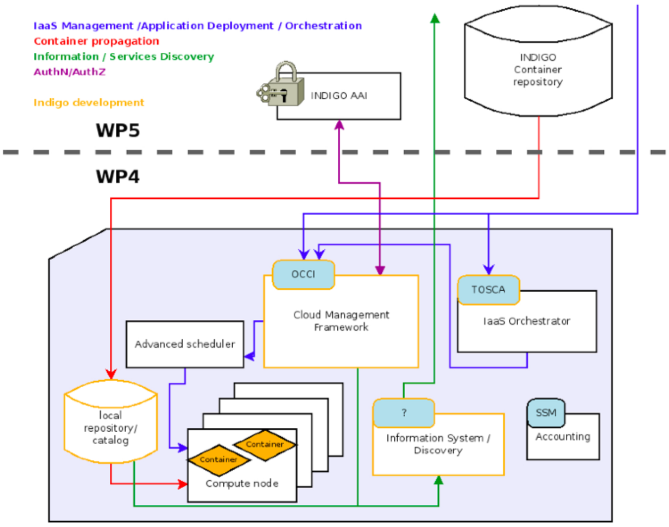

IaaS Management /Application Deployment / Orchestration
Container propagation
Information / Services Discovery
AuthN/AuthZ
Indigo development
INDIGO AAI
INDIGO Container repository
WP5
WP4
OCCI
Cloud Management Framework
TOSCA
IaaS Orchestrator
Advanced scheduler
local repository/ catalog
Container
Container
Compute node
?
Information System / Discovery
SSM
Accounting

Figure 7. Infrastructure view of the INDIGO architecture

For OpenStack and OpenNebula, the top two CMF's on the market, INDIGO, in collaboration with the corresponding Open Source communities, is spending significant efforts to make containers first-class citizens and, concerning APIs and management, indistinguishable from traditional VMs. While in OpenStack integration of Nova-Docker will introduce support for Docker containers, for OpenNebula, additional developments are required. In particular, the project developed OneDock[10], which introduces Docker as an additional hypervisor for OpenNebula, maintaining full integration with the Open- Nebula APIs and web-based portal (SunStone).

With the pressure of optimizing computer center resources but at the same time providing fair, traceable and legally reproducible services to customers, available cloud schedulers need to be improved. Therefore, the project is focusing on the support of spot-instances allowing brokering resources based on SLAs and prices. Technically this feature requires the CMF to be able to preempt active instances based on priorities. On the other hand, to guarantee an agreed usage of compute cycles integrated over a time interval, we need to invest in the evaluation and development of fair-share schedulers integrated in CMFs. This requires a precise recording of already used cycles and the corresponding readjustment of permitted current and future usage per individual or group. The combination of both features allows resource providers to partition their resources in a dynamic way, ensuring an optimized utilization of their infrastructures.

The middleware also provides local site orchestration features by adopting the TOSCA standard in both OpenStack and OpenNebula, with similar and comparable functionalities.

While in the cloud computing area, the specification of service qualities, e.g. number and power of CPUs, the amount of RAM and the performance of network interfaces, is already common sense, negotiating fine grained quality of service in the storage area, in a uniquely defined way, is not offered yet. Therefore, the high level objective of the storage area is to establish a standardized interface for the management of Quality of Services (QoS) and Data Life Cycle in Storage (DLC). Users of e-infrastructures will be enabled to query and control properties of storage areas, like access latency, retention policy and migration policies with one standardized interface. A graphical representation of the components is shown in Figure 8.

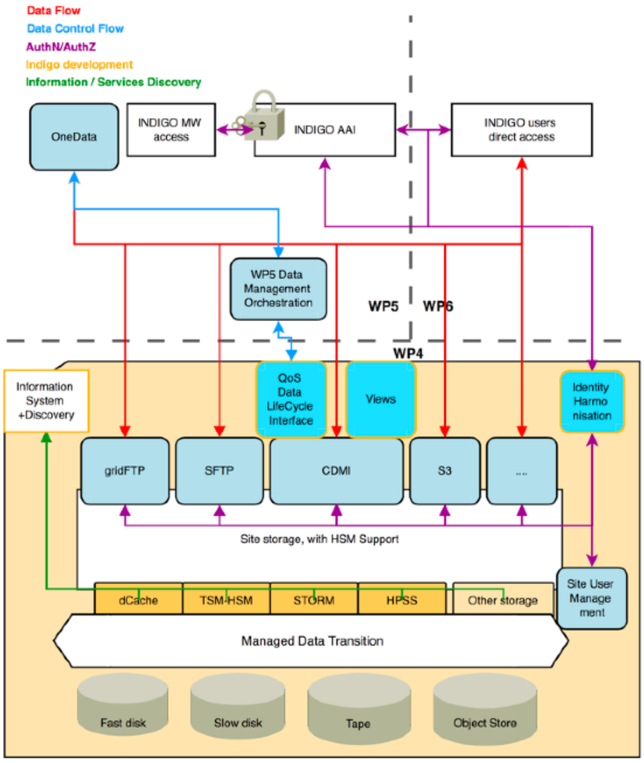


Figure 8. Storage services from the INDIGO architecture perspective

As with all infrastructure services, the interface is supposed to be used by either the PaaS storage federation layer or by user applications utilizing the infrastructure directly. This will be pursued in a component-wise approach. Development will focus on QoS and interfaces for existing storage components and transfer protocols that are available at the computer centers. Ideally, the Storage QoS component can be integrated just like another additional component into existing infrastructures.

3.4. Interfacing with the user

The INDIGO architecture needs to address the challenge of guaranteeing a simple and effective final usage, both for software developers and application running. A key component with a big impact on the end-user experience is the authentication and authorization meachanism employed to access the e-infrastructures. The identity harmonisation problem, described in more detail in the projects' AAI architecture document[11], has many aspects that need to be tackled:

- ability to authenticate users coming with di erent credentials
- ability to recognize which credentials are linked to which individuals, and provide a unique identi er linked to the individual (orthogonal to the dfferent credentials used)
- ability to link attributes to the identity that can be used to define and enforce authorisation policies
- ability to provision identity information and authorisation policies to relying services

To address these points the project intends to develop a service called *Identity Access Management (IAM)*, which provides a central solution for identity harmonisation, user authentication and authorisation. In particular, it provides a layer where identities, enrollment, group membership and other attributes management as well as authorization policies on distributed resources can be managed in a homogeneous way leveraging the supported federated authentication mechanisms (see Figure 9).

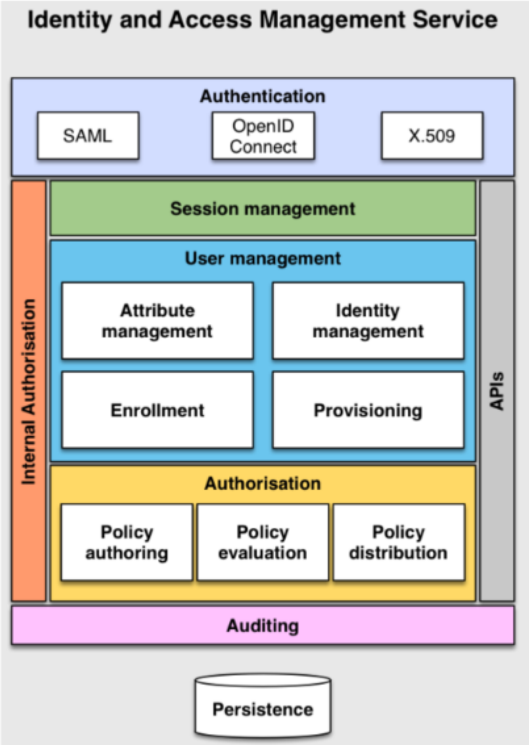


Figure 9. Architecture of the Identity Access Management INDIGO service

The IAM service supports standard authentication mechanisms as SAML, OpenID-connect (OIDC) and X.509. The user identity information collected in this way is then exposed to relying services through OpenID-connect interfaces. In a way, the IAM acts as a credential translator for relying services, harmonizing the different identity of the users and exposing them using a single standardized protocol. This approach simplifies integration at services as it doesn't force each service to understand and support each authentication mechanism used by users.

3.5. Graphical User Interfaces

We have developed the tools needed for the development of APIs to access the INDIGO PaaS framework. It is via such APIs that the PaaS features can be exploited via Portals, Desktop Applications or Mobile Apps. Therefore, the main goals of such endeavour from a high level perspective are to:

- Provide User Friendly front-ends demonstrating the usability of the PaaS services;
- Manage the execution of complex workflows using PaaS services;
- Develop Toolkits (libraries) that will allow the exploitation of PaaS services at the level of Scientific Gateways, desktop and mobile applications;
- Develop an Open Source Mobile Application Toolkit that will be the base for development of Mobile Apps (applied, for example, to a use case provided by the Climate Change community).

The architectural elements of the user interface (see Figure 11) can be de- scribed as follows:

- FutureGateway Portal: it provides the main web front-end, enabling most of the operations on the e-infrastructure. A general-purpose instance of the Portal will be available to all users.
- FutureGateway Engine: it is a service intermediating the communication between e-Infrastructures and the other user services developed. It incorporates many of the functionalities provided by the Catania Science Gateway Framework[12], extended by others specific to INDIGO. It exposes a simple RESTful API for developers building portals, mobile and desktop applications.
- Scientific Workflows Systems: these are the scientific workflow management systems orchestrating data and job flow. We have selected Ophidia, Galaxy, LONI and Kepler as the ones more demanded by the user communities.
- Wfms plug-ins - these are plug-ins for the Scientific Workflow Systems that will make use of the FutureGateway Engine REST API, and will provide the most common set of the functionalities. These plug-ins will be called differently depending on the Scientific Workflow system (modules, plug-ins, actors, components).
- Open Mobile Toolkit - these are libraries that make use of the FutureGateway Engine REST API, providing the most common set of the functionalities that can be used by multiple domain-specific mobile applications running on different platforms. Foreseen libraries include support for iOS and Android and, if required, for WindowsPhone implementations.
- INDIGO Token Translation Service Client - The Token Translation Service client enables clients that do not support the INDIGO-token to use the INDIGO AAI architecture.

3.6. Unified Data Access

The main goal in providing unified data access at the PaaS level is providing users with seamless access to data. The challenge resides in hiding the complexities and heterogeneities between the infrastructures where the data is actually being stored.

The INDIGO PaaS provides three data management services, Onedata, FTS and Dynafed, that allow accessing federated data in an unified way. Depending on how data are stored/accessible, they will be made available through a different services in a way which is transparent to the user (see Figure 10). In order to access and manage data, we will exploit the interfaces provided by the infrastructure layer:

- Posix and WebDAV for data access.
- GridFTP for data transfer.
- CDMI for the Metadata Management.
- REST APIs to expose QoS features of the underline storage.

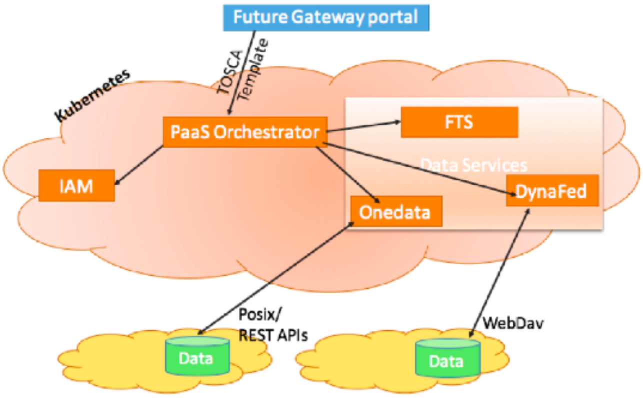


Figure 10. Data access high-level description and involved services

4. Conclusions and Future work

The INDIGO Architecture, based on a three level approach (User Interface, PaaS layer and Infrastructure Layer), is able to fulfil the identified requirements projects users communities. At the Infrastructure site level the architecture provides new scheduling algorithms for open source Cloud frameworks. Also very importantly, it provides dynamic partitioning of batch versus cloud resources at the site level. By implementing the Cloud bursting tools available in the architecture, the site can also have access to external infrastructures. The infrastructure site is also enhanced with full support to containers, where the local containers repositories can as well be securely synchronized with dock- erhub external repositories, facilitating enormously the automatic instantiation of applications. From the point of view of data, the architecture is able to integrate local and remote Posix access for all types of resources: bare metal, virtual machines or containers. In particular it provides transparent mapping of object storage to Posix, and a transparent Gateway to existing lesystem (like GPFS or LUS- TRE). Data access is also enhanced with the support to WebDAV, GridFTP and CDMI access.

The plan is to deliver a first release of the platform by July 2016, implementing the most important features to let users deploy their services and applications across a number of testbed provided by the INDIGO partners, and provide developers with an initial feedback. The first INDIGO-DataCloud release (Fig. 11) will provide open source components for:

- Data center solutions, allowing data and compute resource centers to increase efficiency and services for customers.
- Data solutions, offering advanced access to distributed data.
- Automated solutions, allowing users to easily specify and deploy complex data and compute resource requirements.

- User-level solutions, integrating scientific applications in programmable front-ends and in mobile applications.
- Common solutions - All the INDIGO components are integrated into a comprehensive Authentication and Authorization Architecture, with support for user authentication through multiple methods (SAML, OpenID Connect and X.509), support for distributed authorization policies and a Token Translation Service, creating credentials for services that do not natively support OpenID Connect.

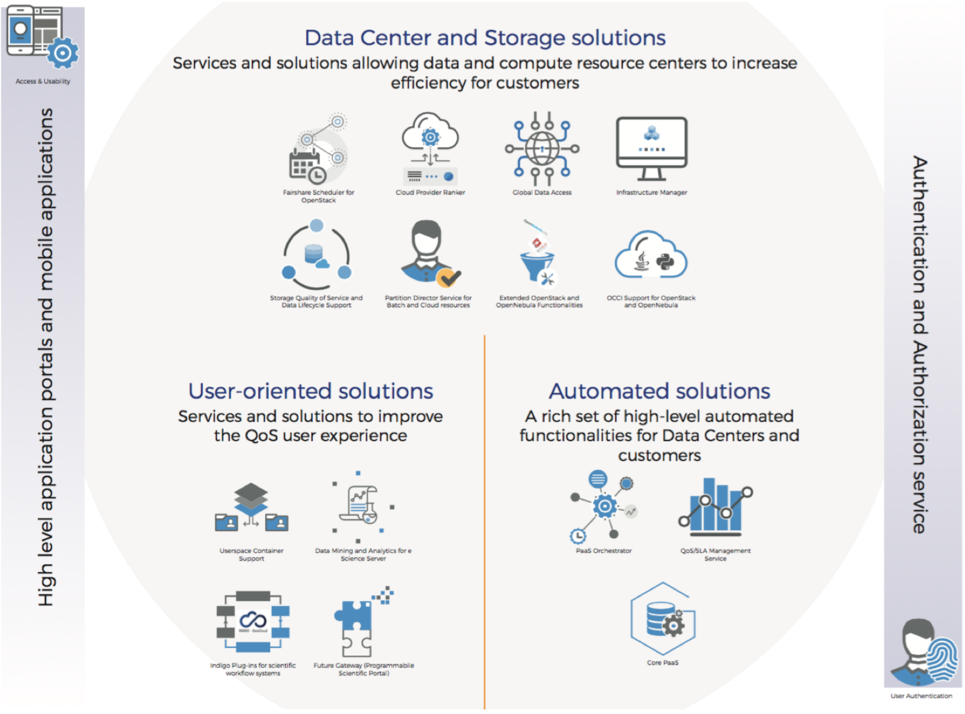

Access & Usability
High level application portals and mobile applications
Data Center and Storage solutions
Services and solutions allowing data and compute resource centers to increase efficiency for customers
Fairshare Scheduler for OpenStack
Cloud Provider Ranker
Global Data Access
Infrastructure Manager
Storage Quality of Service and Data Lifecycle Support
Partition Director Service for Batch and Cloud resources
Extended OpenStack and OpenNebula Functionalities
OCCI Support for OpenStack and OpenNebula
User-oriented solutions
Services and solutions to improve the QoS user experience
Userspace Container Support
Data Mining and Analytics for e Science Server
Indigo Plug-ins for scientific workflow systems
Future Gateway (Programmabile Scientific Portal)
Automated solutions
A rich set of high-level automated functionalities for Data Centers and customers
PaaS Orchestrator
QoS/SLA Management Service
Core PaaS
Authentication and Authorization service
User Authentication

Figure 11. INDIGO-1 (MidnightBlue) Solutions

The INFN-CNAF teams are involved not only in the management of the project, leading the Consortium coordination and Project Management, having the leadership of the WP3.2 task *Software release and maintenance* and WP5.2 task *Security and Authorization*, but also in the development and deployment activities like:

- the development of the Common Solutions, as main contributor of the Identity and Access Management (IAM) Service - a layer where identities, enrolment, group membership, attributes and policies to access distributed resources and services can be managed in a homogeneous and interoperable way.
- the Data Center Solutions,
  - developing the Dynpart, the Partition Directore Service for Batch and cloud resources. It facilitates the management of a hybrid data center that provides both batch-system based services and cloud-based services.
  - testing the CloudProviderRanker, a standalone REST WEB Service which ranks cloud providers basing on rules implemented with the Drools framework
- the Automated Solutions, testing and deploying on the project testbeds the core component of the PaaS layer, the Orchestrator, together with its dependencies: the QoS/SLA Management Service

- the testing and verification of the Data Service Solution, deploying the Onedata (OneZone and OneProvider) and the CDMI (Storage Quality of Service and Data Lifecycle support solution) on the pilot services testbeds

As activities of the next year we envisage not only to continue to contribute to the development and improvement of the software components, but also in providing most of project solutions on the Cloud@CNAF infrastructure.

References

[1] EC Expert Group Report on Cloud Computing *http://cordis.europa.eu/fp7/ict/ssai/docs/future-cc-2may-finalreport-experts.pdf*
[2] OpenStack *https://www.openstack.org/*
[3] OpenNebula *http://opennebula.org*
[4] Taverna *http://www.taverna.org.uk*
[5] Ophidia *http://ophidia.cmcc.it*
[6] Microservices *https://en.wikipedia.org/wiki/Microservices*
[7] Kubernetes *http://www.infoq.com/articles/scaling-docker-with-kubernetes*
[8] Docker *https://www.docker.com/*
[9] INDIGO official deliverable D2.1 *https://www.indigo-datacloud.eu/documents/initial-requirements-research-communities-d21*
[10] ONEdock *https://github.com/indigo-dc/onedock*
[11] AAI architecture document *https://owncloud.indigo-datacloud.eu/index.php/s/sUTRpymjANAX0Hd*
[12] Catania Science Gateway *http://www.catania-science-gateways.it*

The INFN Information System

Stefano Bovina[1], **Marco Canaparo**[1], **Enrico Capannini**[1], **Fabio Capannini**[1], **Samuele Cattabriga**[1], **Claudio Galli**[1], **Guido Guizzunti**[1] **and Stefano Longo**[1]

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

E-mail: stefano.bovina@cnaf.infn.it, marco.canaparo@cnaf.infn.it, enrico.capannini@cnaf.infn.it, fabio.capannini@cnaf.infn.it, samuele.cattabriga@cnaf.infn.it, claudio.galli@cnaf.infn.it, guido.guizzunti@cnaf.infn.it, stefano.longo@cnaf.infn.it

Abstract.

The Information System Service's mission is the implementation, management and optimization of all the infrastructural and application components of the administrative services of the Institute.

In order to guarantee high reliability and redundancy the same systems are replicated in an analogous infrastructure at the National Laboratories of Frascati (LNF).

The Information System team manages all the administrative services of the Institute, both from the hardware and the software point of view and they are in charge of carrying out several software projects.

The core of the Information System is made up of the salary and HR systems.

Connected to the core there are several other systems reachable from a unique web portal: firstly, the organizational chart system (GODiVA); secondly, the accounting, the time and attendance, the trip and purchase order and the business intelligence systems. Finally, there are other systems which manage: the training of the employees, their subsidies, their timesheet, the official documents, the computer protocol, the recruitment, the user support etc.

1. Introduction

The INFN Information System project was developed in 2001 to digitize and manage all the administrative and accounting processes of the INFN Institute, and to carry out a gradual dematerialization of documents.

In 2010, INFN decided to transfer the accounting system, based on the Oracle Business Suite (EBS) and the SUN Solaris operating system, from the National Laboratories of Frascati (LNF) to CNAF, where the SUN Solaris platform was migrated to a RedHat Linux Cluster and implemented on commodity hardware.

The Service Information System was officially established at CNAF in 2013 with the aim to develop, maintain and coordinate many IT services which are critical for INFN. Together with the corresponding office in the National Laboratories of Frascati, it is actively involved in fields related to INFN management and administration, developing tools for business intelligence and research quality assurance; it is also involved in the dematerialization process and in the provisioning of interfaces between users and INFN administration.

The information system service team at CNAF is currently composed of 8 people, both developers and system engineers.

Over the years, other services have been added, leading to a complex infrastructure that covers all aspects of people's life working at INFN.

2. Production services

2.1. INFN ERP system

The INFN Enterprise Resource Planning system (ERP) has been hosted at CNAF since 2011 and is based on the Oracle E-Business Suite.

This system is used to track all the information about business trips and purchase orders and its software tools are developed internally. Oracle E-Business Suite consists of a collection of tools with an Oracle Database as backend and is the main application used by INFN accounting offices.

While the application layer is available only to a restricted range of IP, the database is used as backend from several web applications with a restricted access.

The infrastructure of Oracle's E-Business Suite is hosted on a RedHat Cluster composed of 4 physical nodes with tree failover domains:

- a failover domain made up of 2 nodes for the database service
- a failover domain made up of 2 nodes for the application service
- a failover domain made up of 4 nodes for the backup service

The backup process require a complete switching off of the ERP system to ensure a consistent cold backup. This backup will be sent to a tape library and to the LNF site in Frascati for accomplishment of disaster recovery requirements.

2.2. Portale missioni

PHP web application based on CakePHP (a fast PHP framework), used to manage INFN employees business trips and purchase orders with all the related approval procedures.

This web application exploits the Oracle Database belonging to Oracle EBS.

2.3. Gestionaleweb

Java web application based on Vaadin (a Java Web UI Framework) and integrated with the Oracle EBS, used to manage electronic invoicing.

2.4. Sincro

Software used to manage the integration between Oracle EBS and banking systems for all payment orders.

2.5. Time and attendance system

The time and attendance system at INFN is a complex application whose main objective is tracking the hours that employees spend at work. Its most important features include:

- Showing the timestamps of the clock in and clock out of every employee
- Indicate the hours worked daily, monthly and every three month
- Showing clearly when the employee have to justify his/her absence
- Keeping track of the available amount of the different justifications demandable by employees in order to allow or not a particular request
- Generating a hierarchical workflow that manages managers approval or refusal of the different employees requests.
- Integration with business intelligence and human resources management systems

- Computation of the hours that an employee can request as overtime.

The "squaring" is the fundamental operation that computes, starting from raw data (timestamps) and regulations: the hours spent, overtimes management, meal tickets, the average quarterly worked hours.

Moreover, the system is in charge of checking if the justification of absence is used according to the regulations and if the amount requested is too little or too much compared to the necessity. The software code that involves the "squaring" operation is the one more subjected to modifications because of changes in government regulations or of different explanations given to the laws.

In 2015, the main activity focused on a reorganization of the flow of information concerning the personal details, the organizational chart, the employees' contracts, the people in charge of each service. Instead of being imported from different systems as it was before, currently, all these data are imported (nightly as default, or on demand) from only one system: the GODiVa service.

2.6. Vamweb

PHP web application called VamWeb, developed externally by Selesta Company, which manages the access and attendant systems.

The software gathers the timestamp of the clock in and clock out of employees and send them to the Time and Attendance System. The VamWeb software system is also used to control the access to some of the institutes canteens.

2.7. Business Intelligence

With the ever-growing necessity to let the public data manageable and available to be consulted by authorities, officers and supervisors, it had been built a Business Intelligence System.

Its architecture gives the chance to view the same bunch of information from different point of view, merging details coming from the three main branches of the services for the researchers and the staff (personal profiles, working details, balance for experiments and purchases) and driving the appraisal on the ongoing flow of procedures inside the offices and the laboratories belonging to the institution.

To sum up, it allows to make analysis and previsions about the progress of the activities, trying to reach the requirement of transparency together with personal data protection solicited by the current regulations.

2.7.1. The environments The Business Intelligence service provides a single machine for its data warehousing purposes and the daily ETL mechanism concerning the creation of the data tables: this is a Linux/Unix OS system, with an Oracle database system on which operate the transactions defined though Java runnable tasks created and shaped using Jaspersoft JETL application (that can independently run on local machines connected remotely to the data warehouse for testing and compiling final versions of the tasks). On the presentation tier, the architecture let have two paired configuration of the Jaspersoft Jasperreports Server on two separated environments, both running under their personal building of Tomcat with a PostgreSQL database for reports and profiling storage: one of them is constantly online to grant the access to the data to all the employees certified while the other, used for testing purposes, can be refreshed simply by copying the dump of the production machine database.

2.7.2. The reporting system There are three kinds of report presented into the service:

- the first kind is used to show content for the accounting of internal offices and the budget plans about the teams of experiments of the organization. Such documents are built on

OLAP cubes and XML views, managed with policies to cut information depending on the charge of the user;

- the second kind allows to examine the information about the people belonging to INFN and its sections, like working days, holidays, work permits and so on. These reports run upon mixed tools in Java and XML, merged into layout built with Jaspersoft Studio Professional
- a third group of reports, made using the same technology of the previous one are provided only for facilities and statistics used by the Informative System.

2.7.3. The policy system All users entering the Business Intelligence main page are immediately able to open reports filled with data filtered taking advantage of the LDAP credentials mechanism provided by the institution and its database of hierarchies, which ensures a set of strings to validate single sign-on credentials for the connection and allows to free browse the reports basing on the granted access to the employee. Those credentials define the user profile and its charge related to the contract and the role inside INFN chain of structures/sections, offices and workgroups. This access system is completed by the usage of the internal authorization list, that can be configured per report to allow the visualization of the single link for showing the dynamic document or under the definition of XML policies to be paired with OLAP cubes.

2.8. Piwik

Web analytics software used to tracks online visits and displays reports on these visits.

2.9. Nexpose

Vulnerability scanner software used to detect systems vulnerability.

3. Infrastructure

The INFN information system is located inside the CNAF Tier1 Datacenter where every hardware component is highly redundant. The connectivity between CNAF and the LNF sites has been accomplished by a Cisco VPN concentrator, that allows encrypted connections between these sites. The storage component is based on EMC VNX-5300 solution with Fibre Channel connectivity. The infrastructure of INFN information system is based on two different solutions:

- RedHat Cluster Suite: composed of some physical nodes for the Oracle Database and the Oracle Business Suite service
- oVirt virtualization manager: to manage KVM virtual machines (since Q4 of 2014)

We are currently managing a total of ninety machines (virtual and physical) and a usable storage of about 40TB (configured in RAID level 5 and 1).

4. Main Activities performed in 2015

4.1. Provisioning

In 2015 the main goals of this task were: the level out, the writing of documents and the automatization of setup of our services.

To achieve these targets, we had to revisit our services configurations and made some tuning. Furthermore, to improve the reproducibility and the documentations of our infrastructure and to introduce automatization we started to use configuration management(CM) tools as Puppet and Foreman as an external node classifier (ENC).

4.2. Monitoring

The next challenge was the replacement of current monitoring tool, Nagios, not designed for dynamic infrastructure and automations.

The chosen tools were Sensu and InfluxDB.

Sensu is specifically designed to solve monitoring challenges introduced by modern infrastructure platforms and allows you to reuse monitoring checks and plugins from legacy monitoring tools like Nagios. Furthermore it is completely integrated with Puppet.

InfluxDB is an open source database written in Go specifically to handle time series data with high availability and high performance requirements.

Both infrastructure, Foreman-Puppet and the monitoring ones, are provided by a cross group of people from different CNAF divisions and we actively take part of it.

4.3. Business Intelligence upgrade

The main task which involved the Business Intelligence system in 2015 was the upgrade of the overall software infrastructure both from the hardware and software point of view. In fact, the system was installed for the first time in 2012 and needed a complete renovation. At the same time, we conducted a migration of all the data both belonging to the datawarehouse and to the presentation tiers. The new infrastructure allows a better usage of memory and a faster execution of the tools for accessing information.

4.4. Oracle EBS Monitoring

Besides the standard EBS tools for monitoring, in 2015 some other PL/SQL tools were developed, registered and scheduled in the db. They send via email reports about the synchronization of data in the accounting system.

4.5. Unique organizational chart

Originally, the definition of the organizational chart of the different INFN branches was managed by some custom EBS forms. This led to some problems in the management and maintenance of this information. After choosing GODiVA as the centralized organizational chart information, in 2015 an import mechanism was implemented,in order to get all the necessary data.

4.6. HR improvements

In 2015 three activities have been carried out as regards the data exchange between the accounting system and the salary system.

Firstly, the projected end contract date have been imported in the accounting system. This is particularly important for the business trip system. Secondly, a data correction activity was necessary in particular it was focused on the national identifiers. Finally, it has been implemented the import of the unique identification number (UUID).

These three activities were the base for the most complex task that took up all of 2015 to be completed. This task consisted of two steps:

- the decoupling of the HR module of the INFN ERP from the former Payment Administration System
- the subsequent coupling of the HR module with the GODiVA System for the import of contract data.

The two systems (Godiva and HR-ERP) implement different data processing logics, so a full revision of the old pl-sql data alignment procedure was required in order to conform to the new Godiva interface. This revision has led to a significant improvement both in terms of efficiency of the data alignement procedure, and of service offered to the final users because it reduced the minimum elapsed time between the first hiring date and the contract replication date in ERP from 15 days to only one day.

4.7. Oracle EBS testing environment

A cloning procedure of the INFN ERP system has been put in place, in order to keep the test and development systems aligned with the production environment.

The cloning procedure is based upon a series of shell scripts which invoke RMAN tool, in order to backup the production database with its archive logs and subsequently restore it on separated systems.

We also set up an additional clone system where update procedures are carried out after the cloning phase, by migrating the underlying database from production version (11gR2) to the most recent version compatible with Oracle E-Business Suite R12.1.3 (12c).

The update procedure on this specific clone system also applies the most recent Oracle patches both to the database and to the application software.

4.8. Disaster Recovery

The Disaster Recovery project aims at protecting data belonging to INFN against external attacks. The project is composed of 3 main steps:

- Geographic copy between CNAF and LNF.
- Geographic alignment of all the databases.
- Geographic copy of all the software applications.

The main services involved are: the Salary system, the accounting system, the time and attendance system, the document management system, the protocol system, GODiVA, the Business Intelligence, the Information System unique web portal and some institutional web site hosted at LNF.

In 2015 the team carried out some monitoring activities, and some modifications to the storage system by rationalizing the busy areas. A complete test has been done starting from the backup data copied from CNAF to LNF. Some tests have been finished about the real time copy at CNAF of the oracle database of the GODiVA application system.

4.9. SSA

Electronic version of the already existing service ”Servizio Salute e Ambiente”

4.10. Docfornitori

Java based web application to manage suppliers related documents. With this project, we begun to study a CI system for our projects, using Docker containers.

4.11. Time and attendance system migration

Among the main activities we have carried out the porting of the ”Time and attendance system” system to a more recent version of Java and the deploy on more up to date systems, based upon Tomcat 7 and CentOS 6.

We also took advantage of the above mentioned activity to eliminate the lock-in on home made tools for building and packaging the application and migrating to standard tools specific for this particular purpose.

Additional Information

Organization

Director

Gaetano Maron

Scientific Advisory Panel

Chairperson	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*
	Pierre-Etienne Macchi	*IN2P3/CNRS, France*

Data Center – Tier1

Head: L. dell'Agnello

Farming	**Storage**	**Networking**	**Infrastructure**	**User Support**
A. Chierici	V. Sapunenko	S. Zani	M. Onofri	D. Cesini
S. Dal Pra	A. Cavalli	L. Chiarelli[1]	M. Donatelli	A. Falabella
G. Misurelli	D. Cesini	D. De Girolamo	A. Mazza	L. Morganti
S. Virgilio	E. Fattibene	F. Rosso		F. Noferini
	D. Gregori			M. Tenti
	M. Pezzi			S. A. Tupputi
	A. Prosperini			
	P. Ricci			

[1] GARR employee relocated at CNAF

Software Development and Distributed Systems

Head: D. Salomoni

Software Development

F. Giacomini
D. Andreotti
M. Caberletti
A. Ceccanti
G. Dalla Torre
M. Favaro
M. Manzali
E. Ronchieri
E. Vianello

Distributed Systems

C. Aiftimiei
R. Bucchi
V. Ciaschini
A. Costantini
D. Michelotto
A. Paolini
M. Panella
S. Taneja
G. Zizzi

National ICT Services

Head: R. Veraldi

S. Antonelli

External Projects and Technology Transfer

Head: M. C. Vistoli

A. Ferraro
B. Martelli

Information System

Head: G. Guizzunti

S. Bovina
M. Canaparo
E. Capannini
F. Capannini
S. Cattabriga
C. Galli
S. Longo
C. Simoni

Director Office

Head: A. Marchesi

Expenditure Centralization Office[2]

Head: M. Pischedda

[2] The office is under the INFN Director General.

Seminars

Jan. 30th	Johannes Gutleber **Towards a Future Circular Collider**
Feb. 20th	Daniele Andreotti **Continuous Integration and Testing with Docker**
Apr. 28th	Elisabetta Ronchieri, Stefano Dal Pra, Vladimir Sapunenko **Report da CHEP 2015**
May 28th	Carter Bullard **High-Performance Network Activity Analytics with Argus**
July 14th	**Il trasferimento tecnologico dell'INFN in Emila Romagna**
July 23rd	**Report da "INFN Visit on Agile Infrastructure" al CERN**
Sept. 15th	Sébastien Valat **MALT, a memory tracker**
Sept. 24th	Stefano Bovina, Giuseppe Misurelli **CNAF Bebop trio: deployment, monitoraggio e analisi dei log as-a-Service**
Oct. 5th	**Report da "INFN Visit on Agile Infrastructure" al CERN, parte seconda**
Oct. 26th	Tim Mattson **Big Data: What happens when data actually gets big?**
Nov. 6th	Matteo Favaro **Browser-Server Communication with WebSockets**
Nov. 12th	Elisabetta Ronchieri **Breve introduzione al linguaggio R**
Nov. 30th	Cristina Aiftimiei, Matteo Panella, Andrea Ceccanti **Report da OpenStack Summit e Docker Conference Europe**
Dec. 21st	Francesco Giacomini **Fondamenti di architettura dei calcolatori**