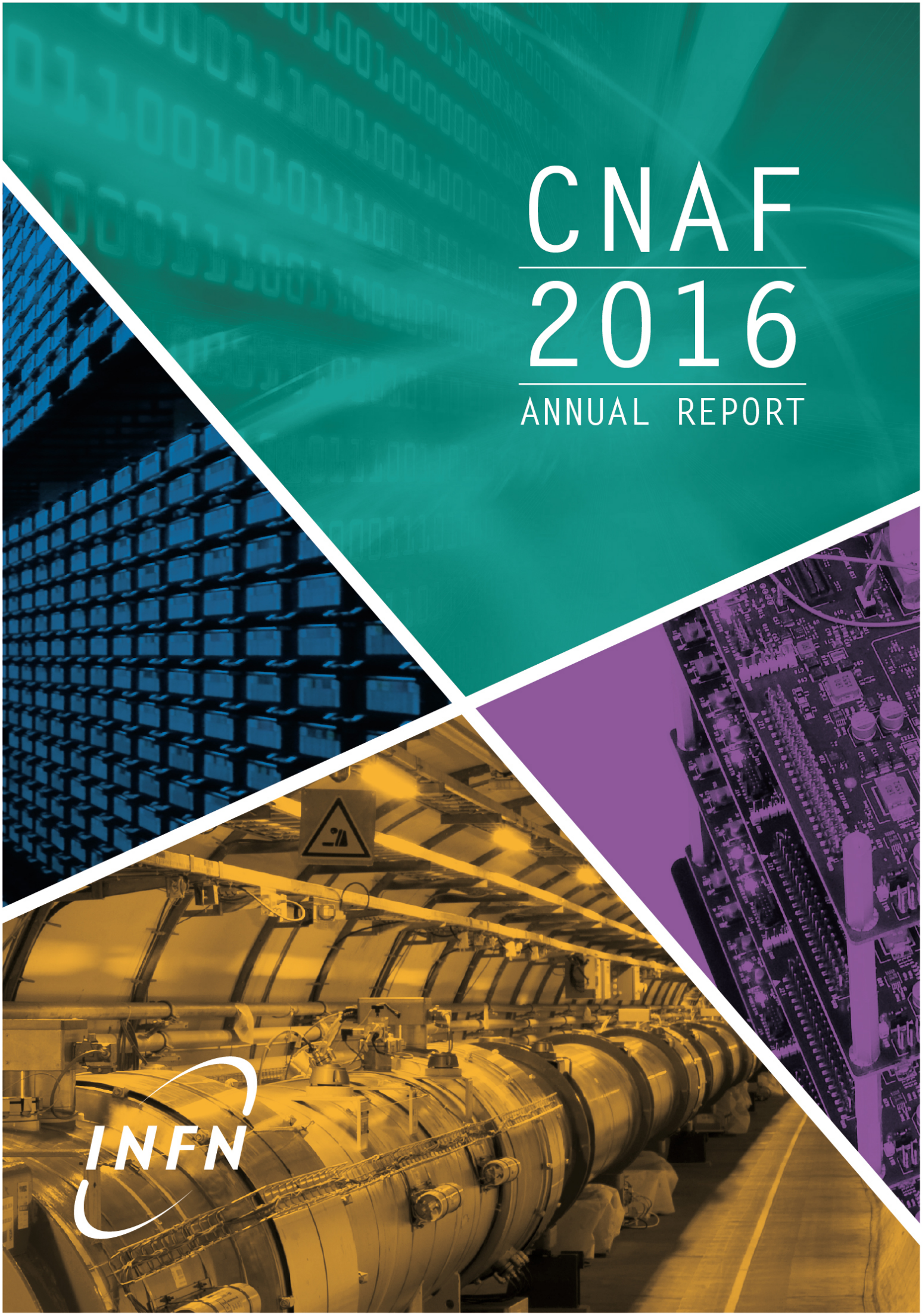

CNAF
2016
ANNUAL REPORT
INFN

INFN-CNAF Annual Report 2016

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

Editors

Luca dell'Agnello
Lucia Morganti
Elisabetta Ronchieri

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 Tier1 and Data center

Research and Developments

Technological transfer and other projects

Additional information

Introduction

The mission of CNAF can be summarized in four points, representing the key pillars on which the Center is based: scientific computing for the INFN research activities; innovation and development; IT services - including administrative services - of interest to all INFN; technology transfer to the public and private sectors. During 2016, all these activities pursued the planned objectives and achieved relevant results.

LHC RUN II resumed stable proton beams at 13 TeV in April 2016, confirming both its ability to exceed the design parameters with an instantaneous luminosity of $1.5 \times 10^{34} cm^{-2} s^{-1}$ and its impressive duty cycle to provide good collisions for physics (80%). For the various WLCG Tiers worldwide, this implied a significant increase in the amount of data to be stored and processed offline. LHC experiments were forced into taking corrective actions in their computing model, and in some cases to a drastic cancellation campaign of old data on tapes and, if possible, to an exploitation of opportunistic computing resources. The new requests affected CNAF as well and we could take advantage of the experience gained in 2015 on the elastic extension of our Tier1 towards remote data centers, both private and public. In this respect, the contribution of the Bari RECAS Center was significant, with up to 3000 cores of remotely available computing power. Due to the crucial role that external resource providers play in the strategies of CNAF, the Center became a member of the HNSciCloud Consortium, an H2020 project aiming at investigating the procurement and usage of private cloud services for our scientific community.

The CNAF Tier1 Data Center (DC) worked properly along the year, providing the pledged resources to the LHC experiments and ranking among the very first places of the WLCG Tier1s in terms of processed jobs and general availability of the center. Moreover, CNAF also provided computing and storage resources to many other particle-physics and astro-particle experiments in which INFN is involved. Currently, more than 30 non-LHC experiments use CNAF Tier1 computing resources, covering experimental setups as diverse as ground-based detectors like Virgo/Ligo to search for gravitational waves, detectors flying in space, like AMS-02 on the International Space Station, to study the universe and its origin, and underground experiments like DarkSide for dark matter searches. A stunning overview of these experiments can be found in this report. Furthermore, through its Cloud @ CNAF facility, the CNAF DC provides computing and storage resources to other research groups, ranging from the Digital Cultural Heritage domain represented by the INFN CHNet network, to the Computational Chemistry and Biomedical domains, to scientists working on experiments like EEE (Extreme Energy Events), where the cloud is used to acquire, store and process events from muon detectors installed in about 50 secondary schools spread over Italy. Finally, the CNAF DC hosts an HPC cluster dedicated to beam simulations to support the design of new particle accelerators, including the High Luminosity LHC.

The number of users exploiting the facilities of our DC is moderately high - over one thousand, and even though they are grouped in only a few dozen experiments, they require constant support and specific advice. For this reason, CNAF decided, just a few years ago, to create a User Support Team, which consists of post-doc students skilled on the computing model of the experiments, on the day-by-day operations of the Tier1, and on the middleware management applied to the experiment-specific issues.

In the middle of 2016, the INDIGO-DataCloud project announced its first public software release, code named midnightblue; INDIGO-DataCloud is an H2020-funded project with 26 partners coordinated by CNAF. The midnightblue release provides many software components addressing existing technical gaps linked to easy and optimal usage of distributed data and compute resources. It is the outcome of an initial phase of requirement gathering involving several scientific collaborations in many and various areas: earth science physics, bioinformatics, cultural heritage, astrophysics, life sciences and climatology. This first release provides a solid contribution towards the definition and implementation of an efficient

European Open Science Cloud. Along this development line, INDIGO-DataCloud provided the key components for a new EU project like EOSCPilot just funded by the EU and other projects (among them EOSCHub, Extreme Data Cloud - XDC and DEEP-HybridDataCloud) that are still under examination of the Commission. The desirable funding of these projects will enable us to pursue this fruitful and strategic development path towards an European Open Science Cloud that will provide a beneficial impact not only to scientific disciplines, but also to the manufacturing world, the tertiary sector and the public administration.

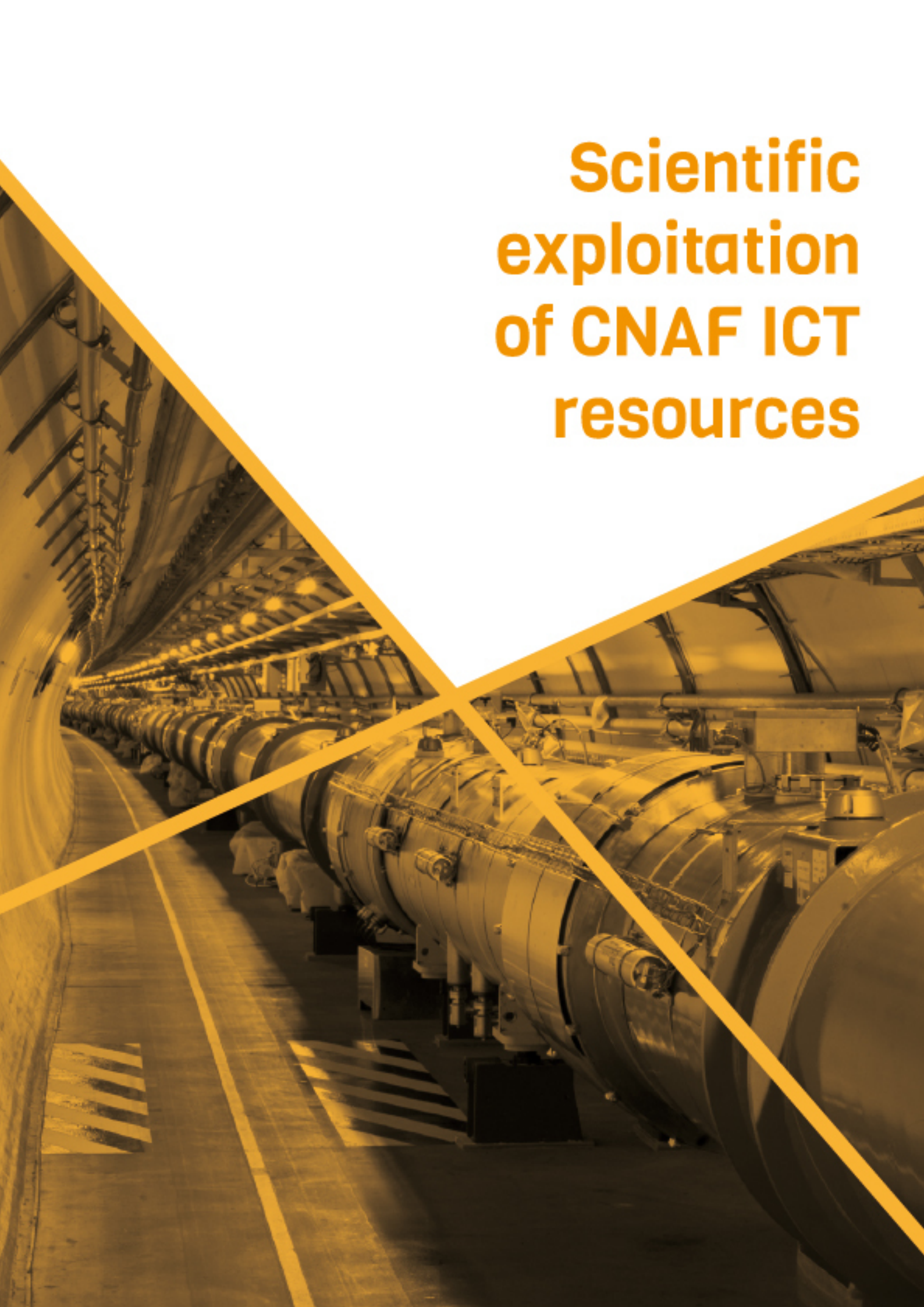
Many other development projects have affected several areas of direct interest to CNAF and to the experiments using the DC. Among others, the COSA project on the use of low-power processors in a data center in order to reduce energy consumptions; the development of important software for some experiments such as LHCb (event builder) or KM3NeT (data acquisition); the development and evolution of Grid components which are still crucial for WLCGs, such as VOMS, StoRM and Argus.

CNAF Knowledge Transfer has been mainly focused on the startup of the new Technology Transfer Laboratory (TTLab), which puts together various heterogeneous competencies of the INFN structures located in the Emilia-Romagna Region (Bologna and Ferrara Sections plus the CNAF Center) with the aim of promoting the transfer of INFN know-how in physics, computing, mechanics and electronics to regional enterprises. Among the first and important outcomes of TTLab in 2016, we recall a commissioned research to validate an innovative immersion cooling system targeted to big data centers and a POR-FESR funded project (OPEN-NEXT) on the use of low-power processors in industrial embedded systems. The assorted skills of TTlab's collaborators naturally led to projects benefiting from the meeting of different cultures; for instance, the organization and optimization of the analysis pipelines of genome sequencing for biophysical scientists. At the end of the year, an activity has started targeting the ISO-27001 certification, which would enable CNAF to store and manage private and sensitive personal data such as medical diagnostic information or genetic data.

In closing, in 2016 CNAF confirmed its attitude to scientific computing by supporting and successfully providing computing and storage resources to most INFN nuclear, subnuclear, and astro-particle physics experiments. The development line on Open Cloud systems, undertaken by CNAF in the last years and based on its decennial experience on distributed systems, already represents a recognized reference at European level and gives us the opportunity to be a major actor in the context of the European Open Science Cloud which is currently being established. All these highly innovative activities, coupled with the efficient Tier1 Data Center, represent a strategic asset for transferring our experience and knowledge to other scientific disciplines, and especially to those with the most direct impact on the lives and health of citizens. More generally, and thanks to the work of our TTLab Technology Transfer Laboratory, we saw in 2016 a positive and promising tendency to transfer our scientific and technological developments to society, but also to our manufacturing and tertiary production system, at regional, national as well as European level.

Gaetano Maron
CNAF Director

Scientific exploitation of CNAF ICT resources



The User Support unit at CNAF

D. Cesini, E. Corni, A. Falabella, L. Lama, L. Morganti, M. Tenti
INFN-CNAF, Bologna, IT
E-mail: user-support@lists.cnaf.infn.it

Abstract. Many different research groups, typically organized in Virtual Organizations (VOs), exploit the Tier-1 Data center 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 five fellows 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, monitoring frameworks and writing guides and documentation for users (see e.g. https://www.cnaf.infn.it/en/users-faqs/);
- 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 INFN 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 2016, we collaborated to renew the data center monitoring system including a complete refactoring of the Tier-1 experiments dashboard.

Moreover, the team was involved in the stress test of the transfer tool "dataclient" (see relative contribution), developed in order to match some typical needs from small experiments in terms of data transfer, namely the need to rely on a secure transfer tool hiding the complexity

of personal certificate management. Under the guidance of the User Support team, "dataclient" was successfully used by KM3Net, CUORE and CUPID-0 experiments during 2016.

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

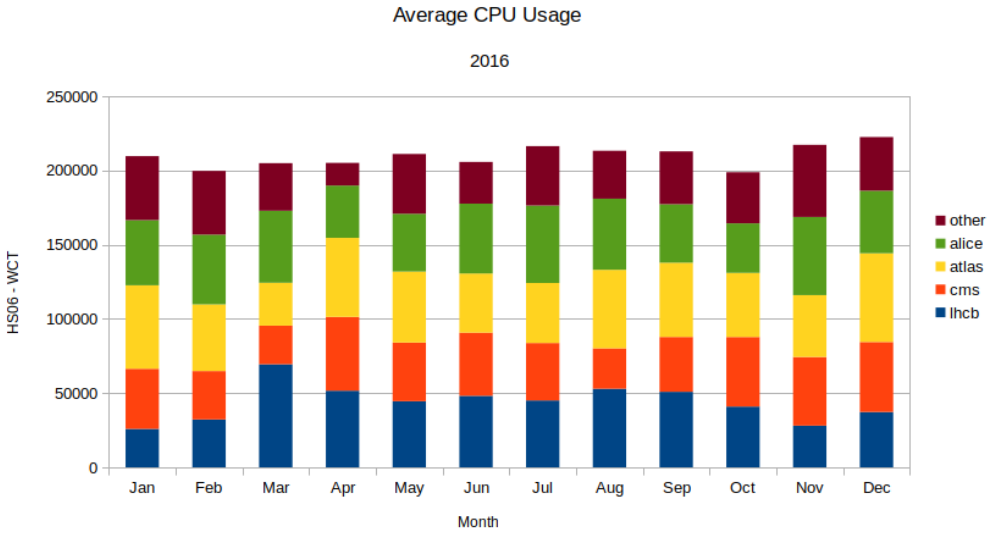


Figure 1. Monthly averaged CPU usage during 2016. Non-LHC experiments are grouped together (*Other*).

2. Supported experiments

Besides the four LHC experiments (ALICE, ATLAS, CMS, LHCb) for which CNAF acts as a Tier-1 site, during 2016 the User Support has also 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, Compass, CTA, Cuore, Cupid, Dampe, DarkSide-50, Enubet, Famu, Gerda, Fermi-LAT, Icarus, LHAASO, Juno, Kloe, KM3Net, LHCf, Magic, NA62, Opera, Padme, Pamela, Panda, Virgo, and XENON.

The following figures show resources pledged and used by the supported experiments during 2016. Fig. 1 refer to CPU, Fig. 2 to disk and Fig. 3 to tape.

On average, the LHC experiments represent three quarters of the total resources funded at CNAF, but many other research groups use CNAF resources significantly.

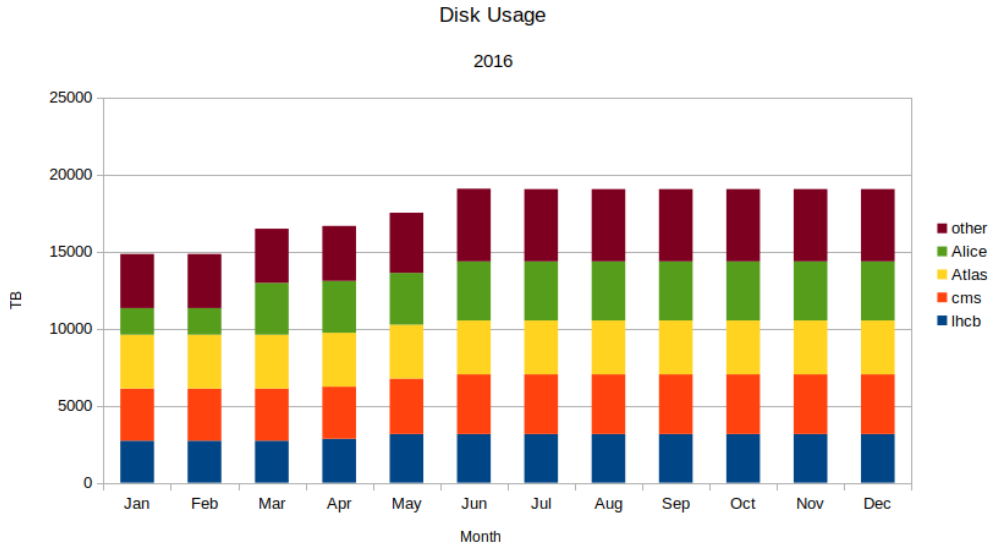


Figure 2. Disk used (TB) by all the supported experiments during 2016. Non-LHC experiments are grouped together (*other*).

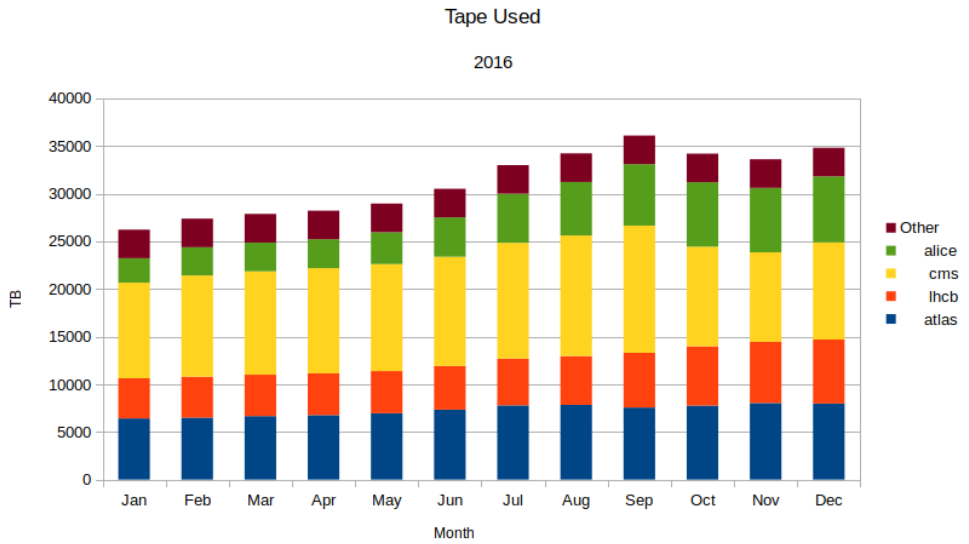


Figure 3. Tape used (TB) by the supported experiments during 2016. Non-LHC experiments are grouped together (*other*).

ALICE computing at the INFN CNAF Tier1

D. Elia[1], F. Noferini[2] and G. Vino[1,3]

[1] INFN, Bari, IT
[2] INFN, Bologna, IT
[3] GARR Consortium, Roma, IT

E-mail: Domenico.Elia@ba.infn.it

Abstract. In this paper the computing activities for the ALICE experiment at the CERN LHC are described, in particular those in connection with the contribution of the Italian community and the role of the Tier1 located at the INFN CNAF in Bologna.

1. Experimental apparatus and physics goal

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

The experimental apparatus consists of a central barrel part, which measures hadrons, electrons and photons, and a forward spectrometer to measure muons. It 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. 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 upgrade activities

The Run2 phase is progressing very well: in 2016 the experiment has operated very efficiently (above 90%), making good use of the excellent availability of the LHC (60%, almost a factor of 2 above the usual fraction). The pp data taking campaign at 13 TeV resumed on 23 April 2016, when stable beam

was declared for the first time. After completing its extensive commissioning which included the installation of 216 readout cards, the TPC was ready to be included in the first runs. With the first data collected during stable beams the ALICE team was able to test the performance of the new readout cards of the TPC as well as to check the data quality after the replacement of the gas of the TPC. Running with the gas mixture Ar:CO2 (90:10) resulted in stable operation in high particle flux. No high voltage trips have occurred at interaction rates up to 900 kHz and the commissioning of the new readout cards (RCU2) were completed. A Run3-like calibration procedure (TPC to ITS+TRD track interpolation) was developed and included in the standard TPC calibration framework. All the remaining ALICE detector subsystems have been performing as expected and taking physics data during the heavy ion and proton-proton-reference running.

At the beginning of November 2016 the total collected pp data sample resulted in about 800 M minimum bias as well as 500 M high-multiplicity rare events. ALICE took data smoothly during the whole pp period at an interaction rate of about 100 kHz, except for a short period at an interaction rate of 300-550 kHz dedicated to muons (SSD was unavailable). The last four weeks of data taking in November were devoted to collect p-Pb collisions at 5 and 8 TeV. In the 5 TeV period ALICE was leveled at a luminosity corresponding to about 20 kHz interaction rate and total samples of about 760 and 400 M events were collected, for minimum bias and centrality trigger selections respectively. For the 8 TeV p-Pb period the interaction rate reached 300 kHz and priority was given to the rare triggers: minimum bias and high-multiplicity samples in this case resulted in 120 and 50 M events, respectively. A total of 8 PB raw data has been collected in 2016.

Along 2016 ALICE has continued to finalise results from Run1 data taking, while also working on new results from the analysis of the Run2 samples. Many new physics results have been obtained from pp, p-Pb and Pb-Pb collisions. More than 30 papers have been submitted to journals in the last year, including in particular the following main topics: centrality evolution of the charged-particle pseudo-rapidity density over a broad pseudo-rapidity range in Pb-Pb at 2.76 TeV [4], anisotropic flow of charged particles in Pb-Pb at 5.02 TeV [5], multi-strange baryon production in p-Pb at 5.02 TeV [6], D-meson production versus multiplicity in p-Pb at 5.02 TeV [7]. Among the most remarkable results, strange and multi-strange hadron production in pp collisions at 7 TeV revealed a strangeness enhancement in high-multiplicity pp collisions, one of several surprising observations suggesting that a Quark Gluon Plasma may be formed in such collisions [8].

The general upgrade strategy for Run3 is conceived to deal with this challenge with expected Pb-Pb interaction rates of up to 50 kHz aiming at an integrated luminosity above 10/nb. 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 between 2014 and 2015. A transition from the R&D phase to the construction of prototypes of the final detector elements is currently taking place. For the major systems, the final prototype tests and evaluations are currently being performed in view of the corresponding production readiness reviews and start of production early 2017. The final prototype of the ALPIDE pixel chip for the ITS upgrade is under evaluation, with very promising results. The preproduction of the GEM foils for the ALICE TPC upgrade is under way and the final TPC chamber prototypes are being assembled. The prototype of the SAMPA chip, the common TPC and Muon frontend electronics, shows very good performance and the submission of the final prototype is in preparation for early 2017.

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 [9]. 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. 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 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. All raw data collected in 2016 has been passed through the calibration stages, including the newly developed track distortion calibration for the TPC, and has been validated by the offline QA process before entering the final reconstruction phase. The ALICE software build system has been extended with additional functionality to validate the AliRoot release candidates with a large set of raw data from different years as well as with various MC generators and configurations. It uses the CERN elastic cloud infrastructure, thus allowing for dynamic provision of resources as needed. The Grid utilization in the accounting period remained high, with no major incidents. The CPU/Wall efficiency remained constant, at about 85% across all Tiers, similar to the previous year. The much higher data rate forseen for Run3, tough, will require a major rethinking of the current computing model in all its components, from the software framework to the algorithms and to the distributed infrastructure. The design of the new computing framework for Run3, started in 2013 and mainly based on the concepts of Online-Offline integration ("O^2 Project"), has been finalized with the corresponding Technical Design Report [10]: 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 2016 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 [11,12].

Still on the R&D side in Italy, the design and development of a site dashboard project started a couple of years ago has been continued in 2016 and is going to be finalized in the first half of 2017. 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 sensors. The Bari dashboard has been exported in all the other sites along 2016: the project has now entered the final phase where a unique centralized dashboard for the ALICE computing in Italy is being 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 presented at CHEP 2016 [13].

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 experiment's 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 2016 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 72% of the ALICE computing activity was Montecarlo simulation, 14% raw data processing (which takes place at the Tier-0 and Tier-1 centres only) and 14% analysis activities: Fig.1 illustrates the share among the different activities in the ALICE running job profile along the last 12 months.

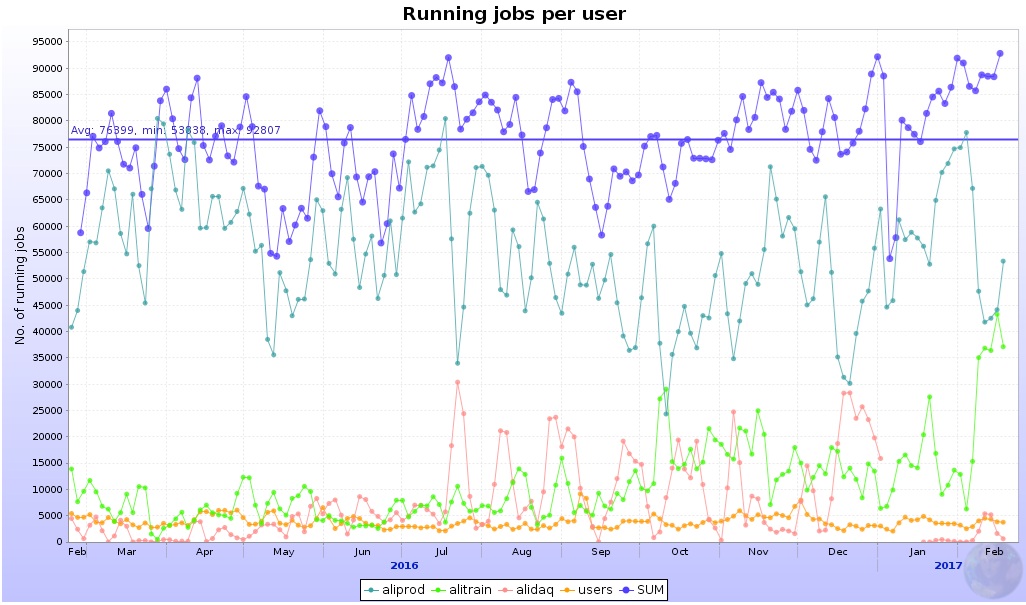


Figure 1. Share among the different ALICE activities in the 2016 running jobs profile.

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 current split of analysis activities, in terms of CPU hours, is about 35% individual jobs and 65% organized trains (5% and 9% of the total ALICE computing activity, respectively).

In 2016, CNAF deployed pledge resources corresponding to about 29 kHS06 CPU, 3900 TB disk and 5500 TB tape storage. The INFN Tier-1 has provided about 5% of the total CPU hours used by ALICE, ranking second of the ALICE Tier-1 sites, following only FZK in Karlsruhe. This amounts to about 30% of the total INFN contribution: it successfully completed nearly 6 million jobs, for a total of more than 20 million CPU hours. Fig.2 and 3 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.

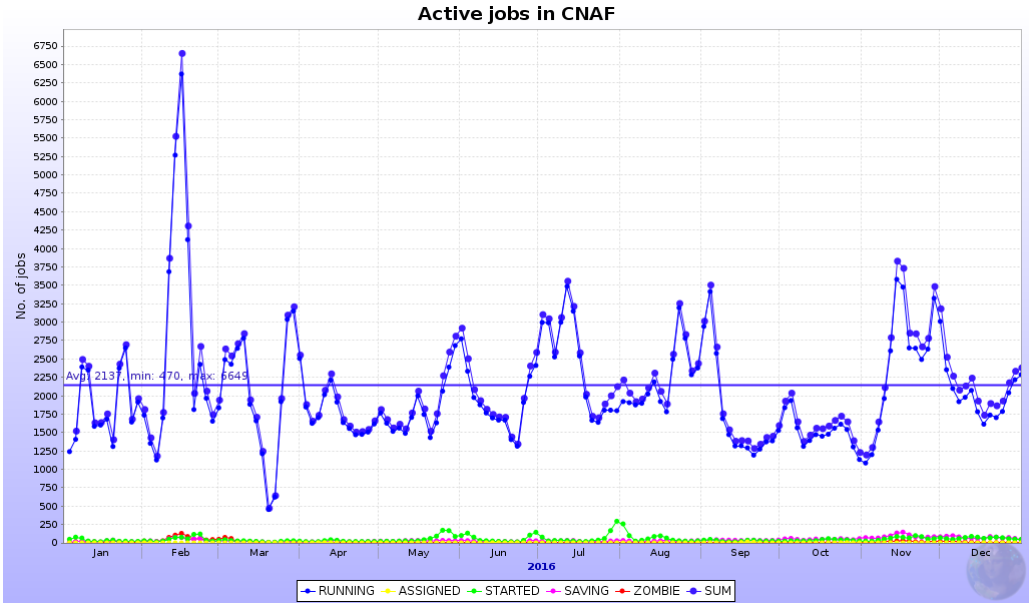


Figure 2. Running jobs profile at CNAF in 2016.

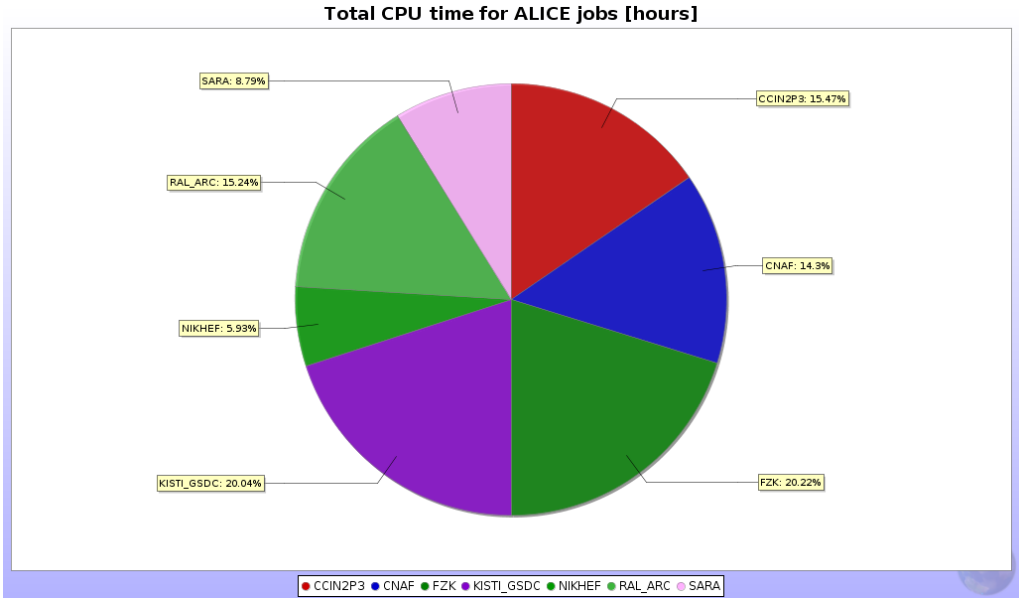


Figure 3. Ranking of CNAF among ALICE Tier-1 centres in 2016.

At the end of the last year ALICE was keeping on disk at CNAF more than 3 PB of data in nearly 60 million files, plus about 6 PB of raw data on custodial tape storage (500 TB in excess with respect to pledge); 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 40 Gb/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] J. Adam *et al*. (ALICE Collaboration), Physics Letters B **754** 373-385 (2016).
[5] J. Adam *et al*. (ALICE Collaboration), Physics Review Letters **116** 132302 (2016).
[6] J. Adam *et al*. (ALICE Collaboration), Physics Letters B **758** 389-401 (2016).
[7] J. Adam *et al*. (ALICE Collaboration), Journal of High-Energy Physics **8** 1-44 (2016).
[8] J. Adam *et al*. (ALICE Collaboration), arXiv:1606.07424v1, submitted to Nature Physics.
[9] P. Cortese *et al.* (ALICE Collaboration), CERN-LHCC-2005-018 (2005).
[10] J. Adam et al. (ALICE Collaboration), CERN-LHCC-2015-006 (2015).
[11] S. Piano et al., Journal of Physics: Conference Series **664** (2015) 022033
[12] D. Elia et al., Journal of Physics: Conference Series **664** (2015) 042013
[13] D. Elia, 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).

AMS-02 data processing and analysis at CNAF

B. Bertucci[1,2], **M. Duranti**[1,2], **D. D'Urso**[3,4,5,*]

[1] Università di Perugia, Perugia, IT
[2] INFN, Perugia, IT
[3] Università di Sassari, Sassari, IT
[4] ASDC, Roma, IT
[5] INFN-LNS, Catania, IT
AMS experiment `http://ams.cern.ch`, `http://www.ams02.org`, `http://www.pg.infn.it/ams/`

E-mail: * `ddurso@uniss.it, 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 $\sim$ 90 billion events up to the end of 2016. CNAF is one of the repositories of the full AMS data set and contributes to the data production and Monte Carlo simulation of the international collaboration. It represents the central computing resource for the data analysis performed by Italian collaboration. In the following, the AMS computing framework, the role of the CNAF computing center and the use of the CNAF resources in 2016 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 $\sim$ 90 billion events up to the end of 2016. The signals from the $\sim$ 300.000 electronic channels of the detector and its monitoring system (thermal and pressure sensors) are reduced on board to match the average bandwidth

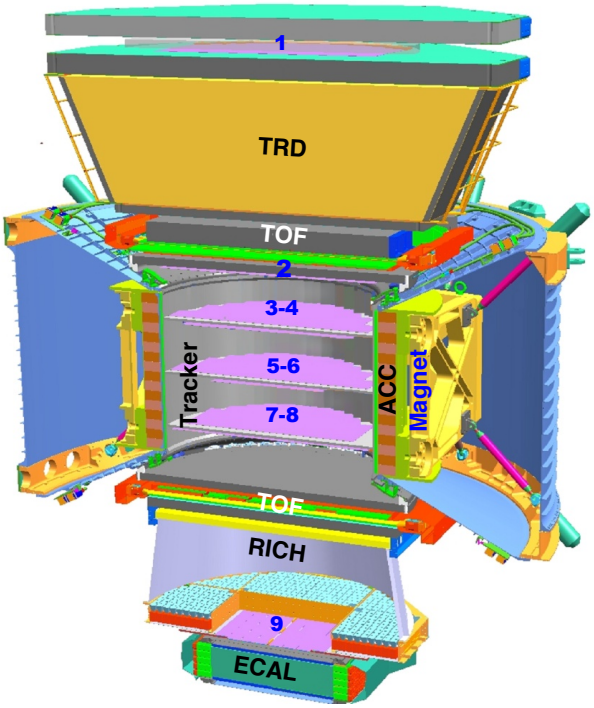


Figure 1. AMS-02 detector consists of nine planes of precision silicon tracker, a transition radiation detector (TRD), four planes of time of flight counters (TOF), a permanent magnet, an array of anticoincidence counters (ACC), surrounding the inner tracker, a ring imaging Cherenkov detector (RICH),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

of ∼10 Mbit/s for the data transmission from space to ground, for a ∼ 100 GB/day of raw data produced by the experiment.

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 to the data production and Monte Carlo simulation of 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 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 2016 there were 6 full data reproductions done. The last 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).

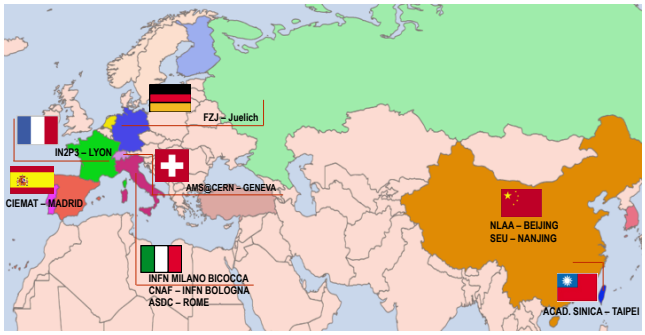


Figure 2. AMS-02 Major Contributors to Computing Resources.

Usually 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 9800 HS06 and 1790 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 30 users are routinely performing the bulk of their analysis at CNAF, transferring to local sites (i.e. their small local computing farm or their laptop) 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 and the AMS farm located in INFN-Perugia, to allow an efficient use of the local computing resources ($\sim$ 550 cores and $\sim$ 200 TB). The solution is based on the use of the General Parallel File System (GPFS) and of the Tivoli Storage Manager (TSM).

4. Data processing strategy at CNAF

At CNAF, on the local filesystem, are available the last data production and a subsample of Monte Carlo production. Due to the high I/O throughput from disk of AMS jobs, $\sim$ 1MB/s

(1000 running jobs may generate $\sim$ 1GB/s of I/O throughput from disk), the AMS Collaboration verified the possibility to access, in an lsf job, all the files of AMS production at CERN via xrootd. To efficiently use all of the resources allocated at CNAF, in terms of disk and CPU time, AMS decided to adopt a double strategy to process AMS data files: lsf jobs using data stored on local filesystem and lsf jobs accessing data files at CERN via xrootd protocol. We limit the number of jobs running on local filesystem to 800 and use all the available computing resourse to process files via xrootd, not saturating the access to local disk. Typical use of CNAF resources is shown in figure 3: the number of pending (green) and running (blue) AMS jobs in the top plot and the input/output (green/blue) network traffic rate, in the bottom plot, as a function of time are displayed. From figure 3, one can observe the increasing in the network traffic correlated to the number of running jobs making use of xrootd.

Currently, AMS server disk can support I/O throughput up to 40 Gb/s and we are verifying our strategy, increasing the number of lsf running on local disk and accessing external files via xrootd only when they are not locally available or if the local filesystem is going to be saturated.

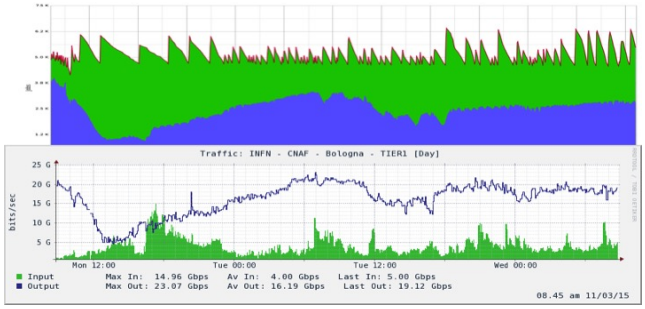


Figure 3. Number of pending (green) and running (blue) AMS jobs, in the top plot, and input (green)/output(blue) network traffic rate, on the lower plot, as a function of time

.

5. Activities in 2016

AMS activities at CNAF in 2016 have been related to data reprocessing, Monte Carlo production and data analysis. Those activities have produced two publications reporting the measurement of the anti-proton [3] and boron to carbon ratio [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 11520 minutes (it allows to run 8 core multi-thread jobs for 1 day) 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 400 jobs as 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 the 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.

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 proton, helium and light nuclei ions for AMS publications.

To support Monte Carlo campaign, special LSF profile has been implemented to allow AMS users to submit multi-thread simulation jobs. The AMS collaboration in 2016 used ∼11000 CPU-years for MC production. In particular in 2016 the collaboration started to face one of the main physics objectives of the experiment: the search for primordial anti-matter, i.e. anti-Helium. Being the anti-Helium more rare than 1 particle over 10 million Helium particles, and so the signal/background so tiny, a large effort to produce a MC sample of the background (i.e. Helium) is needed in order to have a statistical meaning sample to search Helium particles being mis-identified as anti-Helium. A large MC Helium production, with 35 billion simulated events, corresponding to ∼ 6000 CPU-years, has been conducted. This effort has been shared among the various AMS collaboration production sites, including CNAF, as shown in Fig.4.

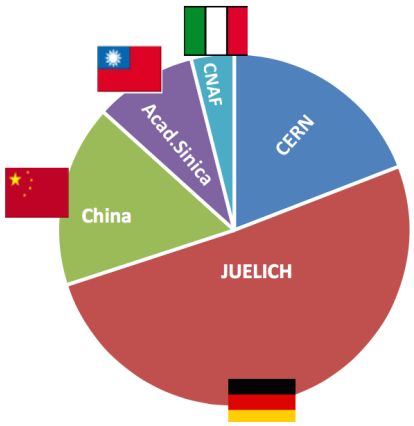


Figure 4. Sharing among the various production sites of the ∼ 6000 CPU-years needed for the anti-Helium analysis.

.

Data analysis

Different analysis are carried on by the Italian collaboration. In 2016, the CNAF resources for user analysis have been devoted to several different topic: the update, with more statistics, of the electron and positron analyses, the study of their time variation as well as the study of the proton and helium flux as function of time, the measurement of the light nuclei abundances (that resulted in the Boron to Carbon flux ratio publication on PRL [4]), the deuteron abundance measurement and the antiproton analysis (resulted in a publication on PRL [3]). The disk resources pledged in 2016, ∼ 1.8 PB, were mostly devoted to the PASS4/PASS6 data sample (∼ 1 PB), MC data sample (∼ 400 TB), selected data streams (∼ 100 TB of pre-selected data used for common electron/positron, antiproton, 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,117 (2016) ,091103.1-10.
[4] M.Aguilar *et al.*, AMS-02 Collaboration, Phys.Rev. Lett,117 (2016) ,231102.1-8.
[5] D. D'Urso & M. Duranti, Journal of Physics: Conference Series, 664 (2015), 072016 .
[6] http://xrootd.org.

ATLAS activities

A. De Salvo
INFN Roma 1, Roma, IT

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). 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 ~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.

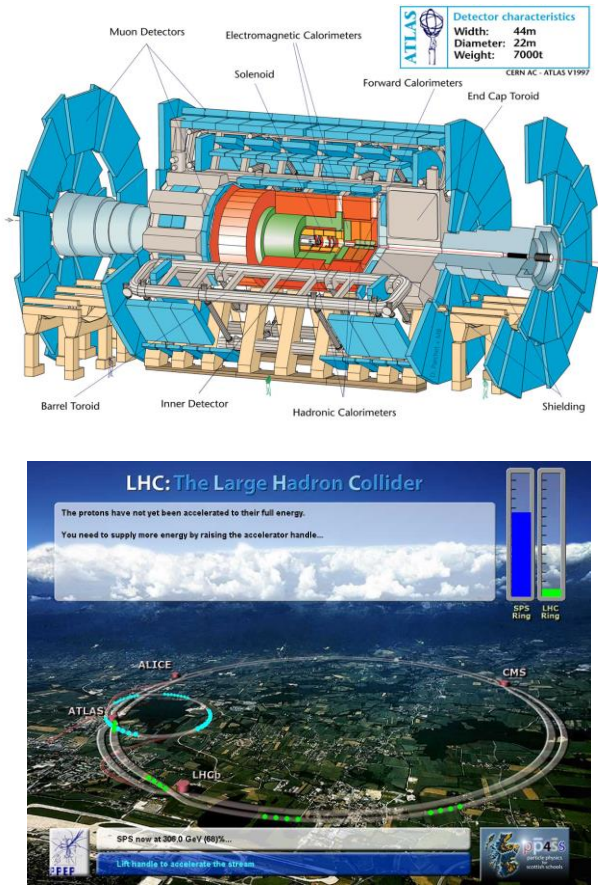


Figure 1. The ATLAS experiment at LHC

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 & 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.

2.1.1. 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 & Computing Project reflects all areas of activity within the project itself. Strong high-level links are established with other parts of the ATLAS organization, such as the T-DAQ Project and Physics Coordination, through cross-representation in the respective steering boards. The Computing Management Board, and in particular the Planning Officer, acts to make sure that software and computing developments take place coherently across sub-systems and that the project as a whole meets its milestones. The International Computing Board assures the information flow between the ATLAS Software & Computing Project and the national resources and their Funding Agencies.

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 a higher importance with respect to normal Tier-2s and can have primary data too. They are also considered satellites of the Tier-1, also identified as nucleus. 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. Since 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 2016 and other analysis.

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 2016, for some specific activities:

1) fine tuning and performance improvements of the Xrootd federation using the StoRM storage system, completely developed by CNAF within the LCG and related projects, funded by EU;
2) 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;
3) improvements of the dynamic model of the multi-core resources operated via the LSF resource management system;
4) network troubleshooting via the Perfsonar-PS network monitoring system, used for the LHCONE overlay network, together with the other T1 and T2 sites;
5) planning, readiness testing and implementation of StoRM for the future infrastructure of WLCG
6) 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 2016, in order to analyse the data of the Run-2 and produce the Monte Carlo data needed for the 2016 run.

The LHC data taking started in April 2016 and, until the end of the operation in December 2016, all the Italian sites, 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. ATLAS collected and registered at the Tier0 ~35.6 fb-1 and ~20 PB of raw and derived data, while the cumulative data volume distributed in all the data centers in the grid was of the order of ~50 PB. The data has been replicated with an efficiency of 100% and an average throughput of the order of ~12 GB/s during the data taking period, with monthly average peaks above 16 GB/s. For just Italy, the average throughput was of the order of 800 MB/s with monthly average peaks around 1000 MB/s. The data replication speed from Tier0 to the Tier2s has been quite fast with a transfer time lower than 4 hours. 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 150k in Dec, with an average CPU efficiency up to more than 80%.

The use of the grid for analysis has been stable on ~37k simultaneous jobs, with peaks around the conferences' periods to over 54k, 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 hierarchy of redirectors, thus reducing the number of failures due to missing or not accessible local files, while also giving the possibility to relax the data management and storage requirements in the sites. The FAX federation is now in production mode and used as failover in many analysis tasks.

The Italian sites also contributed to the development of the http/webdav federation, where the access to the storage resources is managed using the http/webdav protocol, in collaboration with the CERN DPM team, the Belle2 experiment, the Canadian Corporate Cloud ant the RAL (UK) site. The purpose is to build a reliable storage federation, alternative to FAX, to access physics data both on the grid and on cloud storage infrastructures (like Amazon S3, MicroSoft Azure, etc). The Italian community is particularly involved in this project and the first results have been presented to the WLCG collaboration.

The Italian community also contributes to develop new tools for distributed data analysis and management. At the end of 2016 the old data management tool DQ2 was definitely dismissed and now only the new Rucio framework is used. Another topic of interest is the usage of new computing technologies: in this field the Italian community contributed to the development and testing of muon tracking algorithms in the ATLAS High Level Trigger, using GPGPU.

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 2016 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.

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] INFN, Lecce, IT

[2] Observatorio Pierre Auger, Malargüe, Argentina

(Full author list : http://www.auger.org/archive/authors_2016_12.html)

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

Abstract. The Pierre Auger Observatory has begun a major Upgrade (AugerPrime), with an emphasis on improved mass composition determination using the surface detectors of the Observatory. AugerPrime, will be upgraded with new scintillation detectors for a more detailed measurement of gigantic air showers. This is required to identify cosmic objects that accelerate atomic particles up to highest energies. In this paper the adopted computing model is summarized and the computing organization of the Italian part of the collaboration is explained. Among the physics results, the measurement of the spectrum and the results from the study of the mass composition are presented.

1. Introduction

The data taken with the Pierre Auger Observatory[1] contributed to a number of steps forward in the field of ultra-high energy cosmic rays (UHECRs). The measurements confirmed with high precision the suppression of the primary cosmic ray energy spectrum at energies above $5\text{x}10^{19}$ eV[2]. This reduction is compatible with the Greisen-Zatsepin-Kuzmin (GZK) effect, but the level of its impact to the cut-off remains unclear. The measured photon and neutrino fluxes limits at ultrahigh energy[3] indicate that top-down mechanisms such as the decay of super-heavy particles cannot be the main producer of the observed particle flux. The distributions of the depth of shower maximum (X_{max}) evaluated for different energy intervals have been used to determine the UHECR composition on Earth, surprisingly evidencing the presence of a large fraction of protons in the energy range of the spectral ankle. At the same time, according to the Auger data, the anisotropy of the arrival directions of these protons cannot be larger than a few percent. Moreover the proton component disappears just below 10^{19} eV where a helium component appears. These transitions indicate that we do not have enough composition-sensitive data to obtain the composition at energies higher than a few times 10^{19} eV.

In order to extend the composition sensitivity of the Auger Observatory into the flux suppression region, an upgrade of the Auger Observatory (named AugerPrime [4, 5]) has been planned. The main aim of AugerPrime is to provide additional measurements of composition-sensitive observables, allowing to determine the primary mass of the highest energy cosmic rays on a shower-by-shower basis. The study of the origin of the flux suppression will provide fundamental constraints on the astrophysical sources and will allow us to determine more precise estimates of gamma-ray and neutrino fluxes at ultra-high energy. The measurement of the flux

contribution of protons will elucidate the physics potential of existing and future cosmic ray, neutrino, and gamma-ray detectors. In order to do so, the aim of AugerPrime is to reach a sensitivity as small as 10% in the flux contribution of protons in the suppression region. The determination of the primary mass composition of ultra-high energy cosmic rays is deeply related to our understanding of extensive air showers and hadronic interactions. In the Auger data, there is a disagreement between the observed and expected muon numbers, therefore it is of fundamental importance to study the hadronic multiparticle production in extensive air showers.

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[6]. 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. Given the upgrade phase of the experiment there is a transition phase also in the simulation of the apparatus to fully include the upgraded detector, and the simulation campaigns executed under GRID are momentarily stop.

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.

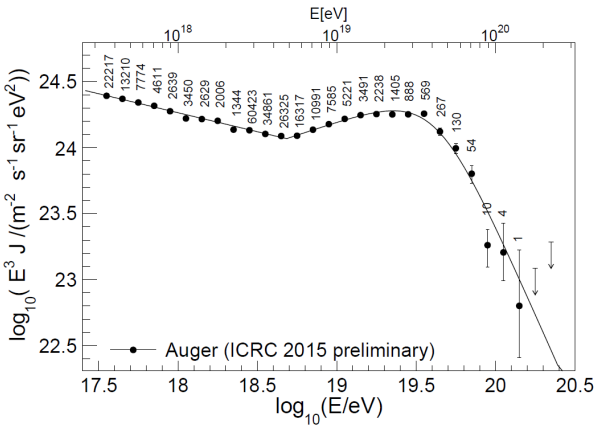


Figure 1. The combined energy spectrum measured by the Auger Observatory, fitted with a flux model. The data points include only the statistical uncertainties. For each data point is reported the number of events.

4. The Energy Spectrum

The events used for the determination of the energy spectrum consist of 4 different sets of data: the SD-1500 vertical events with zenith angle up to 60°, the SD-1500 inclined events with zenith angle between 60° and 80°, the SD-750 vertical events and the hybrid events. The hybrid set of data contains events detected simultaneously by the fluorescence telescopes and by at least one WCD.

The first step in the procedure used for the determination of the spectrum is the evaluation of the energy of the events. The FD allows the measurement of the electromagnetic energy released by the shower in the atmosphere as a function of the atmospheric depth. The total primary energy is then derived by integrating this longitudinal profile over the depth range and adding an estimate of the so-called invisible energy carried into the ground by high-energy muons and neutrinos.

The SD samples the shower particles that reach the ground. The intensities of the signals registered in the WCD are used to quantify the shower size and the impact point of the shower axis on the ground.

The absolute calibration of the SD sets of data is inferred from a high-quality subset of hybrid events (full details in [7, 8, 9]).

The final step in the procedure used for the determination of the energy spectrum is a precise evaluation of the exposure. Above the energy for full detector efficiency, the calculation of the SD exposure is based on the determination of the geometrical aperture of the array for the corresponding zenith-angle interval and of the observation time. The choice of a fiducial trigger based on active hexagons allows one to exploit the regularity of the array, and to compute the aperture simply as the sum of the areas of all active hexagons. The calculation of the exposure for the hybrid set of data is more complex. It relies on a detailed time-dependent Monte Carlo simulation which exactly reproduces the data taking conditions and includes the response of the hybrid detector [10].

The energy spectrum reported in figure 1 has been obtained by combining the four independent sets of data. They are combined using a method that takes into account the systematic uncertainties of the individual measurements (see details in [11]).

The characteristic features of the combined energy spectrum, shown in figure 1, have been quantified by fitting a model that describes a spectrum by a power-law below the ankle

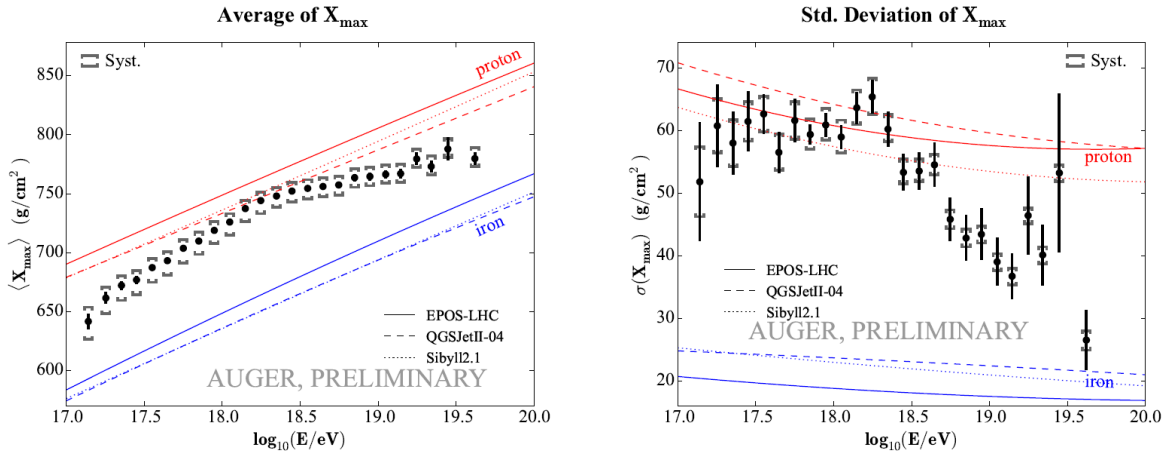


Figure 2. Average and RMS of the X_{max} compared to the model predictions for an all-proton and an all-iron composition.

$J(E) = J_0(E/E_{ankle})^{\gamma_1}$ and power-law with a smooth suppression at the highest energies:

$$J(E) = J_0 \left(\frac{E}{E_{ankle}}\right)^{-\gamma_2} \left[1 + \left(\frac{E_{ankle}}{E_s}\right)^{\Delta\gamma}\right] \left[1 + \left(\frac{E}{E_s}\right)^{\Delta\gamma}\right]^{-1}$$

Here, γ_1 and γ_2 are the spectral indexes below and above the ankle energy E_{ankle} respectively, E_s is the energy at which the differential flux falls to one-half of the value of the power-law extrapolation from the intermediate region, $\Delta\gamma$ gives the increment of the spectral index beyond the suppression region, and J_0 is the normalization of the flux, taken as the value of the flux at $E = E_{ankle}$.

The energy spectrum can also be exploited to study the distribution of cosmic-ray sources by searching for a flux variation with declination (δ) of the incoming directions. This study is of particular interest to the discussion of the difference seen in the suppression region between the spectra measured by Auger and by the Telescope Array experiment [12], which, despite being still compatible within the quoted systematic uncertainties of both experiments, is not understood so far.

5. The Mass Composition

Different observables can be used to obtain information on the primary composition, the most direct of which is the depth of maximum development of the longitudinal shower profile (X_{max}), measured by the FD. X_{max} is related to the depth of the first interaction of the primary and to the subsequent development of the shower. For this reason, the interpretation in terms of composition is complicated by the large uncertainties in the hadronic interaction models used in the simulations. The average of the X_{max} for different energies of the primary and its RMS can be directly compared to the predictions of air shower simulations using recent post-LHC hadronic interaction models, as shown in figure 2.

Our measurements are clearly at variance with model predictions for pure composition; assuming no change in hadronic interactions at these energies, they point to a composition getting heavier above the ankle.

6. The Upgrade Program of the Experiment

Taking data until the end of 2024 will double the present surface detector (SD) event statistics and reduce the total statistical uncertainty at the highest energies. With the planned upgraded detector running for 7 years we can expect about 700 events above $3x10^{19}$ eV and more than 60 above $6x10^{19}$ eV for zenith angles less than 60°. "Horizontal air showers"will add about 30% to the exposure and thus to the number of expected events. Accounting for a detector resolution of 15% or better in determining the number of muons, this would allow a separation of a fraction as small as 10% of protons from intermediate and heavy primaries. The key question is whether we can use additional information on the separation between the electromagnetic and muonic shower components for improving the estimate of the mass of the primary particles adding an extra measurement of the particles in the EAS independent of the measurements made with the water-Cherenkov detectors (WCD). To achieve the maximum advantage from this additional measurement, the shower should be sampled in the position of the WCD with a detector that has a different response to the basic components of the EAS. Moreover, the additional detector has to be reliable, easy to realize and install, and has to have minimal maintenance. Overall, the expectations from air shower simulations strongly indicate the feasibility of composition determination at the highest energies. It can be expected that, if the detector resolution in determining the number of muons and the X_{max} is smaller or of the order of the shower fluctuations, the primary mass can be inferred on an event-by-event basis.

References

[1] A. Aab et al. [Pierre Auger Coll.], The Pierre Auger Cosmic Ray Observatory, Nucl. Instrum. Meth. A798 (2015) 172.
[2] A. Aab et al. [Pierre Auger Coll.], The flux of ultra-high energy cosmic rays after ten years of operation of the Pierre Auger Observatory, in proceedings of "34th International Cosmic Ray Conference" PoS(ICRC2015)171.
[3] A. Aab et al. [Pierre Auger Coll.], Updates on the neutrino and photon limits from the Pierre Auger Observatory, in proceedings of "34th International Cosmic Ray Conference" PoS(ICRC2015)1103.
[4] A. Aab et al. [Pierre Auger Coll.],Upgrade of the Pierre Auger Observatory (AugerPrime), in proceedings of "34th International Cosmic Ray Conference" PoS(ICRC2015)686.
[5] A. Aab et al. [Pierre Auger Coll.], The Pierre Auger Observatory Upgrade AugerPrime Preliminary Design Report, in arXiv:1604.03637 [astro-ph.IM] (2015).
[6] J. Knapp and D. Heck 1993 *Extensive Air Shower Simulation with CORSIKA, KFZ Karlsruhe* **KfK 5195B**
[7] The Pierre Auger Collaboration, JCAP **1408**, 08, 019 (2014)
[8] R. Pesce for the Pierre Auger Collaboration, in *Proc. of ICRC 2011* (Proceedings, Beijing China, 2011) 0214
[9] D. Ravignani for the Pierre Auger Collaboration, in *Proc. of ICRC 2013* (Proceedings, Rio de Janeiro Brazil, 2013) 0693
[10] The Pierre Auger Collaboration, Astropart. Phys. **34**, 368 (2011)
[11] I. Valiño for the Pierre Auger Collaboration, in *Proc. of ICRC 2015* (Proceedings of Science, The Hague Netherlands, 2015)
[12] Pierre Auger and Telescope Array Collaborations, I.C. Maris et al., in *Proc. of UHECR2014* (Utah USA, 2015)

The Borexino experiment at the INFN CNAF Tier1

A. C. Re
on behalf of the BOREXINO collaboration
Università degli Studi and INFN, Milano, IT
E-mail: alessandra.re@mi.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 and results in a shielding capacity against cosmic rays of 3,800 meter water equivalent (m.w.e.).

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. Other significant publications are about the non-solar neutrino physics and concern the observation of anti-neutrinos from the Earth (the geoneutrinos) and the test of electric charge conservation.

During 2016 the Borexino collaboration carried out studies about the the search for neutrino and antineutrino events correlated with 2350 gamma-ray bursts [2] and about the time periodicities of the ^{7}Be solar neutrino interaction rate [3].

Still in 2016 the Borexino collaboration has devoted great efforts to perform preliminary sensitivity studies and antineutrino Monte Carlo simulations in order to get ready for the SOX project start. Thanks to its unprecedented radiopurity, the Borexino experiment offers in fact a unique opportunity to perform a short-baseline neutrino oscillation study: the SOX (Short distance neutrino Oscillations with boreXino) project. The SOX experiment [4] aims at the complete confirmation or at the 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. In case of a negative result SOX will anyway be able to close a long-standing debate about the reality of the neutrino anomalies, will probe the existence of new physics in low energy neutrino interactions and will provide a measurement of the neutrino magnetic moment. 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 antineutrino source is currently under production in Mayak (Russia) and the SOX experiment is expected to start in spring 2018 and takes 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 6 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 300 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 and/or a massive Monte Carlo generation.

4. Conclusions

During next years the amount of CNAF resources needed and used by the Borexino-SOX collaboration is expected to substantially increase. On the one hand Borexino will continue in its rich solar neutrino program with the ambitious target of CNO neutrino flux measurement; on the other hand it 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. 2017*Astropart. Phys.* **86-01** 11.
[3] Agostini M. et al. 2017 *submitted to Astropart. Phys., available on arXiv: 1701.07970 [hep-ex].*
[4] Bellini G. et al. 2013 *JHEP* **8** 038.

The Cherenkov Telescope Array

L. Arrabito[1] and C. Bigongiari[2]

[1], Laboratoire Univers et Particules de Montpellier, Université de Montpellier, CNRS/IN2P3, Montpellier, FR

[2], INAF Osservatorio Astronomico di Roma, Monte Porzio Catone (RM), IT

E-mail: arrabito@in2p3.fr, ciro.bigongiari@oa-roma.inaf.it

Abstract.

The Cherenkov Telescope Array (CTA) is an ongoing worldwide project to build a new generation ground based observatory for Very High Energy (VHE) gamma-ray astronomy. CTA will feature two arrays of Imaging Atmospheric Cherenkov Telescopes (IACTs), one in each Earth hemisphere, to ensure the full sky coverage and will be operated as an open observatory to maximize its scientific yield. Each array will be composed of tens of IACTs of different sizes to achieve a ten-fold improvement in sensitivity, with respect to current generation facilities, over an unprecedented energy range which extends from a few tens of GeV to a hundred of TeV. Imaging Cherenkov telescopes have already discovered tens of VHE gamma-ray emitters providing plentiful of valuable data and clearly demonstrating the power of the imaging Cherenkov technique. The much higher telescope multiplicity provided by CTA will drive to highly improved angular and energy resolution, which will permit more accurate morphological and spectrographical studies of VHE gamma-ray sources. CTA project combines therefore guaranteed scientific return, in the form of high precision astrophysics, with considerable potential for major discoveries in astrophysics, cosmology and fundamental physics.

1. Introduction

Since the discovery of the first VHE gamma-ray source, the Crab Nebula [1] by the Whipple collaboration in 1989, ground-based gamma-ray astronomy has undergone an impressive development which drove to the discovery of more than 190 gamma-ray sources in less than 30 years [2]. Whenever a new generation of ground-based gamma-ray observatory came into play gamma-ray astronomy experienced a major step in the number of discovered sources as well as in the comprehension of the astrophysical phenomena involved in the emission of VHE gamma radiation. Present generation facilities like H.E.S.S. [3], MAGIC [4] and VERITAS [5] already provided 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 and Active Galactic Nuclei, clearly demonstrating the huge physics potential of this field, which is not restricted to pure astrophysical observations, but allows significant contributions to particle physics and cosmology too, see [6, 7] for recent reviews.

The impressive physics achievements obtained with the present generation instruments as well as the technological developments regarding mirror production and new photon-detectors triggered many projects for a new-generation gamma-ray observatory by groups of astroparticle physicists around the world which later merged to form the CTA consortium [8]. CTA members are carrying on a worldwide effort to provide the scientific community with a state-of-the-art

ground-based gamma-ray observatory, allowing exploration of cosmic radiation in the very high energy range with unprecedented accuracy and sensitivity.

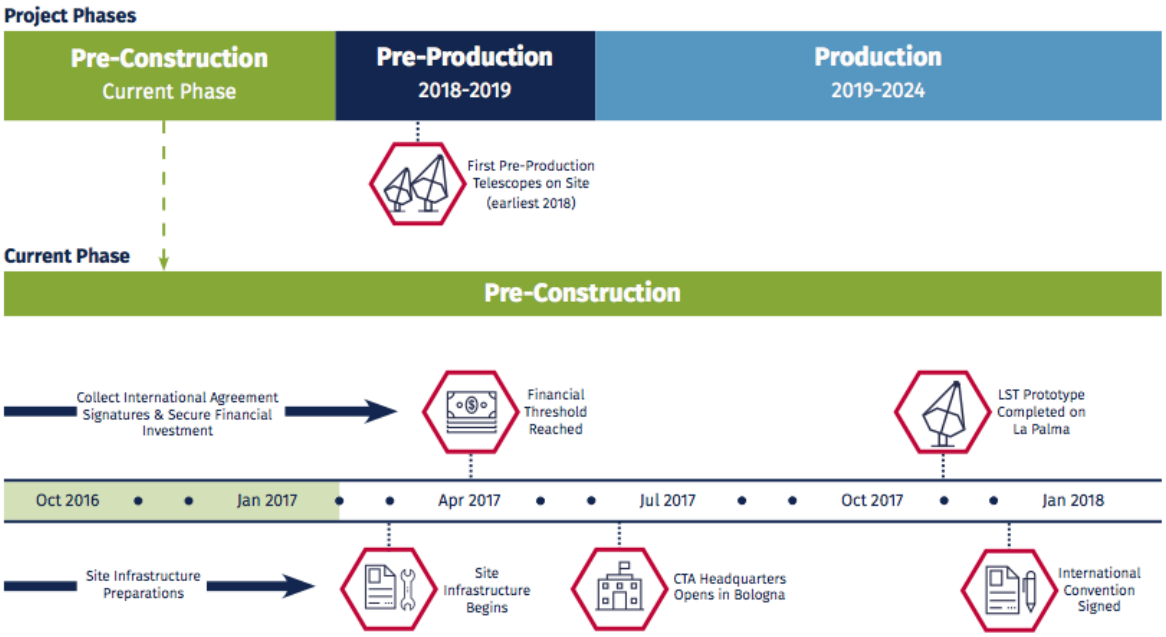


Figure 1. CTA project time line.

VHE 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 achieve a full sky-coverage the CTA observatory will consist of two arrays of IACTs, one in both Earth hemispheres. The northern array will be placed at the Observatorio del Roque de Los Muchachos on La Palma Island, Spain, while the southern array will be located in Chile at the ESO site close to the Cerro Paranal. The two sites were selected after years of careful consideration of extensive studies of the environmental conditions, simulations of the science performance and assessments of construction and operation costs. Each array will be composed by IACTs of different sizes to achieve an overall ten-fold improvement in sensitivity with respect to current IACT arrays while extending the covered energy range from about 20 GeV to about 300 TeV. The southern hemisphere array will feature telescopes of three different sizes to cover the full energy range for a detailed investigation of galactic sources, and in particular of the Galactic center, without neglecting observations of extragalactic objects. The northern

hemisphere array instead will consist of telescopes of two different sizes only covering the low energy end of the above-mentioned range (up to some tens of TeV) and will be dedicated mainly to northern extragalactic objects and cosmology studies.

The CTA observatory with its two arrays will be operated by one single consortium and a significant and increasing fraction of the observation time will be open to the general astrophysical community to maximize CTA scientific return.

The CTA project has entered the pre-construction phase, see figure 1, and the first telescope is being deployed at La Palma site while detailed geophysical characterization of the southern site is ongoing. First data are expected to be delivered in 2017 by some telescope prototypes already deployed at testing sites and hopefully by the telescope under construction at La Palma. CTA Observatory is expected to become fully operational by 2024 but precursors mini-arrays are expected to operate already in 2018.

A detailed description of the project and its expected performance can be found in a dedicated volume of the Astroparticle Physics journal [9], while an update on the project status can be found in [10]. CTA is included in the 2008 roadmap of the European Strategy Forum on Research Infrastructures (ESFRI), is one of the Magnificent Seven of the European strategy for astroparticle physics by ASPERA, and highly ranked in the strategic plan for European astronomy of ASTRONET.

2. Computing Model

In the pre-construction phase the available computing resources are used mainly for the simulation of atmospheric showers and their interaction with the Cherenkov telescopes of the CTA arrays to evaluate the expected performance and optimize many construction parameters. The simulation of the atmospheric shower development, performed with Corsika [11], is followed by the simulation of the detector response with sim_telarray [12], a code developed within the CTA consortium. It is worthwhile to notice that thanks to the very high rejection of hadronic background achieved with the IACT technique, huge samples of simulated hadronic events are needed to achieve statistically significant estimates of the CTA performance. About 10^{11} 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.

Given these large storage and computing requirements, the Grid approach was chosen to pursue this task and a Virtual Organization for CTA was created in 2008 and is presently supported by 19 Grid sites spread over 7 countries, with resources of the order of 12000 of available logical CPUs and more than 3.5 PB of storage.

The CTA production system currently in use [13] is based on the DIRAC framework [14], which has been originally developed to support the production activities of the LHCb (Large Hadron Collider Beauty) experiment and today is extensively used by several particle physics and biology communities. DIRAC offers powerful job submission functionalities and can interface with a palette of heterogeneous resources, such as grid sites, cloud sites, HPC centers, computer clusters and volunteer computing platforms. Moreover, DIRAC provides a layer for interfacing with different types of resources, like computing elements, catalogs or storage systems.

A massive production of simulated data has been carried on in 2016 aimed initially at the array layout optimization and later at the full characterization of the chosen layout. Simulated data have been analyzed with two different analysis chains to crosscheck the results.

About 3.3 million of GRID jobs have been executed in 2016 for such task corresponding to about 126 millions of HS06 hours of CPU power and 1.5 PB of disk space. CNAF contributed to this effort with about 17 millions of HS06 hours and 265 TB of disk space, corresponding to 13% of the overall CPU power used, see figure 2 and the 18% of the disk space, resulting the fourth contributor both in terms of CPU time and storage.

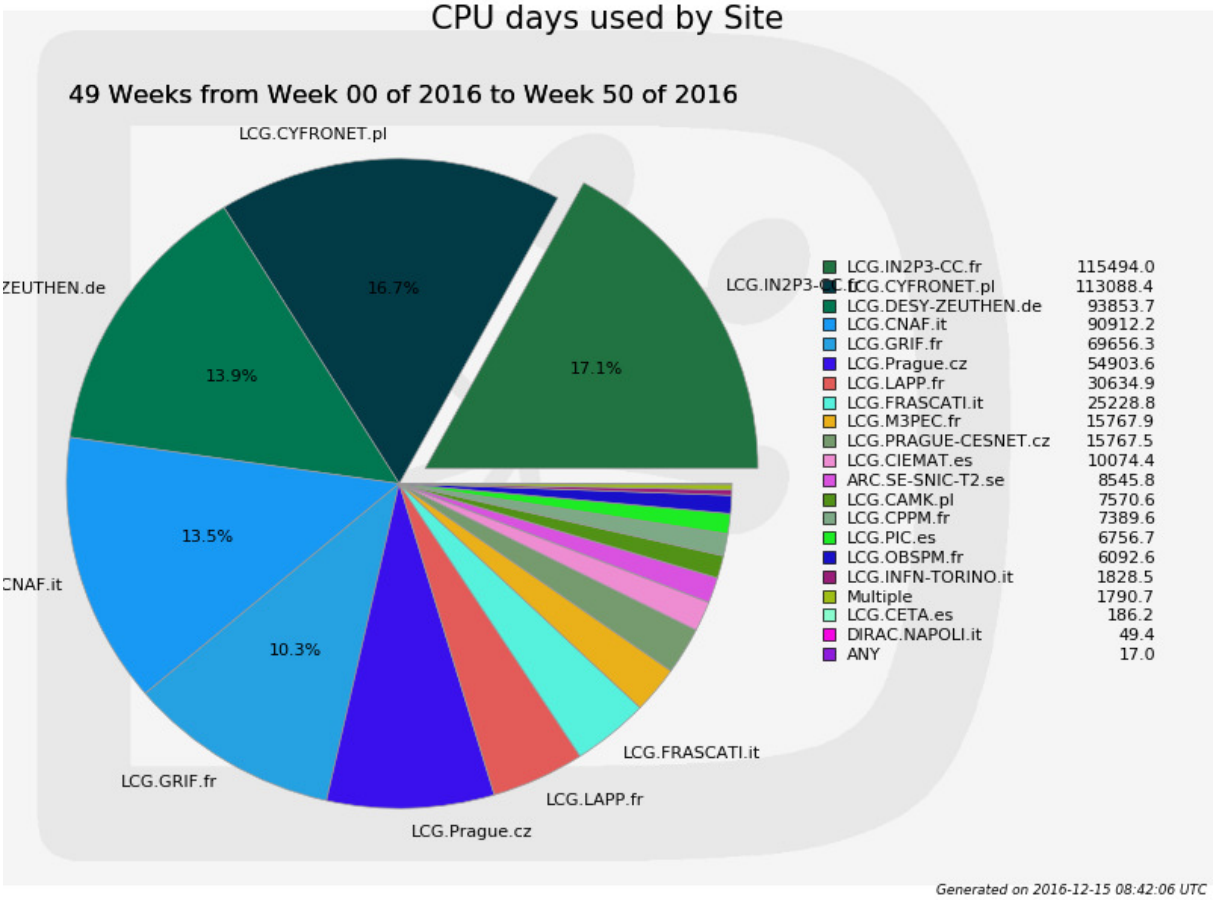


Figure 2. CPU power provided in 2016 by Grid sites in the CTA Virtual Organization.

References

[1] Weekes T C *et al.* 1898 "Observation of TeV gamma rays from the Crab nebula using the atmospheric Cerenkov imaging technique" *ApJ* **342** 379-95

[2] TevCat web page http://tevcat.uchicago.edu

[3] H.E.S.S. web page https://www.mpi-hd.mpg.de/hfm/HESS/

[4] MAGIC web page https://magic.mppmu.mpg.de

[5] VERITAS web page http://veritas.sao.arizona.edu

[6] de Naurois M and Mazin D "Ground-based detectors in very-high-energy gamma-ray astronomy" Comptes Rendus - Physique **16** Issue 6-7, 610-27

[7] Lemoine-Goumard M 2015 "Status of ground-based gamma-ray astronomy" Conf. Proc of 34^{th} International Conference on C, 2015, The Hague, PoS ICRC2015 (2016) 012

[8] CTA web page https://www.cta-observatory.org/about/cta-consortium/

[9] Hinton J, Sarkar S, Torres D and Knapp J 2013 "Seeing the High-Energy Universe with the Cherenkov Telescope Array. The Science Explored with the CTA" *Astropart. Phys.* **43** 1-356

[10] Bigongiari C 2016 "The Cherenkov Telescope Array" Proc. of Cosmic Ray International Seminar (CRIS2015), 2015, Gallipoli, *Nucl. Part. Phys. Proc.* **279281** 174-81

[11] Heck D, Knapp J, Capdevielle J N, Schatz G and Thouw T 1998 "CORSIKA: a Monte Carlo code to simulate extensive air showers" Forschungszentrum Karlsruhe GmbH, Karlsruhe (Germany), Feb 1998, V + 90 p., TIB Hannover, D-30167 Hannover (Germany)

[12] Bernlhör K 2008 "Simulation of imaging atmospheric Cherenkov telescopes with CORSIKA and sim_telarray" *Astropart. Phys* **30** 149-58

[13] Arrabito L, Bregeon J, Haupt A, Graciani Diaz R, Stagni F and Tsaregorodtsev A 2015 "Prototype of a production system for Cherenkov Telescope Array with DIRAC" Proc. of 21^{st} Int. Conf.e on Computing in High Energy and Nuclear Physics (CHEP2015), 2015, Okinawa, *J. Phys.: Conf. Series* **664** 032001

[14] Tsaregorodtsev A *et al.* 2014 "DIRAC Distributed Computing Services" Proc. of 20^{st} Int. Conf.e on Computing in High Energy and Nuclear Physics (CHEP2013) *J. Phys.: Conf. Series* **513** 032096

The CMS Experiment at the INFN CNAF Tier1

T. Boccali
INFN, Pisa, IT

E-mail: Tommaso.Boccali@cern.ch

Abstract. After a brief description of the CMS Computing operations during LHC RunII, the CMS utilization at Tier-1 CNAF is presented, with a focus on recent developments and studies.

1. Introduction

The CMS Experiment at CERN collects and analyses data from the pp collisions in the LHC Collider.

The first physics Run, at centre of mass energy of 7-8TeV, started in late March 2010, and ended in February 2013; more than 25 fb^{-1} of collisions were collected during the Run. RunII, at 13 TeV, is currently ongoing: it started in 2015, and is expected to continue up to the end of 2018.

During the first two years of RunII, LHC has been able to largely exceed its design parameters: already in 2016 instantaneous luminosity reached 1.5 10 34 $cm^{-2}s^{-1}$, 50% more than the planned "high luminosity" LHC phase. The most astonishing achievement, still, is a huge improvement on the fraction of time LHC can serve physics collision, increased form ~35% of RunI to more than 80% in some months on 2016.

The most visible effect, computing wise, is a large increase of data to be stored, processed and analysed offline, with 2016 allowing for the collection of more than 40 fb^{-1} of physics data.

2. RunII computing operations

During RunII, the computing 2004 model designed for RunI has greatly evolved. The MONARC Hierarchical division of sites in Tier0, Tier-1s and Tier-2s, is still present, but less relevant during operations. All simulation, analysis and processing workflows can now be executed at virtually any site, with a full transfer mesh allowing for point-to-point data movement, outside the rigid hierarchy.

Remote access to data, using WAN-aware protocols like XrootD and data federations, are used more and more instead of planned data movement, allowing for an easier exploitation of CPU resources.

Opportunistic computing is becoming a key component, with CMS having explored access to HPC systems, Commercial Clouds, and with the capability of running its workflows on virtually any (sizeable) resource we have access to.

The main problem CMS Computing had to face during 2016 operations is due to the scarcity of storage resources, from two distinct factors:

1. A ~10% under pledge in the deployed resource with respect to the model and to RRB recommendations;

2. The unexpected increase in LHC luminosity and availability.

These two factors have caused a crisis around July, when nearly half of the CMS Tier-1 sites had disk areas full, and with tape storage usage exceeding 90% of the total available.

CMS had to deploy emergency measures in order to keep a regular offline data flow, and to cope with transfers of data to distributed sites:

1. An emergency tape clean-up of up to 30 PB was requested in September;

2. Some trigger streams were no more made available to users in RAW and RECO formats

3. A fine tuning of CERN/EOS was needed, in collaboration with the IT Department, in order not to fill Tier-0 buffers with outgoing data.

The measures were successfully in keeping the system operational, and by September the problems had cleared.

3. CMS WLCG Resources

CMS Computing model has been used to request resources for 2016 RunII data taking, with total requests (Tier-0 + Tier-1s + Tier-2s) exceeding 1400 kHS06, 87 PB on disk, and 140 PB on tape.

Historically, Italy contributed to CMS computing with 12% of the Tier-1 and Tier-2 resources. For 2016 it meant deploying

- 48 kHS06 CPU at CNAF, and 84 kHS06 shared between the 4 Tier-2s;
- 3960 TB of disk at CNAF, and 4560 TB shared between the 4 Tier-2s;
- 12000 TB of tape at CNAF.

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 at moments large CPU over pledge, and in general has consistently resulted as the second Tier-1 in CMS as number of processed hours, after the US Tier-1; the same holds for total number of processed jobs, as shown in Figure 1.

The tape resource has been used at levels exceeding 90% by September 2016, followed by a 4.5 PB deletion. By March 2017, a 90% level was reached again.

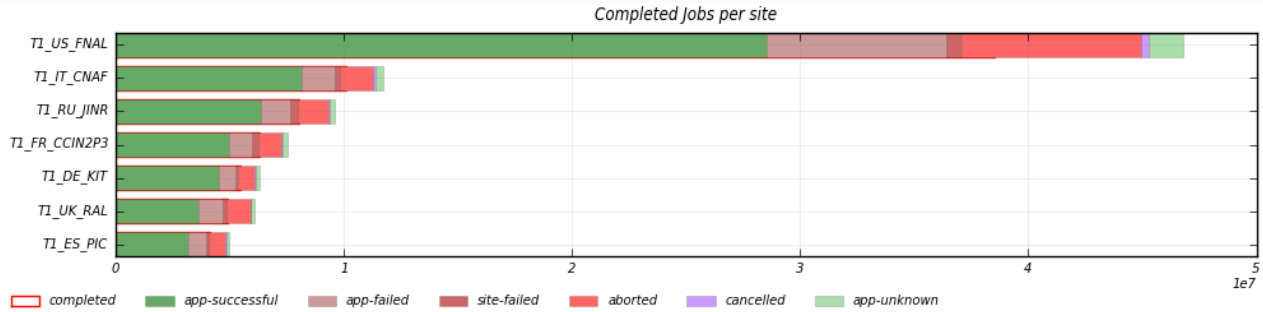


Figure 1. Jobs processed at CMS Tier1s during 2016.

4. Expected resource growth

The LHC collider is planning another 2 years of RunII data taking, at a centre of mass energy of 13 TeV. Year-by-year increases, which would have been large in presence of the reference computing model, have been reduced in the 2017 actual request. Still, CNAF is expected to grow substantially in 2017, with a resource request for 72 kHS06, 6380 TB of Disk, and 21 PB for tape. At the time of writing, pledges by CNAF are lower than the usual 12% share. For what concerns 2018, final numbers are not available at the moment; anyhow, CMS requests should not increase substantially, with CPU remaining stable, and tape increasing by ~10%. Final numbers will be available within months.

5. Activities in opportunistic computing

The CNAF CMS team has been very active throughout 2016 in testing and validating the use of opportunistic resources for CMS.

Two activities have been followed in 2016:

- A test of Aruba Commercial cloud for CMS usage. A few hundred cores have been put in to production, as an elastic extension of the CNAF LSF batch system. An in-house developed tunnelling solution allows the extension of CNAF IPs on remote hosts, limiting the utilization of the tunnel to services. Data access is granted via standard network path, utilizing XrootD remote access via the CMS federation. Results have been reported at major computing conferences [1, 2, 3].
- A test using a Microsoft Grant obtained by CNAF (20k$), lasting till July 2017.

In general, CNAF and CMS are very interested in tests and strategies usable in order to expand elastically the Tier-1 site to external resources.

A second expected phase is planned using the same remote resources, but using a less intrusive technical solution which avoids tunnels; it will be reported in next year's report.

6. Conclusions

The CMS Collaboration expressed in many occasions its praise to the CMS CNAF Tier1, which has consistently been in the top 2 Tier1 sites for resource utilization, resource pledges and availability. CNAF represents an important asset for the CMS Collaboration, and all the expectations are towards an even greater role inside the CMS Computing. The studies CNAF is carrying out for the utilization of new technologies are of utmost importance for CMS, since the utilization of opportunistic resources will be more and more relevant for LHC Experiments.

References

[1] Elastic CNAF DataCenter extension via opportunistic resources. PoS(ISGC 2016)031, V. Ciaschini, S. Dal Pra, L. dell'Agnello, A. Chierici, D. de Girolamo, V. Sapunenko, T. Boccali, A. Italiano.

[2] Extending the farm on external sites: the INFNTier-1 experience. Presented at CHEP 2016, A. Chierici, D. Cesini, L. dell'Agnello, S. Zani, T. Boccali, V. Sapunenko.

[3] Dynfarm: A Dynamic Site Extension, accepted for publication in CHEP2016 Proceeding, V. Ciaschini, D. de Girolamo.

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 presently in pre-operation at the Laboratori Nazionali del Gran Sasso of INFN, in Italy. 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 has been completed in 2016.

1. The experiment

The main goal of the CUORE experiment [1] is to search for Majorana neutrinos through the neutrinoless double beta decay (0νDBD): $(A, Z) \rightarrow (A, Z + 2) + 2e^-$. The 0νDBD has never been observed so far and its half life is expected to be higher than 10^{25} y. CUORE searches for 0νDBD in a particular isotope of Tellurium (^{130}Te), using thermal detectors (bolometers). A thermal detector is a sensitive calorimeter which measures the energy deposited by a single interacting particle through the temperature rise induced in the calorimeter itself. This is accomplished by using suitable materials for the detector (dielectric crystals) and by running it at very low temperatures (in the 10 mK range) in a dilution refrigerator. In such condition a small energy release in the crystal results in a measurable temperature rise. The temperature change is measured by means of a proper thermal sensor, a NTD germanium thermistor glued onto the crystal. The bolometers act at the same time as source and detectors for the sought signal. The CUORE detector is an array of 988 $^{nat}TeO_2$ bolometers, for 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. CUORE will reach a sensitivity on the ^{130}Te 0νDBD half life of 10^{26} y, thus starting to cover the inverted neutrino mass hierarchy region. During the 2016 the installation of the readout electronics and data acquisition system has been completed and the bolometer towers were installed in the cryostat. The cool down started in December 2016 and the CUORE experiment is currently in pre-operation phase.

2. CUORE computing model and the role of CNAF

The CUORE raw data consist in Root files containing the continuous data stream of $\sim$ 1000 channels recorded by the DAQ at sampling frequencies of 1 kHz. Triggers are implemented via software and saved in a custom format based on the ROOT data analysis framework. 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 about 20 TB/y of raw data are expected. A full copy of data is maintained at CNAF and preserved also on tape.

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. 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 main instance of the CUORE database is located on a computing cluster at the Laboratori Nazionali del Gran Sasso and a replica is synchronized at CNAF. The full analysis framework at CNAF is working and kept up to date to official CUORE softwre release.

In 2016 the CNAF resources have been extensively used for Monte Carlo simulations, both for CUORE-0 and CUORE. The code is based on the GEANT4 package, for which the 4.9.6 and the 10.xx up to 10.03 releases have been installed. Monte Carlo simulations are mainly used to evaluate the experimental background. For CUORE-0, high statistics simulations have been run for specific cases, to be used to model the background measured by the experiment. The result of such work was the measurement of the 2-neutrino half-life for the ^{130}Te, with an increased sensitivity with respect to previous results [2]. For CUORE an extensive use of CNAF was done to run Monte Carlo simulations for each possible source expected to contribute to the background in the Region of Interest for the 0νDBD decay. The goal of this work is the evaluation, at the present knowledge of material contaminations, of the background index reachable in the ROI by the experiment. Depending on the specific efficiency of the simulated radioactive sources (sources located outside the lead shielding are really inefficient), the Monte Carlo simulation could exploit from 5 to 500 computing nodes, with durations up to some weeks.

Actually the stored amount of simulated and real data is of about 3.6 Tb.

References

[1] Artusa D *et al.* (CUORE) 2015 *Adv.High Energy Phys.* **2015** 879871 (*Preprint* 1402.6072)
[2] Alduino C *et al.* (CUORE) 2017 *Eur. Phys. J.* **C77** 13 (*Preprint* 1609.01666)

CUPID-0 experiment

CUPID-0 collaboration

E-mail: stefano.pirro@lngs.infn.it

Abstract. With their excellent energy resolution, efficiency, and intrinsic radio-purity, cryogenic calorimeters are primed for the search of neutrino-less double beta decay (0νDBD). The sensitivity of these devices could be further increased by discriminating the dominant alpha background from the expected beta like signal. The CUPID-0 collaboration aims at demonstrating that the measurement of the scintillation light produced by the absorber crystals allows for particle identification and, thus, for a complete rejection of the alpha background. The CUPID-0 detector, assembled in 2016 and now in commissioning, consists of 26 $Zn^{82}Se$ scintillating calorimeters corresponding to about 2×10^{25} 0νDBD emitters.

The configuration of the CUPID-0 data processing environment on the CNAF computing cluster is almost complete, and a more intense use of resources is expected in 2017.

1. The experiment

Neutrino-less Double Beta Decay (0νDBD) is a hypothesized nuclear transition in which a nucleus decays emitting only two electrons. This process can not be accommodated in the Standard Model, as the absence of emitted neutrinos would violate the lepton number conservation. Among the several experimental approaches proposed for the search of 0νDBD, cryogenic calorimeters (bolometers) stand out for the possibility of achieving excellent energy resolution ($\sim$0.1%), efficiency ($\geq$80%) and intrinsic radio-purity. Moreover, the crystals that are operated as bolometers can be grown starting from most of the 0νDBD emitters, enabling the test of different nuclei. The state of the art of the bolometric technique is represented by CUORE, an experiment composed by 988 bolometers for a total mass of 741 kg, presently in detector commissioning phase at Laboratori Nazionali del Gran Sasso. The ultimate limit of the CUORE background suppression resides in the presence of α-decaying isotopes located in the detector structure. The CUPID-0 project [1] was born to overcome the actual limits. The main breakthrough of CUPID-0 is the addition of independent devices to measure the light signals emitted from scintillation in ZnSe bolometers.The different properties of the light emission of electrons and α particles will enable event-by-event rejection of α interactions, suppressing the overall background in the region of interest for 0νDBD of at least one order of magnitude. The detector is composed by 26 ZnSe ultrapure $\sim$ 500g bolometers, enriched at 95% in ^{82}Se, the 0νDBD emitter, and faced to Ge disks light detector operated as bolometers. CUPID-0 is hosted in a dilution refrigerator at the Laboratori Nazionali del Gran Sasso and started the commissioning phase in Jan 2017. In 2016 the first three enriched bolometers were tested, proving that their performance in terms of energy resolution, background rejection capability and intrinsic radio-purity complies with the requirements of CUPID-0 [2]

2. CUPID-0 computing model and the role of CNAF

The CUPID-0 computing model is similar to the CUORE one, being the only difference in the sampling frequency and working point of the light detector bolometers. The complete data stream is saved in root files.Trigger is software generated. 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 arranged in runs, each run lasting about one day. Raw data are transferred from the DAQ computers to the permanent storage area (located at CNAF) at the end of each run. A full copy of data is preserved also on tape.

The 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. 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 main instance of the database is located at CNAF and the full analysis framework is installed and almost ready for being used. A web page for the offline reconstruction monitoring is also maintained.

During 2017 a more intense usage of the CNAF resources is expected, both in terms of computing resourced and storage space.

References

[1] Beeman J W *et al.* 2016 *J. Low. Temp. Phys.* **184** 852–858
[2] Beeman J W *et al.* (LUCIFER) 2015 *Eur. Phys. J.* **C75** 591 (*Preprint* 1508.01709)

DAMPE data processing and analysis at CNAF

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

[1] INFN, Perugia, IT
[2] Università di Sassari, Sassari, IT
[3] ASDC, Roma, IT
[4] INFN-LNS, Catania, IT
[5] INFN, Bari, IT
[6] University of Geneva, Departement de physique nuclaire et corpusculaire (DPNC), Genève, CH
DAMPE experiment `http://dpnc.unige.ch/dampe/`, `http://dampe.pg.infn.it`

E-mail: * `ddurso@uniss.it, 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. CNAF computing center is the mirror of DAMPE data outside China and the main data center for Monte Carlo production. It also supports user data analysis of the Italian DAMPE Collaboration.

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 one double layer of scintillating plastic strips detector, that serve as anti-coincidence detector and to measure particle charge, 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 position on 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 3, 4 and 5 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 $\times$ 19.5 cm^2 large, each read out by a photomultiplier. 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

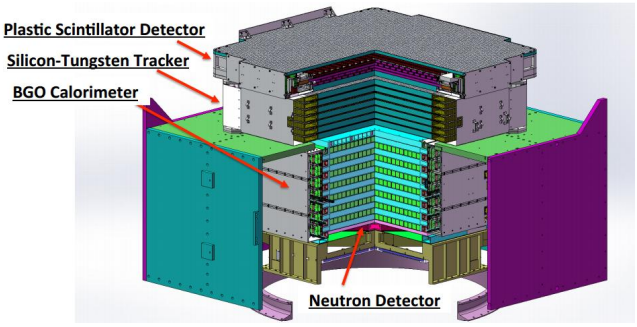


Figure 1. DAMPE telescope scheme: a double layer of the plastic scintillator strip detector (PSD); the silicon-tungsten tracker-converter (STK) made of 6 tracking double layers; the imaging calorimeter with about 31 radiation lengths thickness, made of 14 layers of Bismuth Germanium Oxide (BGO) bars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

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 heavier 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 proton and heavy ions 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 heavier 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. On the European side, the DAMPE collaboration consists of research groups from INFN and University of Perugia, Lecce and Bari, and from the Department of Particle and Nuclear Physics (DPNC) at the University of Geneva in Switzerland.

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, $\sim$ 7 GB/day), and a second stream that include the orbital and slow control information (1F data stream, $\sim$ 7 GB/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, $\sim$ 85 GB/day) is produced. The total amount of data volume produced per day is $\sim$ 100 GB. Data processing and reconstruction activities are currently supported by a computing farm consisting of more then 1400 computing cores, able to reprocess 3 years DAMPE data in 1 month.

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] was 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 run on the selected computing farm. The design and implementation of this workflow system was heavily influenced by the implementation of the Fermi-LAT data processing pipeline [4] and the DIRAC computing framework [5].

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 input 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. Once generated, a secondary job is initiated which performs digitization and reconstruction of existing MC data with a given release for large amounts of MC data in bulk. This process is set-up via a cron-job at DPNC and occupies up to 200 slots in a 6-hour limited computing queue.

2.3. Data availability

DAMPE data are available to the Chinese Collaboration through the PMO institute, while they are kept accessible to the European Collaboration transferring them from PMO to CNAF and from there to the DPNC. Every time a new 1B, 1F or 2A data files are available at PMO, they are copied, using the gridFTP [6] protocol, to a server at CNAF, gridftp-plain-virgo.cr.cnaf.infn.it, into the DAMPE storage area. From CNAF, every 4 hours a copy of each stream is triggered to the Geneva computing farm via rsync. 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. Data verification and copy processes are managed through a dedicated user interface (UI), ui-dampe.

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 maximum bandwidth of 100 Mb/s. So the PMO-CNAF copy processed is used for daily data production.

To transfer towards Europe data in case of DAMPE data re-processing and to share in China Monte Carlo generated in Europe, a dedicated DAMPE server has been installed at the institute for high energy physics, IHEP, in Beijing which is connected to CERNT with a 1Gb/s bandwidth. Data synchronization between this server and PMO is done by manually induced hard-drive exchange.

To simplify user data access overall the Europe, an XRootd federation has been implemented and is currently under test at CNAF and at the RECAS farm in Bari: an XRootd redirector has been set up in Bari and an XRootd server has been installed at CNAF to allow users to read and write from/to the endpoints of the federation.

3. CNAF contribution

The CNAF computing center is the mirror of DAMPE data outside China and the main data center for Monte Carlo production. In 2016, a dedicated user interface, ui-dampe, 100 TB of disk space and 3000 HS06 of CPU time have been allocated for the DAMPE activities. Furthermore from September to November additional 6000 HS06 have been available for the DAMPE Monte

Carlo production. On the base of what implemented for AMS02, an integrated solution for DAMPE data access among CNAF, the ASI Science Data Center (ASDC) [10] in Rome and Perugia, have been set, based on the use of the General Parallel File System (GPFS).

4. Activities in 2016

DAMPE activities at CNAF in 2016 have been mainly related to data transfer and Monte Carlo production.

4.1. Data transfer

A daily activity of data transfer from PMO to CNAF and thereafter from CNAF to GVA have been performed all along the year. Daily transfer rate has been of about 100 GB per day from PMO to CNAF more 100 GB from CNAF to PMO. The step between PMO and CNAF is performed, as seen in previous sections, via gridftp protocol. Two strategies have been, instead, used to copy data from CNAF to PMO: via rsync from the UI and via rsync managed by lsf jobs.

DAMPE data have been reprocessed three times along the year and a dedicated copy task has been fulfilled to copy the new production releases, in addition to the ordinary daily copy.

4.2. Monte Carlo Production

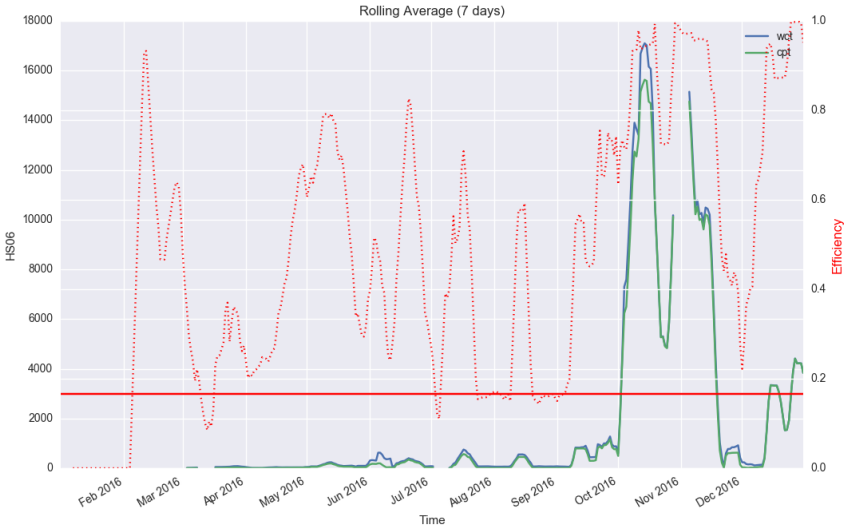


Figure 2. CPU time consumption, in terms of HS06, CPU Time in green (cpt) and Wall Clock Time in blue (wct) as a function of time. Job efficiency is overimposed in red.

As main data center for Monte Carlo production, CNAF has been strongly involved in the Monte Carlo campaign. Dedicated LSF profiles have been implemented to allow DAMPE Collaboration to submit multi-thread simulation jobs. For a couple of months, CNAF supported the effort of the Monte Carlo campaign with an extra pledged of 6000 HS06. At CNAF almost 300 thousands jobs have been executed for a total of about 3 billions of Monte Carlo events. In the figure 2, the CPU time consumption, in terms of HS06, is shown as function of time:

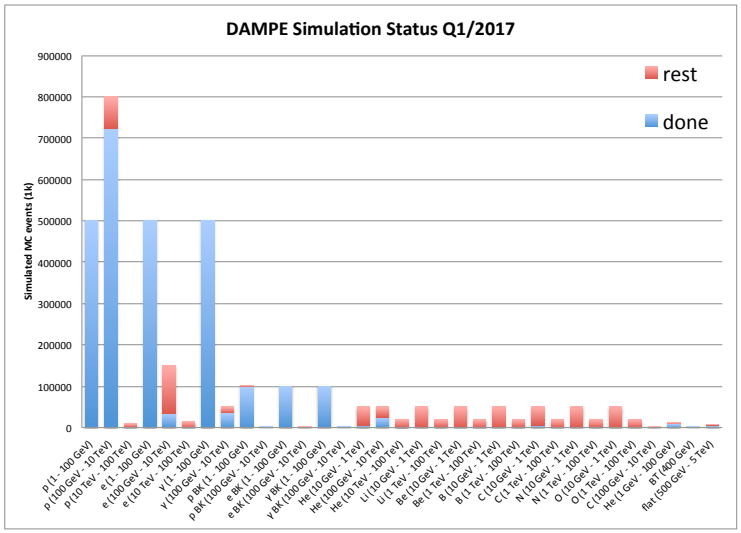


Figure 3. Status of simulation production: completed samples (blue) and samples to be completed (red).

in green the CPU Time (cpt) and in blue the Wall Clock Time (wct); in red job efficiency is overimposed.

In figure 3 is shown the status of simulation production: blue samples completed, red samples still to be completed.

Our Monte Carlo simulation efforts are still continuing based on plans outlined in 2016. In figure 3 is shown the status of the overall simulation production as defined in 2016 for DAMPE. The majority of these samples are heavy ion MCs at very high energies which are expected to be particularly CPU intense.

5. Acknowledgments

The DAMPE mission was founded by the strategic priority science and technology projects in space science of the Chinese Academy of Sciences and in part by National Key Program for Research and Development, and the 100 Talents program of the Chinese Academy of Sciences. In Europe the work is supported by the Italian National Institute for Nuclear Physics (INFN), the Italian University and Research Ministry (MIUR), and the University of Geneva. We extend our gratitude to CNAF-T1 for their continued support also beyond providing computing resources.

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
[10] http://www.asdc.asi.it

DarkSide-50 experiment at CNAF

S. Bussino, S. M. Mari, S. Sanfilippo

INFN and Università degli Studi Roma 3, Roma, IT

E-mail: bussino@fis.uniroma3.it; stefanomaria.mari@uniroma3.it; simone.sanfilippo@roma3.infn.it

Abstract.

DarkSide is a direct dark matter search program operating in the underground Laboratori Nazionali del Gran Sasso (*LNGS*) and it is searching for the rare nuclear recoils (possibly) induced by the so called Weakly Interacting Massive Particles (*WIMPs*). It is based on a Time Projection Chamber with liquid Argon (*LAr-TPC*) from underground sources, with a (46.4 ± 0.7)kg active mass. Moreover the DarkSide-50 (*DS-50*) LAr-TPC is installed inside a 30 t organic liquid scintillator neutron veto, which is in turn installed at the center of a 1kt water Cherenkov veto for the residual flux of cosmic muons. DS-50 has been taking data since November 2013 with Atmospheric Argon (*AAr*) and, since April 2015, it is still working with Underground Argon (*UAr*) highly depleted in radioactive ^{39}Ar. The exposure of 1422 kg d of AAr has demonstrated that the operation of DS-50 for three years in a background free condition is a solid reality, thank to the excellent performance of the pulse shape analysis. The first release of results from an exposure of 2616 kg d of UAr has shown no dark matter candidate events. This is the most sensitive dark matter search performed with an Argon-based detector, corresponding to a 90% CL upper limit on the WIMP-nucleon spin-indipendent cross section of $2 \times 10^{-44} cm^2$ for a WIMP mass of 100 GeV/c^2.

1. The DarkSide-50 experiment

The existence of dark matter is now established from different gravitational effects, but its nature is still a deep mystery. One possibility, motivated by other considerations in elementary particle physics, is that dark matter consists of new undiscovered elementary particles. A leading candidate explanation, motivated by supersymmetry theory (*SUSY*), is that dark matter is composed of as-yet undiscovered Weakly Interacting Massive Particles (*WIMPs*) formed in the early universe and subsequently gravitationally clustered in association with baryonic matter [1]. Evidence for new particles that could constitute WIMP dark matter may come from upcoming experiments at the Large Hadron Collider (*LHC*) at CERN or from sensitive astronomical instruments that detect radiation produced by WIMP-WIMP annihilations in galaxy halos. The thermal motion of the WIMPs comprising the dark matter halo surrounding the galaxy and the Earth should result in WIMP-nuclear collisions of sufficient energy to be observable by sensitive laboratory apparatus. WIMPs could in principle be detected in terrestrial experiments through their collisions with ordinary nuclei, giving observable low-energy <100 keV nuclear recoils. The predicted low collision rates require ultra-low background detectors with large (0.1-10 ton) target masses, located in deep underground sites to eliminate neutron background from cosmic ray muons. The DarkSide experiment is the first to employ a Liquid Argon Time Projection Chamber (*LAr-TPC*) with low levels of ^{39}Ar, together with innovations in photon

detection and background suppression.

The DarkSide-50 detector is installed in Hall C at Laboratori Nazionali del Gran Sasso (*LNGS*) at a depth of 3800 m.w.e., and it will continue to taking data up to the end of 2017. The project will continue with DarkSide-20k (*DS-20k*) and *Argo*, a multi-ton detector with an expected sensitivity improvement of two orders of magnitude.

DS-50 and DS-20k detectors are based on the same detection principles. The target volume is hosted in a dual phase TPC that contains Argon in both phases, liquid and gaseous, the latter on the top of the former one. The scattering of WIMPs or background particles in the active volume induces a prompt scintillation light, called S1, and ionization. Electrons which not recombine are drifted by an electric field of 200 V/cm applied along the z-axis. They are then extracted into gaseous phase above the extraction grid, and accelerated by an electric field of about 4200 V/cm. Here a secondary larger signal due to electroluminescence takes place, the so called S2. The light is collected by two arrays of 19 3"-PMTs on each side of the TPC corresponding to a 60% geometrical coverage of the end plates and 20% of the total TPC surface. The detector is capable of reconstructing the position of the interaction in 3D. The z-coordinate, in particular, is easily computed by the electron drift time, while the time profile of the S2 light collected by the top plate PMTs allows to reconstruct the x and the y coordinates. The LAr-TPC can exploit Pulse Shape Discrimination (*PSD*) and the ratio of scintillation to ionization (S1/S2) to reject β/γ background in favor of the nuclear recoil events expected from WIMP scattering [4, 5].

Events due to neutrons from cosmogenic sources and from radioactive contamination in the detector components, which also produces nuclear recoils, are suppressed by the combined action of the neutron and cosmic rays vetoes. The first one in particular is a 4.0 meter-diameter stainless steel sphere filled with 30 t of borated liquid scintillator act as Liquid Scintillator Veto (*LSV*). The sphere is lined with *Lumirror* reflecting foils and it is equipped with an array of 110 Hamamatsu 8"-PMTs with low-radioactive components and high-quantum-efficiency photocathodes. The cosmic rays veto, on the other hand, is an 11m-diameter, 10 m-high cylindrical tank filled with high purity water act as a Water Cherenkov Detector (*WCD*). The inside surface of the tank is covered with a laminated *Tyvek-polyethylene-Tyvek* reflector and it is equipped with an array of 80 ETL 8"-PMTs with low-radioactive components and high-quantum-efficiency photocathodes.

The exposure of 1422 kg d of AAr has demonstrated that the operation of DS-50 for three years in a background free condition is a solid reality, thank to the excellent performance of the pulse shape analysis. The first release of results from an exposure of 2616 kg d of UAr has shown no dark matter candidate events. This is the most sensitive dark matter search performed with an Argon-based detector, corresponding to a 90% CL upper limit on the WIMP-nucleon spin-indipendent cross section of $2 \times 10^{-44} cm^2$ for a WIMP mass of 100 GeV/c^2 [6].

2. DarkSide-50 at CNAF

The data readout in the three detector subsystems is managed by dedicated trigger boards: each subsystem is equipped with an user-customizable FPGA unit, in which the trigger logic is implemented. The inputs and outputs from the different trigger modules are processed by a set of electrical-to-optical converters and the communication between the subsystems uses dedicated optical links. To keep the TPC and the Veto readouts aligned, a pulse per second (*PPS*) generated by a GPS receiver is sent to the two systems, where it is acquired and interpolated with a resolution of 20 ns to allow offline confirmation of event matching.

To acquire data, the DarkSide detector uses a DAQ machine equipped with a storage buffer of 7 TB. Data are processed and send within the raw ones to CNAF servers. At CNAF data are housed on a server disk of about 700 TB capacity and on a 300 TB tape for backup purposes. Raw data from CNAF, and processed ones from LNGS are then transferred to Fermi National

Laboratories Grid (*FNAL*). At FNAL data are reprocessed and stored on two different servers: a 32 cores, 64 GB RAM for a total capacity of 70 TB, and a second one of 4 cores and 12 GB RAM deputed to provide several services: home space, web server hosting, repository for code (data processing and Monte Carlo) and documents, run database and the DarkSide wiki. Data processed at FNAL, are transferred back to CNAF servers for further analysis and other manipulations. The INFN Roma 3 group has an active role to maintain and follow, step by step, the overall transferring procedure and to arrange the data management.

3. The future of DarkSide

The DarkSide program will continue with DarkSide-20k and Argo, with 30 t and 300 t of liquid Argon respectively. While Argo is currently being designed, DS-20k is expected to taking data in 2020. The optical sensors will be Silicon Photon Multiplier (*SiPM*) matrices with very low radioactivity. However, the goal of DS-20k is a background free exposure of 100 ton-year of liquid Argon which requires further suppression of ^{39}Ar background with respect to DS-50. The project *URANIA* involves the upgrade of the UAr extraction plant to a massive production rate suitable for multi-ton detectors. The project *ARIA* instead involves the construction of a very tall cryogenic distillation column in the Seruci mine (Sardinia, Italy) with the high-volume capability of chemical and isotopic purification of UAr.
The projected sensitivity of DS-20k and Argo reaches a WIMP-nucleon cross section of $10^{-47}\ cm^2$ and $10^{-48}\ cm^2$ respectively, for a WIMP mass of 100 GeV/cm^2, exploring the region of the parameters plane down to the irreducible background due to atmospheric neutrinos.

References

[1] M. W. Goodman, E. Witten, Phys. Rev. D **31** 3059 (1985);
[2] H. H. Loosli, Earth Plan. Sci. Lett. **63** 51 (1983);
[3] P. Benetti et al. (WARP Collaboration), Nucl. Inst. Meth. A **574** 83 (2007);
[4] P. Benetti et al. (WARP Collaboration), Astropart. Phys. **28** 495 (2008);
[5] M. G. Boulay, A. Hime, Astropart. Phys. **25** 179 (2006);
[6] D. D'Angelo et al. (DARKSIDE Collaboration), Il nuovo cimento C **39** 312 (2016).

The *EEE Project* activity at CNAF

C. Aiftimiei, E. Fattibene, A. Ferraro, B. Martelli, D. Michelotto, F. Noferini, M. Panella and C. Vistoli

INFN-CNAF, Bologna, IT

E-mail: andrea.ferraro@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. This project requires a peculiar data management infrastructure to collect data registered in stations very far from each other and to allow a coordinated analysis. Such an infrastructure is realized at INFN-CNAF, which operates a Cloud facility based on the OpenStack opensource Cloud framework and provides Infrastructure as a Service (IaaS) for its users. In 2014 EEE started to use it for collecting, monitoring and reconstructing the data acquired in all the EEE stations. For the synchronization between the stations and the INFN-CNAF infrastructure we used 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. We present the architecture and the functionalities of this data management system that provides a flexible environment for the specific needs of the EEE project.

1. Introduction

One of the main goal of the EEE Project is to involve young students in a high-level scientific enterprise. Therefore the setup of the experiment is very peculiar and requires new solutions for the data management. For this pourpose the EEE Project joined the CNAF cloud facility in 2014 to create its own data collection center. In fact the CNAF cloud provides a flexible environment based on OpenStack [1] opensource Cloud framework which allows to allocate on demand resources adapted to the need of the experiment and to collect data from the telescopes which are distributed in a wide territory. In the CNAF cloud infrastructure a project (tenant) was provided to deploy all the virtual services requested by the EEE experiment.

2. Data Transfers

After a pilot run in 2014, the EEE project performed a first global run, Run-1, involving 35 schools in a coordinated data aquisition. After the success of Run-1 EEE planned to perform one coordinated run each year during the period of the courses in high schools. During all the runs[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

[1] Pilot run from 27-10-2014 to 14-11-2014, Run-1 from 02-03-2015 to 30-04-2015, Run-2 from 06-11-2016 to 20-05-2016 and Run-3 started on 01-11-2016.

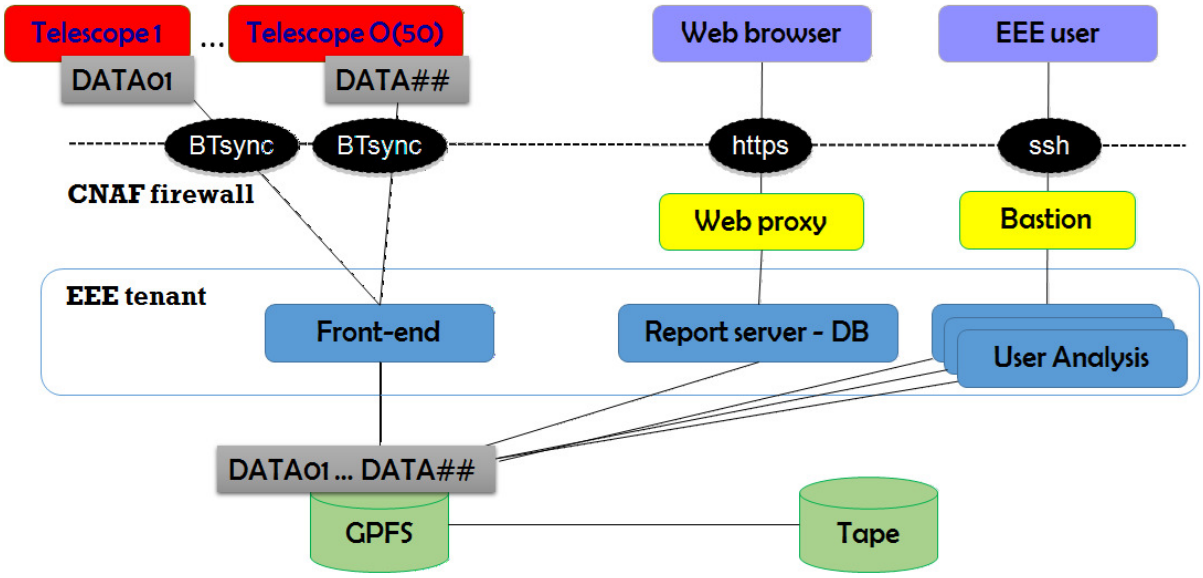


Figure 1. Architecture of the EEE tenant at CNAF.

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.

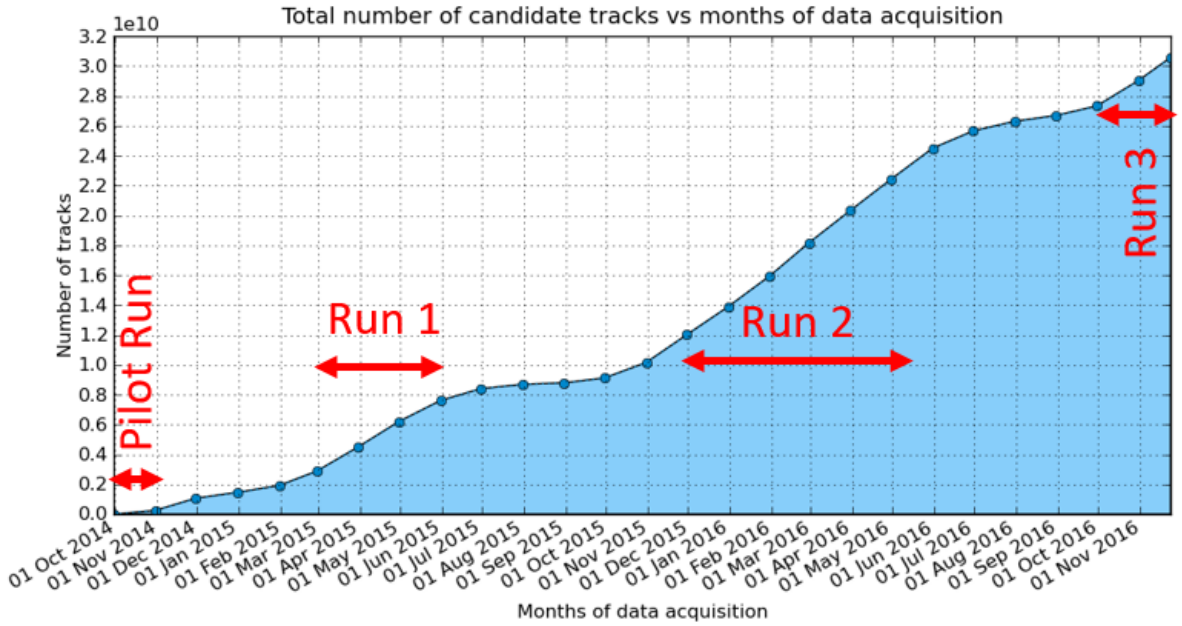


Figure 2. Statistics for the EEE Runs in 2015 and 2016. The cumulative number of collected tracks as function of time is reported. Different coordinated run periods are reported as well.

In the Run-2 perdiod we collected about 14 billion tracks produced py cosmic ray showers, corresponding to 5 TB data transferred at CNAF. The total amount of data collected in Run-2 and previous runs results in about 10 TB and the total number of tracks collected so far is

greater than 30 billion. In Fig. 2 a summary of the data flow performances during Run-2 is reported.

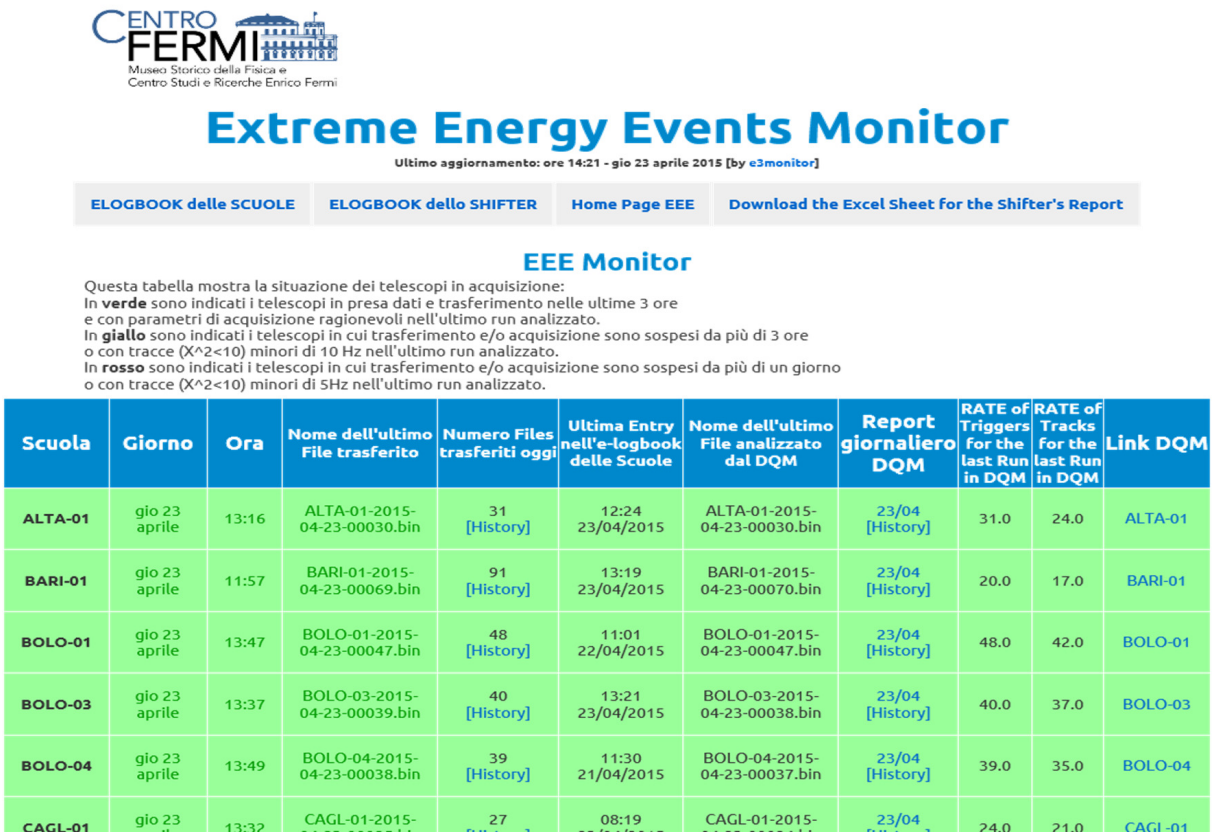


Figure 3. A screenshot of the EEE monitor page [4]. Data Quality Mointor (DQM) plots are provided in real time as well the status of the connection of each school.

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 well consolidated 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.).

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.

ENUBET at CNAF

A. Longhin
INFN, Padova, IT
E-mail: andrea.longhin@pd.infn.it

Abstract. The challenges of precision neutrino physics require measurements of absolute neutrino cross sections at the GeV scale with exquisite (1%) precision. This precision is presently limited to by the uncertainties on neutrino flux at the source. A reduction of this uncertainty by one order of magnitude can be achieved monitoring the positron production in the decay tunnel originating from the K_{e3} decays of charged kaons in a sign and momentum selected narrow band beam. This novel technique enables the measurement of the most relevant cross-sections for CP violation (ν_e and $\bar{\nu}_e$) with a precision of 1% and requires a special instrumented beam-line. Such non-conventional beam-line will be developed in the framework of the ENUBET Horizon-2020 Consolidator Grant, recently approved by the European Research Council. We present the Project, the first experimental results on ultra-compact calorimeters that can embedded in the instrumented decay tunnel and the advances on the simulation of the beamline. A rich program of detector R&D and accelerator-related activities are planned in 2016-2021. CNAF support for ENUBET has just started. We will summarize the current usage of the CNAF resources and perspectives.

1. Conventional and monitored neutrino beams

In conventional beams the prediction of the neutrino flux is based on a full simulation of the beamline considering proton-target interactions, the reinteraction of secondaries, their tracking and decay. The simulation is constrained by beam monitoring devices and ancillary measurements (proton intensity, horn currents, beam-target misalignment etc.) while dedicated hadro-production experiments provide the particle yields from the target. In spite of these constraints, the flux prediction remains heavily dependent on ab-initio simulations leading to large systematics ($\mathcal{O}$(7-10%)). In conventional beams the decay tunnels are almost completely passive regions.

A very precise measurement of the ν_e ($\bar{\nu}_e$) flux can be achieved by directly monitoring the production of $e^{+(-)}$ in the decay tunnel from K_{e3} decays ($K^{+(-)} \to \pi^0 e^{+(-)} \nu_e(\bar{\nu}_e)$) in a sign and momentum-selected narrow band beam[2]. The positron rate is a direct handle to measure the flux of ν_e since these quantities are directly connected through the K decay kinematics and the geometries of the neutrino and positron detectors. The positron rate is monitored in real time but, unlike "tagged ν beams" proposed since the 60's, leptons are not associated to the observed ν on an event-by-event basis using timing coincidences. The required time resolutions for this application are of $\sim$ 10 ns thus well within reach of current technologies.

2. The ENUBET project

The positron monitoring approach will be tested in a conclusive manner by ENUBET (*Enhanced NeUtrino BEams from kaon Tagging*)[3]. The project is aimed at the design and construction

of a detector capable of performing positron identification in K_{e3} decays, while operating in the harsh environment of a ν beam decay tunnel. The project will address all accelerator challenges of kaon tagged beams and study the precise layout of the K/π focusing and transport system. ENUBET has been approved by the ERC (Consolidator Grant, P.I. A. Longhin) for a 5 y duration (since 1 June 2016) and a 2.0 MEUR budget. The ENUBET technology is very well suited for short baseline experiments and might enable a new generation of ν cross section experiments where the ν_e source could be controlled at the 1% level. It could also be exploited for sterile-ν experiments, especially if present anomalies would be soon confirmed. In addition, ENUBET is an important step towards a "time-tagged ν beam" where ν_e interactions could be time correlated with an e^+ in the decay tunnel. In the ENUBET reference design (see Fig. 1, left), the positron tagger is a hollow cylinder surrounding a fraction of the decay tunnel. It is composed of a calorimetric section for e^+ tagging and $\pi^\pm$ rejection (in red) and a system of light tracking devices in the innermost region to reject showers from γ conversions (photon veto, yellow). The secondary beam is designed to have an average momentum of 8.5 GeV and a $\pm$ 20% momentum bite. The momentum of the hadron beam and the length of the decay tunnel are chosen to:

(i) reduce the fraction of K decaying in the transfer line;
(ii) allow for e^+ identification though purely calorimetric techniques;
(iii) match the ν spectrum of interest for future long baseline experiments;
(iv) maximize the ν_e flux from K and reduce the component of ν_e from μ decays;
(v) allow for a small emittance of the entering beam (few mrad over 10×10 cm^2) to prevent undecayed secondaries or μ from π decay from hitting the tagger.

Employing a 500 t ν detector (e.g. ICARUS at Fermilab or ProtoDune-SP/DP at CERN) located 100 m from the entrance of the decay tunnel and a 30 GeV (450 GeV) proton driver, 5×10^{20} (5×10^{19}) protons on target would allow to have a sample of 10^4 tagged ν_e^{CC} interactions.

The positron tagger can be safely operated in terms of pile-up, if local particle rates are below $\sim$1 MHz/cm^2. This can be achieved with multi-Hz slow extractions with a $\mathcal{O}$(ms) duration. Extractions significantly longer than 10 ms are disfavored if focusing of secondaries is achieved by magnetic horns. This constraint could be removed by designing a very efficient focusing system based on DC operated magnets. Within ENUBET, proton extraction schemes compatible with accelerators at CERN, Fermilab and J-PARC will be investigated.

3. Development and tests of the positron tagger prototypes

The basic calorimetric unit (Ultra-Compact Module - UCM) is made of five, 15-mm thick, iron layers interleaved by 5-mm thick plastic scintillator tiles. The transverse dimension is 3×3 cm^2 for a 10 cm length (4.3 X_0). Nine wavelength shifting (WLS) fibers crossing the UCM are

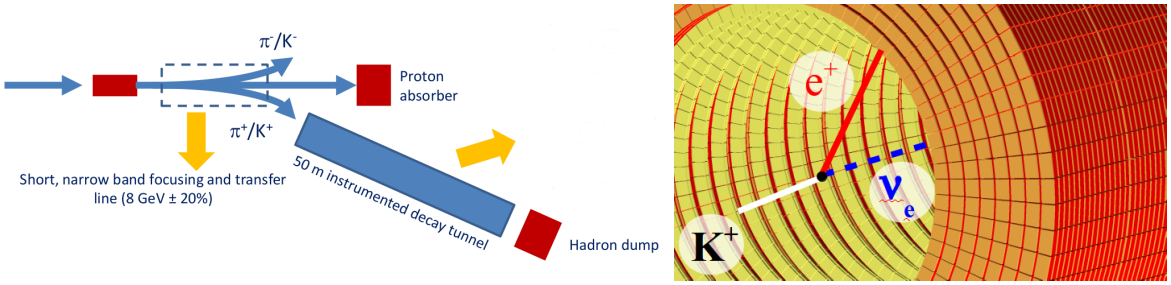


Figure 1. (Left) Layout of the beamline. (Right) A section of the positron tagger. The UCM modules are shown in red and the photon veto detector in yellow.

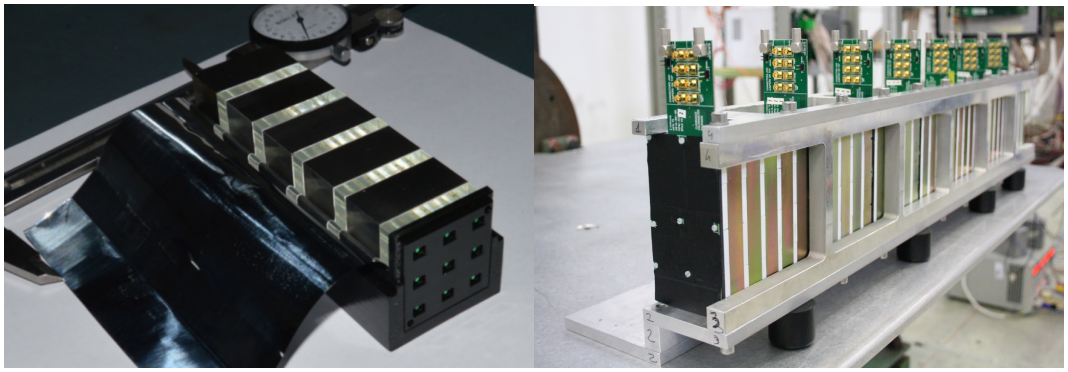


Figure 2. (Left) A single UCM. (Right) An array of 56 UCM units under test at the CERN-PS T9 area in November 2016. The PCBs holding the SiPM arrays are also visible.

.

connected directly to 1 mm^2 SiPM through a plastic holder (Fig. 2, left). Unlike conventional shashlik calorimeters this scheme avoids the occurrence of large passive regions usually needed to bundle the fibers and route them to a common photo-sensor. SiPMs are hosted on a PCB (Fig. 2, right) and the output signals are summed and routed to the FE electronics (fast digitizers) by Copper-Kapton lines. This readout offers outstanding flexibility in terms of choice of granularity and homegeneity in the longitudinal shower sampling.

UCM modules have been tested with cosmic rays in spring 2016 and characterized with charged particles (e/π) in the 1-5 GeV range at the CERN-PS East Area facility in July and November 2016 [4]. An multivariate analysis based on a Geant4 simulation confirms that the UCM e^+ identification and $\pi^{\pm,0}$ rejection is appropriate for ENUBET: an efficiency of 49% for e^+ is obtained with $\pi^\pm$ and π^0mis-identification probabilities of 2.9% and 1.2%, respectively. During the November 2016 CERN test beam a module composed of 56 UCM in 7 longitudinal layers ($\sim 30\ X_0$) and of an outer module acting as an energy catcher, were exposed to e^-/π^- with grazing incidence at various incidence angles. These data, which are currently being analysed, will allow to test the expected $e/\pi^\pm$ performance with data under realistic conditions. In 2017 new prototypes will be tested to assess the recovery time and radiation hardness requirements. A campaign of measurements is also foreseen at the INFN-LNF BTF and at INFN-LNL to test the response to low-energy electrons/photons and neutrons respectively. Finally a full demonstrator (3 m in length, 180° coverage) including the photon veto (Fig. 1, right) for the identification of γ originating from π^0 will be assembled and tested at CERN.

4. ENUBET at CNAF

The activity of ENUBET at CNAF is still in its infancy since it just began in the final part of 2016. A storage of 10 TB has been asked with the goal of hosting data collected during the 2016 test beams in July and November. A dedicate virtual machine (ui-enubet) is already being used by ENUBET users to perform analysis on these datasets using ROOT based user codes. The DAQ scheme which we foresee is based on triggerless acquisition of digitized waveforms from the UCM modules over a time scale of several ms. Given the relatively large amount of DAQ channels of the ENUBET demonstrator (3800) we are developing custom digitizer boards and performing studies to optimize the granularity and resolution of the digitization as well as proper data transfer protocols. Raw data are currently stored in the form of ROOT files. Further processing is performed on the raw data to extract a set of relevant parameters for each waveform based on algorithms which are being developed. Higher level analysis (energy and time resolution, saturation effects, electron pion separation studies) are performed on this reduced data set allowing faster running times and feedback. We foresee at least four weeks of data

taking at the CERN-PS per year (already assigned in 2017 by CERN). The plan is to continue the activity of storage and analysis of the future enlarged experimental data and eventually require more refined computing resources (i.e. access to computing queues) depending on the development of the needs of our users community. Further possibilities which we are considering is hosting the computing needs which will emerge as a consequence of the ongoing studies on the hadronic beamline design or the tagger simulation efforts.

5. Conclusions

The final goal of ENUBET (2016-2021) is to demonstrate that e^+-monitored ν_e beams can be built using existing technologies and hosted at CERN, Fermilab or J-PARC for a new generation of experiments on neutrino cross sections. The results obtained in the preparatory and initial phase of the project are very encouraging. Full simulations support the viability and effectiveness of the calorimetric approach whereas prototypes tests in 2016 demonstrate that shashlik calorimeters with longitudinal segmentation can fulfill the requests for e^+ monitoring in the relevant energy region of few GeV.

6. Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union Horizon 2020 research and innovation programme (grant agreement No 681647).

References

[1] IOP Publishing is to grateful Mark A Caprio, Center for Theoretical Physics, Yale University, for permission to include the `iopart-num` BibTeXpackage (version 2.0, December 21, 2006) with this documentation. Updates and new releases of `iopart-num` can be found on `www.ctan.org` (CTAN).
[2] A. Longhin, L. Ludovici and F. Terranova, Eur. Phys. J. C 75, 155 (2015).
[3] A. Berra et al. [ENUBET Coll.], CERN-SPSC-2016-036; SPSC-EOI-014.
[4] A. Berra et al., Nucl. Instrum. Methods A 830, 345 (2016).
[5] A. Meregaglia et al., Talk at IPRD16, 3-6 Oct. 2016, Siena. Italy.
[6] C. Brizzolari et al., Talk at IEEE (NSS/MIC), 29 October - 5 November 2016, Strasbourg, France.

FAMU: studies of the muon transfer process in a mixture of hydrogen and higher Z gas

E. Mocchiutti on behalf of the FAMU Collaboration
INFN, Trieste, IT
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, preliminary measurements of the muon transfer rate from hydrogen to higher Z gas are needed. In 2015 and 2016 the FAMU collaboration had two data taking sessions at the Rutherford Appleton Laboratory (UK) aimed at the measurement of the temperature dependence of the transfer rate from muonic hydrogen to oxygen. Preliminary results are presented in this paper.

1. FAMU goals

The final aim of this experiment is a precision spectroscopic measurement of the hyperfine splitting (hfs) in the 1S state of muonic hydrogen $\Delta E_{hfs}(\mu^- p)_{1S}$ – providing crucial information on proton structure and muon-nucleon interaction in order to provide the proton Zemach radius r_Z with higher precision than previously possible [1]. This will allow disentangling discordant theoretical values and evidencing any level of discrepancy that may exist between values of r_Z as extracted from normal and muonic hydrogen atoms. It will also set a needed cornerstone result about not yet explained anomalies within the proton charge mean square radius r_{ch} [2, 3]. By measuring the $\Delta E_{hfs}(\mu^- p)_{1S}$ transition wavelength with unprecedented precision $\delta\lambda/\lambda < 10^{-5}$ this experiment will establish new limits on the proton structure parameters and shed light on the low momentum limit of the magnetic-to-charge form factor ratio [4, 5]. The physical process behind this experiment is the following: muonic hydrogen atoms are formed in an appropriate gas target containing a mixture of hydrogen and a higher-Z gas. A muonic hydrogen atom in the ground state, after absorbing a photon of energy equal to the hyperfine-splitting resonance-energy $\Delta E_{hfs} \approx 0.182$ eV is very quickly de-excited in subsequent collision with the surrounding $H2$ molecules. At the exit of the collision the muonic atom is accelerated by $\sim 2/3$ of the excitation energy ΔE_{hfs}, which is taken away as kinetic energy. The experiment will observe the products of a reaction whose rate depends on the kinetic energy of the muonic atoms. The observable is the time distribution of the characteristic X-ray emitted from the muonic atoms formed by muon transfer from hydrogen to the atom of the admixture gas $(\mu^- p)+Z \rightarrow (\mu Z)^*+p$ and its response to variations of the laser radiation wavelength. The $(\mu^- p)_{1S}$ hfs resonance is recognized by the maximal response to the tuned laser wavelength of the time distribution of X-ray K-lines cascade from $(\mu Z)^*$ cascade. By means of Monte Carlo simulations based on the existing data it has been shown that the described method will provide the expected results [6]. The experiment takes place at the Rutherford Appleton Laboratories (UK) where the RIKEN–RAL muon complex [7] is located, the only facility in the world able to provide the pulsed muon

beam which is needed for the experiment.

2. FAMU computing model

As previously done [8], signals from all detectors were digitized and recorded real time event by event using a trigger signal. The trigger signal is 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. In 2016 during seven days of data taking more than 10^7 events were recorded, processed on-line at RAL and transferred for a more deep analysis at CNAF via GridFTP.

At CNAF data are also compared to results coming from the simulation, since CNAF hosts the collaboration Monte Carlo production site. In 2016 most of the Monte Carlo production concerned the study of the background for the transfer rate measurement.

3. Measurement of the transfer rate

In 2016 the apparatus consisted of a gas target filled with a mixture of hydrogen and heavier gasses surrounded by the detectors. In front of the cryogenic gas target a hodoscope was used to measure the shape and timing of the muon beam. On the sides 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.

During the first tests in 2014 at the beam delivery Port 4 of the RIKEN-RAL facility, the detection system and the beam condition allowed to establish a satisfactory background situation [9, 10]. Since the efficiency of the method is bound to the collisional energy dependence of the muon transfer rate, the main focus in 2015-16 has been a detailed experimental analysis of the muon transfer mechanism by measuring the muon transfer rate at various temperatures.

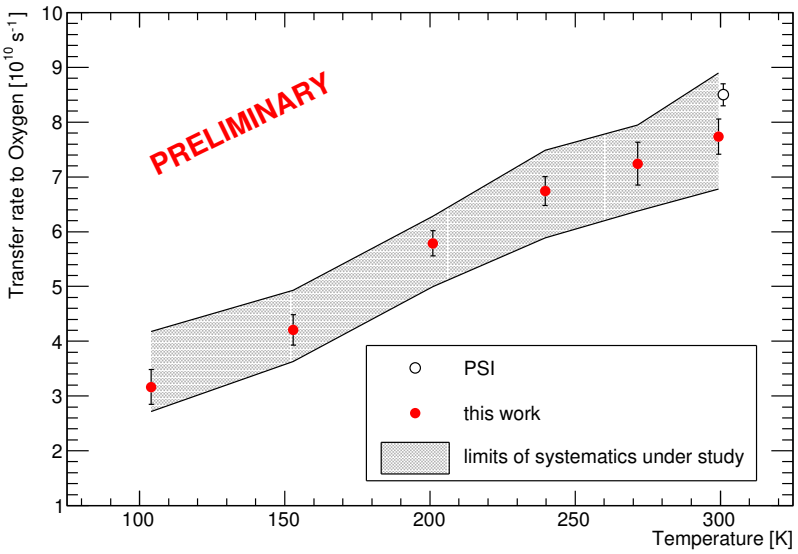


Figure 1. Transfer rate from muonic hydrogen to oxygen as function of the target temperature.

While for many gases the transfer rate at low energies is nearly constant, there is experimental evidence of a relevant energy dependence for Oxygen [10]. Our team has performed at PORT4 the needed dedicated study of muon transfer from hydrogen to the atom of an admixture gas, and in particular to oxygen, at temperatures between 300 and 100 K to confirm the energy

dependent muon transfer rate for oxygen and the profitability of this method. Results are shown in Fig. 1. Filled circles represent this work results at six fixed temperatures. Statistical error bars are reported while the gray band show the limits of the systematic errors still under study. The open circle at 300 K is the only published measure [11] of the transfer rate from muonic hydrogen to oxygen, which is in agreement with our results. While the final data are still subject to careful verification the presently extracted behavior of the transfer rate confirms the expectations and the proposed experimental method.

To achieve this result great care has been put in the simulation [12], to finalize the design and the construction of the experimental layout with particular regard to the cryogenic high-purity gas-target able to work at temperatures below 50K and pressures up to 40 atmospheres. The tasks were: to minimize the material at the beam entrance window of the thermally isolated structure so to keep a minimal induced spread of the low momentum beam to minimize the thickness of the lateral walls so to allow high transparency to the X-rays of the muonic cascade of interest and finally to coat the internal vessel with thin layers of high Z material in order to promote fast nuclear muon capture for non hydrogen muonic atoms and minimize the non prompt noise induced by decay electrons.

4. Conclusions

The first measurement of the transfer rate from muonic hydrogen to oxygen has been performed by the FAMU collaboration in 2016. CNAF continues to play a major role in the computing of the FAMU experiment. Most of the computing resources and all the storage capabilities 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 also would like to thank CRIOTEC - Impianti S.R.L. (Chivasso, TO, Italy) for the extremely professional and proficuous collaboration in the realization of the cryogenic and pressurized target.

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.

References

[1] Adamczak A et al 2015 *Phys. Lett. A* **379**
[2] Pohl R et al 2010 *Nature* **466**
[3] Antognini A et al 2013 *Science* **339**
[4] Karshenboim S G, McKeen D, and Pospelov M 2014 *Phys. Rev. D* **90**
[5] Karshenboim S G 2014 *Phys. Rev. A* **91**
[6] Bakalov D et al 2015 *Phys. Lett. A* **379**
[7] Matsuzaki T, Ishida K, Nagamine K, Watanabe I, Eaton G and Williams W 2001 *Nucl. Inst. Meth. Phys. Res. A* **465**
[8] Mocchiutti E 2016 *INFN-CNAF Annual Report 2015*
[9] Adamczak A et al 2016 *J. of Inst.* **11**
[10] Vacchi A et al 2016 *to appear in RIKEN Accelerator Progress Report 2015-16*
[11] Werthmüller A et al 1998 *Hyperfine Interactions* **116**
[12] Danev P et al 2016 *J. of Inst.* **11**

The Gerda experiment

K. von Sturm on behalf of the Gerda Collaboration
Università di Padova, Padova, IT
E-mail: vonsturm@pd.infn.it

Abstract. The GERmanium Detector Array (GERDA) experiment at Laboratori Nazionali del Gran Sasso (LNGS) of INFN searches for neutrino-less double beta ($0\nu\beta\beta$) decay in ^{76}Ge. High-purity germanium (HPGe) detectors enriched in ^{76}Ge are operated bare in liquid argon (LAr). The LAr cryostat is contained inside a water Cerenkov muon veto which additionally shields from neutrons and gammas. Since December 2015 GERDA is running in its second experimental stage (Phase II). Detectors with improved pulse shape discrimination (PSD) capabilities are deployed and an active veto is installed; it uses the scintillation light of the LAr to flag background events. With the unblinding of the first 10.8 kg yr of data collected in Phase II the ambitious design background index of 10^{-3} cts/(keV kg yr) was reached. This makes GERDA the first experiment in the field expected to operate in the limit of zero-background for the entire Phase II data taking. With no signal detected a lower half-life limit of $5.3 \cdot 10^{25}$ yr (90% C.L.) is given (median sensitivity $4.0 \cdot 10^{25}$ yr). In this note we present a short description of the GERDA Phase II setup, status and the activities carried out at CNAF related to data analysis, background modeling and Monte Carlo simulations.

1. The Gerda Phase II setup

One prediction of many extensions of the standard model of particle physics is the existence of neutrino-less double beta ($0\nu\beta\beta$) decay [1]. This hypothetical process violates lepton number by two units and its existence would prove that neutrinos have a Majorana mass component. Exceptional background control and rejection is needed to explore the parameter space of $0\nu\beta\beta$ decay exceeding 10^{25} yr in half-life.

The GERmanium Detector Array (GERDA) [2] is one of the leading experiments in the field. It searches for $0\nu\beta\beta$ decay in ^{76}Ge using enriched high-purity germanium (HPGe) detectors immersed bare in liquid argon (LAr). A water tank surrounding the LAr cryostat serves as passive shield and as active water Cerenkov muon veto [3]. In December 2015 GERDA has started its second data taking phase after a major upgrade [4]. The setup incorporates 7 enriched (15.8 kg) and 3 natural (7.6 kg) semi-coaxial detectors and 30 enriched Broad energy germanium (BEGe) detectors (20.0 kg) which exhibit an enhanced pulse shape discrimination (PSD) performance [5]. The detectors are mounted in seven strings each contained in a nylon enclosure commonly referred to as 'mini-shroud' [4]. They prevent radioactive ions, namely ^{42}K daughter of ^{42}Ar, to migrate close to the detector surfaces and at the same time they let LAr scintillation photons pass. This is important for the newly implemented LAr instrumentation: a curtain of 800 m of light guiding fibers which are read on both ends by Si photomultipliers (SiPMs) [6] and 16 radio-pure PMTs for cryogenic operation in two arrays above and below the germanium detector array [7] observe light emitted when particles deposit energy in the LAr. The fibers are coated

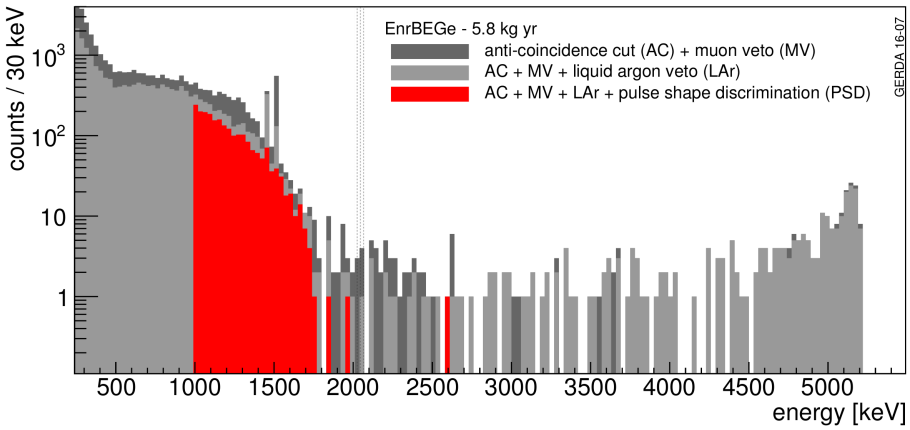


Figure 1. BEGe energy spectrum of the first 5.8 kg yr of exposure taken in GERDA Phase II. The spectrum is shown with anti-coincidence and muon-veto cut only (dark gray), with LAr cut added (light gray) and with LAr and PSD cuts added (red). The ROI ($Q_{\beta\beta} \pm 25$ keV) is indicated by vertical dashed lines.

with a wavelength shifter to match the range of the SiPMs. In total 15 channels of SiPMs and the 16 channels of the PMTs are read when the germanium array triggers. The introduced dead time due to the LAr veto system is about 2.3 %.

In general, PSD is complementary to the LAr veto and potentially rejects different background. It identifies multiple-site events depositing energy in multiple locations in one detector and detector surface events with single-energy depositions [8]. In figure 1 the energy spectrum of the first 5.8 kg yr of Phase II BEGe data is shown with the different veto cuts applied. The high energy region which is populated by surface alpha events is completely vetoed by PSD and ^{42}K is highly suppressed by the LAr veto (see gamma line at 1525 keV in figure 1).

An exceptional background index in the region of interest has been achieved with the first 10.8 kg yr of unblinded Phase II data after PSD and LAr veto cuts. With the BEGe detectors the ambitious design goal has been fully achieved with a background index of $\approx 10^{-3}$ cts/keV kg yr. With no signal detected the preliminarily given lower half-life limit is $5.3{\cdot}10^{25}$ yr at 90% C.L. with a median sensitivity of $4.0 \cdot 10^{25}$ yr [9, 10]. The experimental goal of GERDA is to reach a half-life sensitivity of 10^{26} yr by collecting about 100 kg yr of data. The BEGe data set is expected to be background-free in the entire data taking period which means that the sensitivity of GERDA will grow almost linearly with the collected exposure. The exceptional performance of GERDA, especially in terms of background reduction, justifies larger germanium experiments.

2. Gerda at CNAF

GERDA being a low background experiment, produces about 4 GB/day of physics data and about 20 GB/calibration which is done roughly twice a month. The policy of the GERDA collaboration requires three backup copies of the raw data (Italy, Germany and Russia) which is stored at the experimental site at INFN's Gran Sasso National Laboratory (LNGS) cluster. CNAF is the Italian backup storage center, and raw data is transferred from LNGS to Bologna in specific campaigns of data exchange. For storage GERDA uses the resources of the GRID through the dedicated virtual organization (VO) `gerda.mpg.de@lfc.italiangrid.it`. About 10 TB of data, mainly Phase I, are currently secured and registered in the GERDA catalog at CNAF.

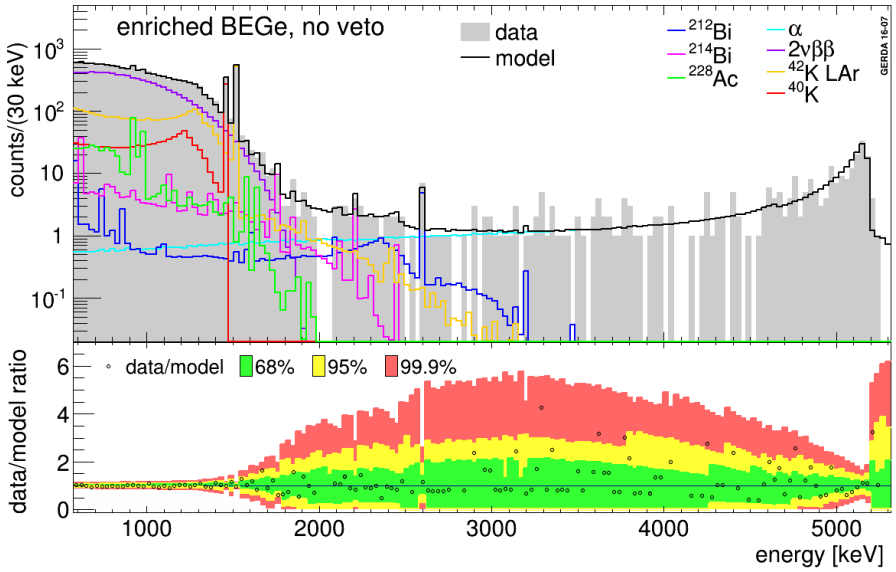


Figure 2. Preliminary minimal Phase II background model. BEGe spectrum before LAr and PSD cuts; Quality and Muon-veto cuts applied. In the bottom plot the ratio between data and model and the Poisson probability for the event distribution in each bin are shown (green 1σ, yellow 2σ, red 3σ).

The GERDA data is organized in a hierarchical tier structure where *tier0* is the unprocessed raw data containing the full traces of all detectors. Each higher tier contains more advanced analysis output. Between *tier0* and *tier1* a data blinding is applied; each event is automatically analyzed for its energy and events in the region (2039 ± 25) keV are removed from further processing. *Tier1* data are saved in Root [11] format based on custom Majorana GERDA Data Objects (MGDO) [12]. In the standard analysis, which is optimized for $0\nu\beta\beta$ analysis, up to *tier2* level three separate analysis chains for germanium, SiPM and PMT data are processed. At *tier3* level the three chains are unified to a single one. However, the modular approach of the GERDA software [13] allows each user to produce custom analysis tiers with non standard analysis chains.

For data analysis all output parameters produced are accessible via a Data Loader developed for the GERDA tiers. It connects all processed data files using their unique keys in a single master event chain. A number of quality cuts are applied which reject pile-up events, discharges and events with multiplicity greater than one (anti-coincidence).

For data processing and analysis a dedicated machine `ui-gerda.cr.cnaf.infn.it` is provided at CNAF. The processed tiers of Phase II data are available for data analysis on the gerda-ui. Moreover, the complete GERDA software is installed and users can load the full suite using a properly set environment. This creates an environment, especially for GERDA members in Italy, that is ready-to-use, in addition to the LNGS machines which usually have a high task load. In GERDA, software developments are versioned and controlled using *git* connected to a private GERDA *github* repository. Installation is managed by the software module management tool *SWMOD* which keeps track of all dependencies (Root, CLHEP, Geant4 [14], etc.) and loads the necessary environment variables. The complete data reconstruction work flow of batch jobs is handled by the *Luigi* python module [15], which offers a web interface to search and filter among all the tasks.

The CPU resources at CNAF are used for Monte Carlo (MC) simulations done with MaGe [16] which is a Geant4 [14] based development of the Majorana and GERDA collaborations. This is an essential part of the background modeling and search for exotic decay modes. The MC simulations are used to produce the prior probability density functions (pdf) for GERDA intrinsic background components, the two-neutrino double beta ($2\nu\beta\beta$) decay spectrum and exotic processes like Majoron decay modes or spectral distortions due to violation of Lorentz invariance. Background decomposition is vital for $0\nu\beta\beta$ analysis and the extraction of the background index in the region of interest. Moreover, by identifying the location and sources of background these can be tackled and or eliminated in further upgrades and successor experiments. Up to $5 \cdot 10^9$ particles in up to 1200 CPUh are simulated per component. The externally parallelized jobs are submitted to the gerda-queue in the Load Sharing Facility (LSF) at CNAF and the full hit information is saved to disk such that the desired spectra can be produced in a post-processing step.

A new background model is being developed for Phase II as constituents of the array were changed and the detector mass was about doubled with respect to Phase I. A part of the Phase II background modeling is conducted at CNAF using the Bayesian analysis toolkit (BAT) [17]. The prior pdfs generated in the MC simulations are scaled in a Bayesian binned maximum likelihood fit and from the posterior distributions the best fit parameters are extracted (for procedural details see [18]). Various source locations and combinations are fit in order to find the best combination and evaluate systematic uncertainties. A preliminary minimal model is shown in figure 2 based on the first 5.8 kg yr of Phase II BEGe data. The model contains isotopes from the ^{238}Ur and ^{232}Th primordial decay chains, which are naturally present in every material, ^{40}K, an irreducible background due to $2\nu\beta\beta$ decay and ^{42}K, daughter of ^{42}Ar. The latter is an important background source for GERDA which is, however, highly suppressed by the LAr veto. Intrinsic contamination due to ^{86}Ga and ^{60}Co are insignificant as the detectors were kept under ground as much as possible during production to prevent cosmic activation of the detector bulk material [5]. Moreover, the half-life of $2\nu\beta\beta$ decay is extracted from this analysis. With $T^{2\nu}_{1/2} = (1.926 \pm 0.094) \cdot 10^{21}$ yr GERDA provides the most precise value on the half-life of $2\nu\beta\beta$ decay of ^{76}Ge [19].

References

[1] H. Päs and W. Rodejohann, Neutrinoless double beta decay, New J. Phys. 17 (2015) 115010
[2] K.-H. Ackermann *et al.* (GERDA Collaboration), Eur. Phys. J. C73 (2013) 2330
[3] K. Freund *et al.* Eur. Phys. J. C76 (2016) 298
[4] B. Majorovits (GERDA collaboration), Physics Procedia 61 (2015) 254
[5] M. Agostini *et al.* (GERDA Collaboration), Eur. Phys. J. C75 (2015) 39
[6] J. Janicskó-Csáthy *et al.*, (2016) (Preprint arXiv:1606.04254)
[7] M. Agostini *et al.*, Eur. Phys. J. C 75 (2015) 506
[8] M. Agostini *et al.* (GERDA Collaboration), Eur. Phys. J. C73 (2013) 2583
[9] B. Schwingenheuer on behalf of the GERDA Collaboration, LNGS (2016)
`https://www.mpi-hd.mpg.de/gerda/public/2016/t16_gerda_phase2_bs.pdf`
[10] M. Agostini on behalf of the GERDA Collaboration, Neutrino Conf. (2016)
`https://www.mpi-hd.mpg.de/gerda/public/2016/t16_neutrino_gerda_ma.pdf`
[11] R. Brun and F. Rademakers, Nucl. Inst. & Meth. in Phys. Res. A 389 (1997) 81
[12] M. Agostini *et al.*, J. of Phys.: Conf. Ser. 375 (2012) 042027
[13] M. Agostini *et al.*, J. of Phys.: Conf. Ser. 368 (2012) 012047
[14] S. Agostinelli *et al.* (GEANT collaboration), Nucl. Inst. Methods A 506 (2003) 250
[15] `https://github.com/spotify/luigi` [2.3.2017]
[16] M. Boswell *et al.*, IEEE Trans. Nucl. Sci. 58 (2011) 1212
[17] A. Caldwell *et al.*, Comput. Phys. Atom. Nucl. 63 (2000) 1282
[18] M. Agostini *et al.* (GERDA Collaboration), Eur. Phys. J. C74 (2014) 2764
[19] M. Agostini *et al.* (GERDA Collaboration), Eur. Phys. J. C 75.9 (2015) 416

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] INFN, Pisa, IT
[2] Department of Physics, University of Trieste, and INFN, Trieste, IT
[3] Laboratoire Univers et Particules, Université de Montpellier, CNRS/IN2P3, Montpellier, FR
[4] Departement de Physique Nucléaire et Corpusculaire, Université de Genève, Geneva, CH

E-mail: michael.kuss@pi.infn.it

Abstract. The *Fermi* Large Area Telescope current generation experiment dedicated to gamma-ray astrophysics is using 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..

By end of 2016, the LAT had registered 510 billion triggers (1800 Hz average trigger rate). An on-board filter analyses the event topology and discards about 80%. Of the 102 billion events that were transferred to ground 990 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 2016

In 2016, 63 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 37. Also external scientists are able to analyse the *Fermi* public data, resulting in about 100 external publications.

In 2016, Fermi findings triggered 6 NASA press releases. It started on 7 January with the announcement of the implementation of a new improved analysis algorithm (called Pass 8) [2]. However, the work for Pass 8 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

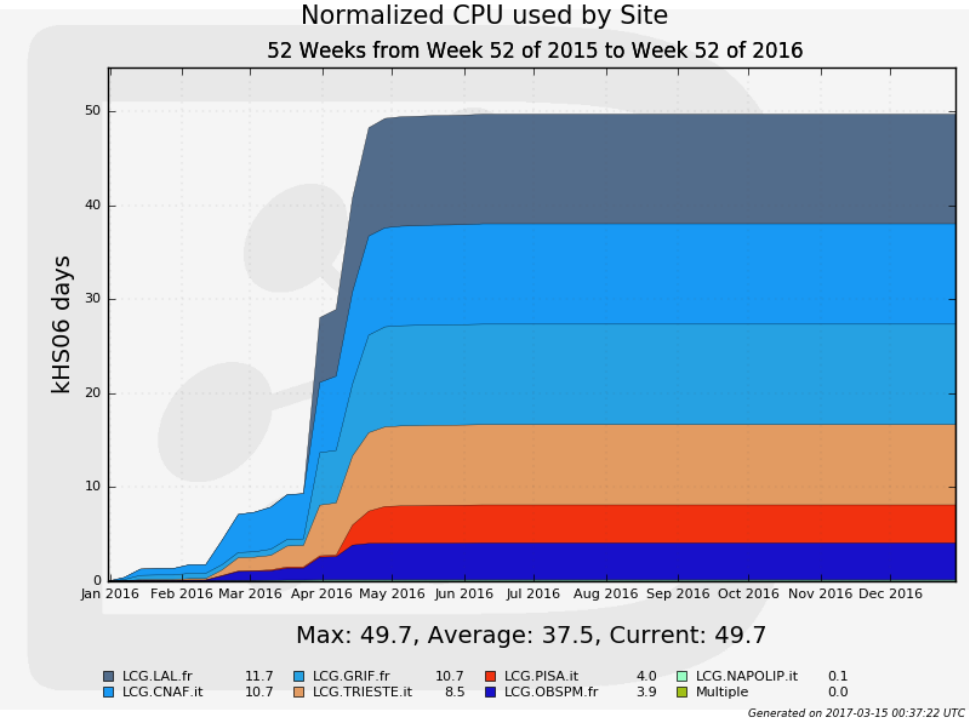


Figure 1. Cummulative usage plot for the VO glast.org in 2016

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 were in the public data base, but now, with 8.5 years of data and Pass 8, they amount to 990 million, thus doubling the number of photons with just 30% more life time. 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, and improved science results were published in 2015 and 2016.

The major scientific achievement of 2016 was the discovery of gravitational waves by the LIGO observatory. The *Fermi* collaboration published 2 papers [3, 4, 5] on the search of an electromagnetic counterpart, which was difficult due to the large search region. With improved sensitivity of the GW obervatories and localization through the addition of Virgo in future it will be possible to test the expectation that stellar BH mergers do not emit much radiation. But gamma-rays can also be produced together with other radiation than gravitational waves. Public data of *Fermi* were used to connect the brightening in gamma-rays of a distant blazar to the detection of a VHE neutrino in the IceCube detector [6, 7]. Instead [8, 9] discuss the link between the emission of infrared and gamma radiation of blazars.

Finally, 2016 saw the publication of two source catalogues, the second *Fermi* LAT Hard Sources (2FHL) catalog [10] of sources seen at energies laregr than 10 GeV, and the first *Fermi* LAT Supernova Remnant (1SNR) catalog [11].

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 (Fig. 1) constitute a substantial fraction of the submitted jobs. However, we also exploit the possibility to submit jobs directly using the grid middleware. About one quarter of the jobs were run at the INFN Tier 1 at CNAF as shown by Fig. 2. The total usage in 2016 was about 150 HS06.

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. However, *Fermi* is already in a phase-out stage. The regular data processing and possible re-processing is managed by SLAC. The not regular MC production is currently taken care by the Lyon farm. Note however that we expect that our usage may rise again, partially due to changes in the computing model associated with long term mission plans of Fermi-LAT, and partially to satisfy specialized simulation needs (e.g. for e^+/e^- analysis and MC production to study the LAT's ability to detect polarized pair production).

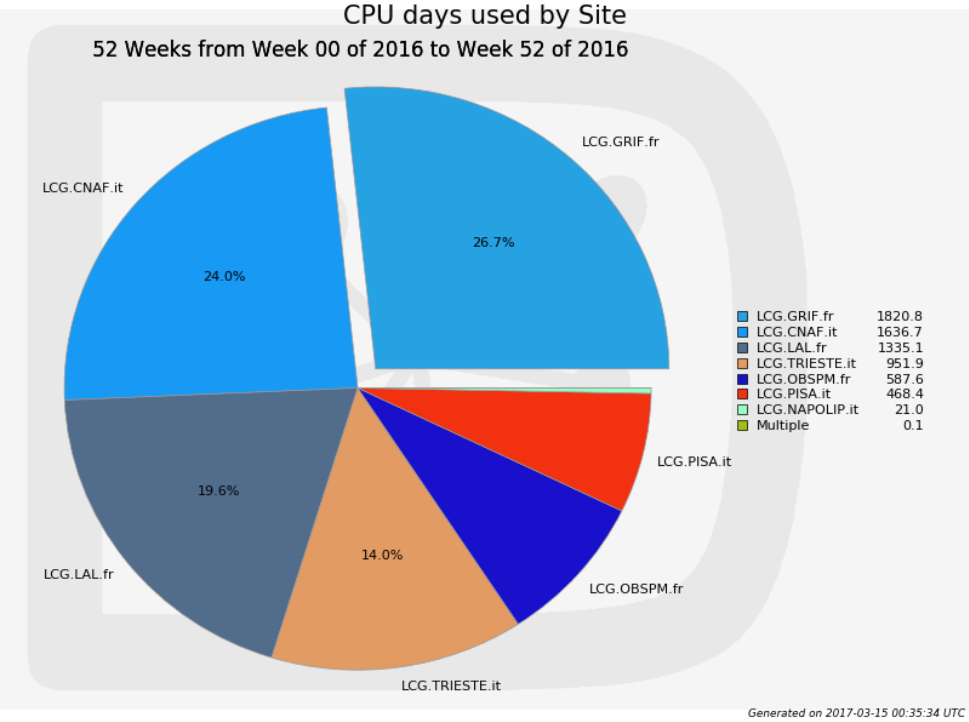


Figure 2. Usage of grid sites by the VO glast.org in 2016

References

[1] Atwood W B et al. 2009 *Astrophysical Journal* **697**, 1071
[2] NASA press release 2016 January 7, http://www.nasa.gov/feature/goddard/2016/nasas-fermi-space-telescope-sharpens-its-high-energy-vision/
[3] Ackermann M et al. 2016 *Astrophysical Journal Letters* **823** no.1, L2
[4] Connaughton V et al. 2016 *Astrophysical Journal Letters* **826** no.1, L6
[5] NASA press release 2016 April 18, https://www.nasa.gov/feature/goddard/2016/nasas-fermi-telescope-poised-to-pin-down-gravitational-wave-sources/
[6] Kadler M et al. 2016 *Nature Physics* **12**, 807
[7] NASA press release 2016 April 28, https://www.nasa.gov/feature/goddard/2016/nasas-fermi-telescope-helps-link-cosmic-neutrino-to-blazar-blast/
[8] Massaro F and D'Abrusco R 2016 *Astrophysical Journal* **827** no.1, 67
[9] NASA press release 2016 August 24, https://www.nasa.gov/feature/goddard/2016/nasas-wise-fermi-missions-reveal-a-surprising-blazar-connection/
[10] Ackermann M et al. 2016 *Astrophysical Journal Supplement* **222** no.1, 5
[11] Acero F et al. 2016 *Astrophysical Journal Supplement* **224** no.1, 8
[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

Juno experiment at CNAF

S. M. Mari, C. Martellini, G. Settanta

INFN and Università degli Studi Roma 3, Roma, IT

E-mail: stefanomaria.mari@uniroma3.it; martellini@roma3.infn.it; giulio.settanta@uniroma3.it

Abstract.

The Jiangmen Underground Neutrino Observatory (JUNO) is a 20 kton multi-purpose underground neutrino detector which was proposed in order to determine, as its primary goal, the neutrino mass hierarchy. The large fiducial volume, together with the excellent energy resolution forseen, would allow to perform also a series of important measurements in the field of neutrinos and astro-particle physics.

To address the mass hierarchy issue, the JUNO detector will be surrounded by a cluster of nuclear power plants at a distance of around 50 km. The resulting reactor antineutrino flux gives the possibility to determine the neutrino mass hierarchy with a significance level of 3-4σ, with six years of running JUNO. The measurement of the antineutrino spectrum with excellent energy resolution will lead also to a precise determination of the solar neutrino oscillation parameters, $sin^2\theta_{12}$ and Δm^2_{21}, with an accuracy below 1%.

The JUNO characteristics make it a suitable detector not only for neutrinos coming from the power plants, but also for neutrinos generated inside the Sun, or during supernovae explosions, or even in the Eart's crust and atmosphere. Other topics of interest potentially accessible to JUNO include the search for sterile neutrinos, proton decay and dark matter annihilation.

Data taking is expected to start in 2020.

1. The JUNO experiment

The standard electroweak model has been proved by many experiments to be a successful theory that not only unifies the electromagnetic and weak interactions, but also explains almost all the phenomena observed in nature, below the electroweak scale. In its original formulation of Weinberg in 1967 [1], neutrinos were assumed to be massless and hence not allowing any lepton flavor mixing. Later observations of a flux deficit for solar neutrinos [2, 3] turned out to be a solid evidence of physics beyond the Standard Model: the deficit was explained in terms of a neutrino flavor oscillation, which takes place due to the fact that neutrinos do have mass.

Neutrinos oscillation is today a well observed phenomenon. By means of a simple extension of the Standard Model, it can be mathematically described in terms of two separated sets of eigenstates, which do not correspond: three flavor eigenstates (ν_e, ν_μ, ν_τ) and three mass eigenstates (ν_1, ν_2, ν_3). The mixing between the eigenstates is described by the three mixing angles θ_{12}, θ_{23}, θ_{13}. So far, the absolute value of the three neutrinos mass is still unknown. The relative square differences Δm^2 have been measured by several experiments, with a precision of few percents, but not the ordering. Such mass hierarchy (i.e. the sign of the square mass difference Δm^2_{13}) is still unknown. The knowledge of this crucial information would have several important implications for fundamental physics.

The Jiangmen Underground Neutrino Observatory (JUNO) is a multi-purpose neutrino experiment, proposed in 2008 to determine the neutrino mass hierarchy by detecting reactor antineutrinos [4]. The detector site has been chosen in order to achieve the best sensitivity to neutrino mass hierarchy. The JUNO complex is currently under construction in China, with a rock overburden above the experimental hall of around 700 m, and is located 53 km away from both Yangjiang and Taishan nuclear power plants. The neutrino detector consists of a 20 kton fiducial mass liquid scintillator (LS), where antineutrinos can interact via inverse beta-decay interactions, producing a positron and a neutron in the final state. Both these secondary particles are then captured and generate scintillation light, which is collected by more than 17.000 photo-multiplier tubes (PMTs) installed on a spherical structure with radius $\simeq$ 20 m. PMTs are submerged in a buffer liquid to protect the LS from the radioactivity of the PMT glass. The scintillator liquid is composed of a mixture of *Linear Alkyl-Benzene* (LAB), 2.5-diphenyloxazole (PPO) and p-bis-(o-methylstyryl)-benzene (bis-MSB), to maximize the collection of light per neutrino event. The central scintillator detector is surrounded by a cylindrical water pool, equipped with $\sim$2000 PMTs to detect the Cherenkov light from cosmic muons and from the environment radioactivity, acting as a veto detector. On top of the water pool there is another muon detector, made of scintillating strips, to accurately identify the muon tracks. The detector design is still developing in the carrying on of R&D.
The $3\%/\sqrt{E[MeV]}$ JUNO energy resolution target will be addressed by maximizing the light yield and the PMT coverage. The current R&D status includes 20" PMTs for the central detector and the water pool, to achieve the high photon statistics.
Given the planned resolution and the neutrino flux from the power plants, the mass hierarchy is expected to be determined in six years of data taking, with a confidence level between 3σ and 4σ. Beside the JUNO main goal, additional measurements can be performed by the detector. The analysis of the reactor antineutrino spectrum, given the large fiducial volume and the excellent energy resolution, can be exploited also for a measurement of the solar oscillation parameters $sin^2\theta_{12}$ and Δm^2_{21} with a sub-percent accuracy, which would represent the most precise measurement in the neutrino solar sector. Supernovae neutrinos can also be observed, inferring important informations on the acceleration processes at the source. The properties of the scintillator can also be exploited to observe the solar neutrino flux, by means of elastic scattering on electrons. The atmospheric neutrino flux will also be accessible, together with an independent measurement of the atmospheric mixing angle θ_{23}. Another component potentially accessible to JUNO is constituted by geoneutrinos, produced by radioactive decays inside the Earth. Exotic searches include non-standard interactions, sterile neutrinos and dark matter annihilation signals.

2. JUNO experiment at CNAF

The Juno DAQ system and the Juno trigger system are now in the final designing phase allowing in the next future a better definition of the computing resources needed to manage the data. The data collected by the experiment will be temporary stored in a computer farm located at the experimental site. A devoted optical fiber link will be used for the data transfer to the IHEP Computing Center located in Beihing where the reconstruction programs will be run. Raw data and processed ones from IHEP will be transferred to CNAF Computing Center. At CNAF data will be reprocessed and stored, and the end user analysis will be supported. Juno will have a duplicate of the raw data at CNAF. The Juno Monte Carlo event production, needed to design the detector and to prepare the data analysis, will be arranged between IHEP and CNAF. A single full sample consists of about 40TBn and it requires about 1000HS06 for more than one month. In the next future a detailed computing model documents will be prepared.

References

[1] S. Weinberg, Phys. Rev. Lett. **19** 1264 (1967);
[2] J. N. Bahcall, N. A. Bahcall and G. Shaviv, Phys. Rev. Lett. **20** 1209 (1968);
[3] J. N. Bahcall, W. F. Huebner, S. H. Lubow, P. D. Parker and R. K. Ulrich, Rev. Mod. Phys. **54** 767 (1982);
[4] F. An et al., J. Phys. G: Nucl. Part. Phys. **43** 030401 (2016).

The KM3NeT neutrino telescope network and CNAF

C. Bozza

Department of Physics of the University of Salerno and INFN Gruppo Collegato di Salerno, Fisciano, IT

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 ordering 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 a diffuse neutrino flux at energies exceeding 1 PeV. The goals of directional identification of the source of high-energy neutrinos and good energy resolution ,together with their small flux, require a detector with a total volume beyond $1\,km^3$. The second line of research is devoted to studying the ordering of neutrino mass eigenstates (Oscillation Research with Cosmics in the Abyss - ORCA). A detector technically 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 ordering (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 and technical reasons 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 with full day operation, 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.

2. Project status

In the ARCA site, 2 DU's are in operation, continuously taking physics data as well as providing the KM3NeT Collaboration with a wealth of technical information including long-term effects and stability indicators. Also data processing is now working in a stable way. Deployment of ORCA DU's is also expected soon. While most of the computing load in previous times was due to simulations for the full building block, several analysis tasks are now running to analyse and asses data quality. Such tasks involve both data from PMT's and slow control data in combination. Consequently, simulations are being enriched with feedback from real data analysis. As a first step, this was done at CC-IN2P3 in Lyon, but usage of other computing centres is increasing and is expected to soon spread to the full KM3NeT computing landscape. This process is being driven in accordance to the goals envisaged in setting up the computing model. The KM3NeT Collaboration is now preparing for the so-called "phase 2.0" aiming at increasing the detector size. The preparatory phase has been funded by an EU Commission grant, starting this year. First deployments for phase 2.0 are foreseen in 2018, following the completion of the "phase 1" detector.

3. 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.

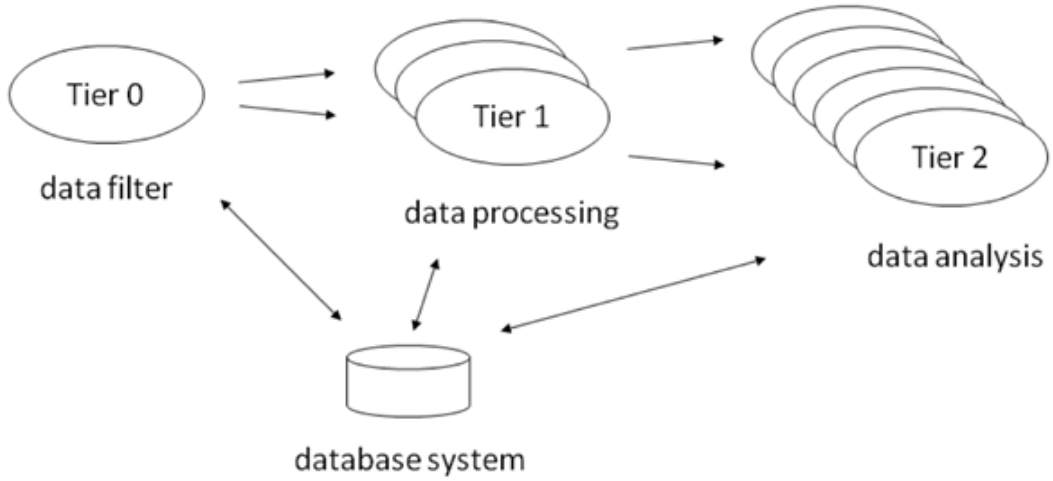


Figure 1. Three-tier model for KM3NeT computing.

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 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.

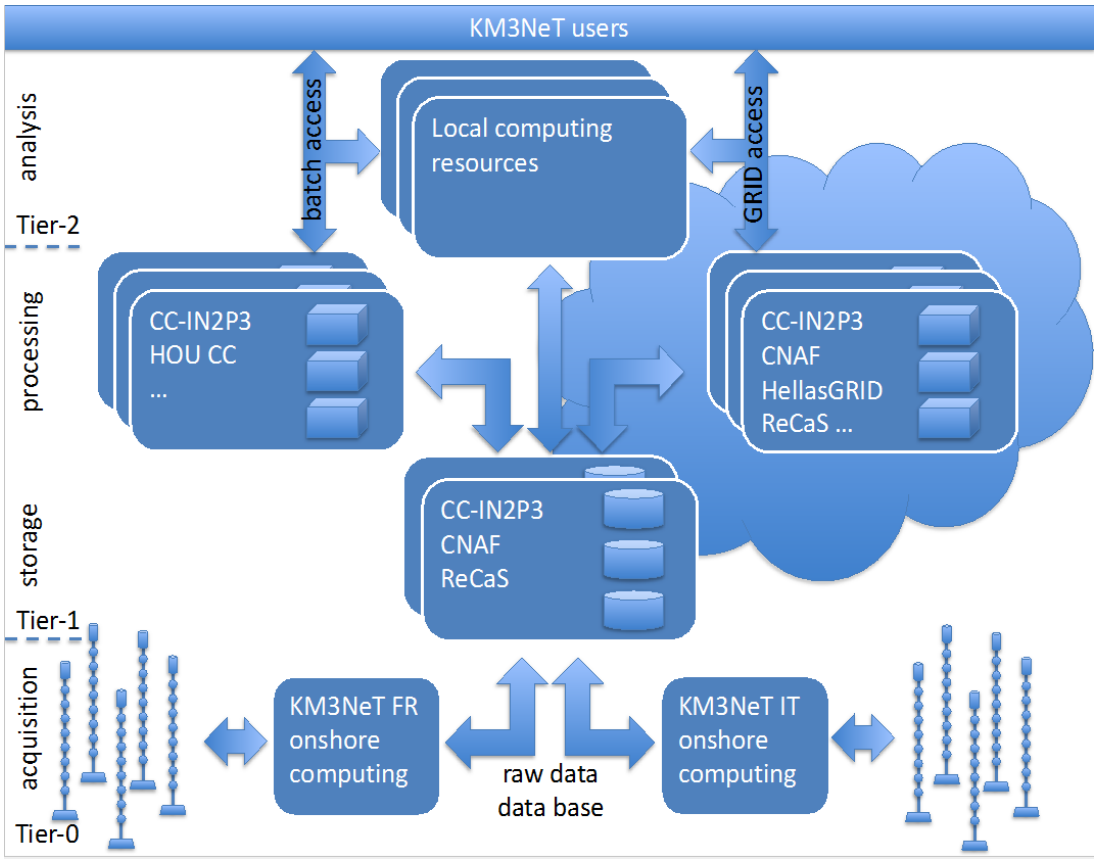


Figure 2. Data flow in KM3NeT computing model.

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. The implementation for the data transfer from CC-IN2P3 to CNAF (via an iRODS-to-gridFTP interface at CC-IN2P3) has been established, which is a techical pre-requisite for mirroring the data sets between those major computing centres. Performances, reliability and fault-tolerance of the link are currently under testing before going into production mode.

4. 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.

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

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

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 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 2018 and the following years. CNAF has already added relevant contributions to KM3NeT in terms of know-how for IT solution deployment, e.g. the Jenkins continuous integration server and the software-defined network at the Tier-0 at the Italian site.

5. 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]

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. An 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 2016 is presented. Emphasis is given to 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 370 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.

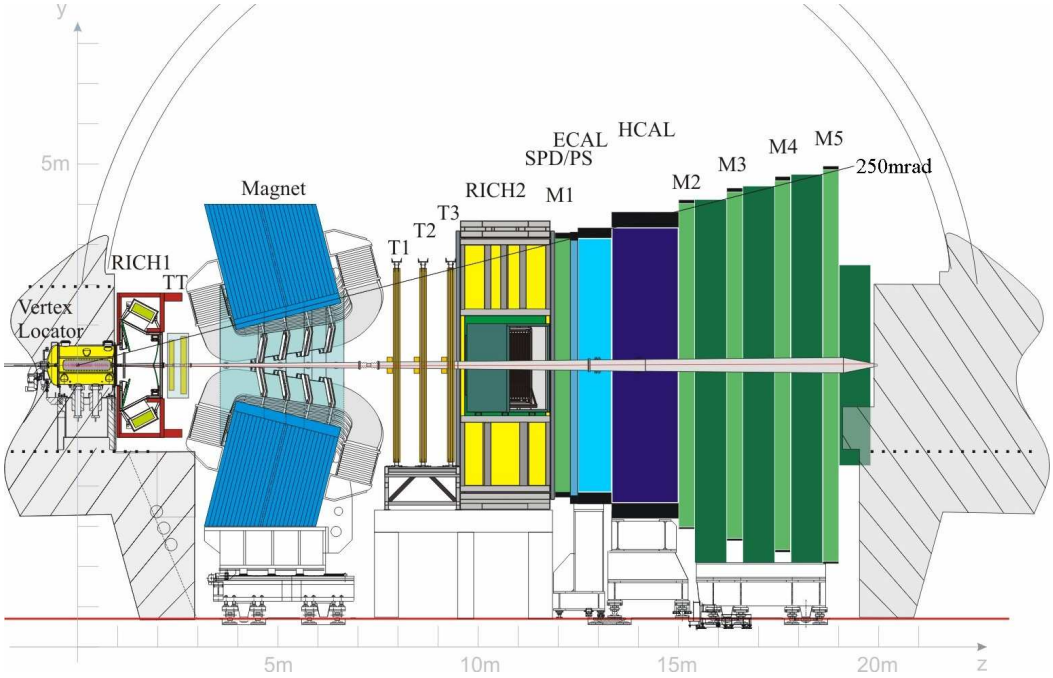


Fig. 1: Sketch of the LHCb detector.

2. Overview of LHCb computing activities in 2016

The usage of offline computing resources involved: (a) the production of simulated events, which runs continuously; (b) running user jobs, which is also continuous; (c) an incremental stripping of Run1 data; (d) validation cycles of the 2015 "TURBO" data; (e) reconstruction of data taken in 2015 in proton-ion collisions; (f) processing of data taken in 2016 in proton and heavy-ion collisions.

Activities related to the 2016 data taking were tested in May and started at the beginning of the LHC physics run in June. The excellent performance of the LHC resulted in a steady processing and export of data transferred from the pit to offline.

LHCb has implemented in Run2 a new trigger strategy, by which the high level trigger 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.

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 on tape. *TURBO* data are subsequently *resurrected*, *i.e.* copied to disk in a micro-DST format that does not require further processing and can be used right away for physics analysis.

By default, an event of the *TURBO* stream contains only information related to signal candidates. In order to ease the migration of analyses using also information from the rest of the event, it was allowed in 2016 data taking to persist the entire output of the online reconstruction (*TURBO++*). This

temporary measure increased the average size of a *TURBO* event from 10kB to about 50kB, with obvious consequences on the required storage space. Efforts are ongoing in order to slim the *TURBO* content by enabling analysts to save only the information that is really needed, with a target *TURBO* event size set to 20kB. Also, work is ongoing on streaming the *TURBO* output, in order to optimize data access.

The LHCb computing model is assuming 10kHz of *FULL* stream and 2.5kHz of *TURBO* stream in proton collisions. In 2016, average rates of 8kHz and 5.5kHz were measured, respectively, resulting in a throughput rate of up to 0.8 GB/s and a transfer rate to the Tier0 of up to 0.6 GB/s during proton collisions. Transfer rates of up to 1.4GB/s were sustained during ion running. Given the LHC live time in 2016 being considerably higher than anticipated, several measures had to be taken in order to cope with the available offline resources. Most notably, the use of storage resources has been optimised by reducing the number of disk-resident copies of the analysis data, and by parking a fraction of TURBO data on tape. Also, the policy of data filtering and reduction (stripping) has been modified in order to mitigate the size on disk of the derived datasets, and to cope with the increase in the stripping rate for real events.

Disk provisions at Tier-2s hosting disk storage (T2D) 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. The total disk space available at T2D sites at the end of 2016 amounted to 4.2PB. The T2D ensemble is thus equivalent to a large LHCb Tier1 site.

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. Clouds and data centers based on virtual machines or containers have been also successfully used, 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.

3. Resource usage in 2016

Table 1 shows the resources pledged for LHCb at the various tier levels for the 2016 period.

2016	CPU (kHS06)	Disk (PB)	Tape (PB)
Tier0	51	7.6	20.6
Tier1	165	15.9	35.0
Tier2	88	2.7	
Total WLCG	**304**	**26.2**	**55.6**

Tab. 1: LHCb 2016 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 RAL computing center. The CNAF contribution is about 40% higher than the pledged one. This achievement has been possible owing to great stability of the center, leading to maximal efficiency in the overall exploitation of the resources.

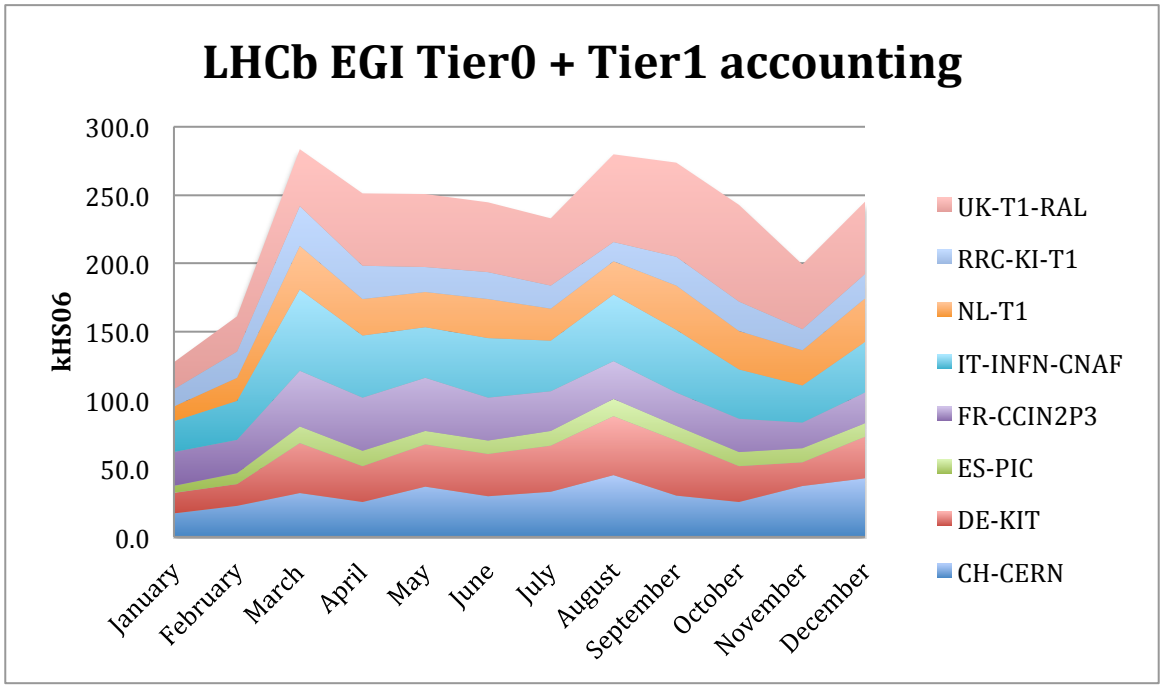


Fig. 2: Monthly CPU work provided by the Tier-0 and Tier-1s to LHCb during 2016.

<Power>	Used (kHS06)	Pledge (kHS06)
CH-CERN	31.4	51
DE-KIT	29.0	24.4
ES-PIC	10.1	9.5
FR-CCIN2P3	28.7	23.0
IT-INFN-CNAF	39.1	28.1
NL-T1	25.5	17.1
RRC-KI-T1	19.1	16.4
UK-T1-RAL	49.9	46.8
Total	**232.8**	**216.3**

Tab. 2: Average CPU power provided by the Tier-0 and the Tier-1s to LHCb during 2016.

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 17.7 PB. Of these, 11.5 PB were due to RAW data taken since June 2016. The rest was due to RDST (3.0 PB) and ARCHIVE (3.2 PB), the latter due to the archival of Monte Carlo productions, stripping cycles, and new Run2 data. The total tape occupancy as of December 31st 2016 is 39.4 PB, 21.5 PB of which are used for RAW data, 8.7 PB for RDST, 9.2 PB for archived data. This is quite in line with the original request of 40.8PB, which should be reached by the end of the WLCG year by an expected growth of 1PB in the ARCHIVE space. The total tape occupancy at CNAF is 5.7PB.

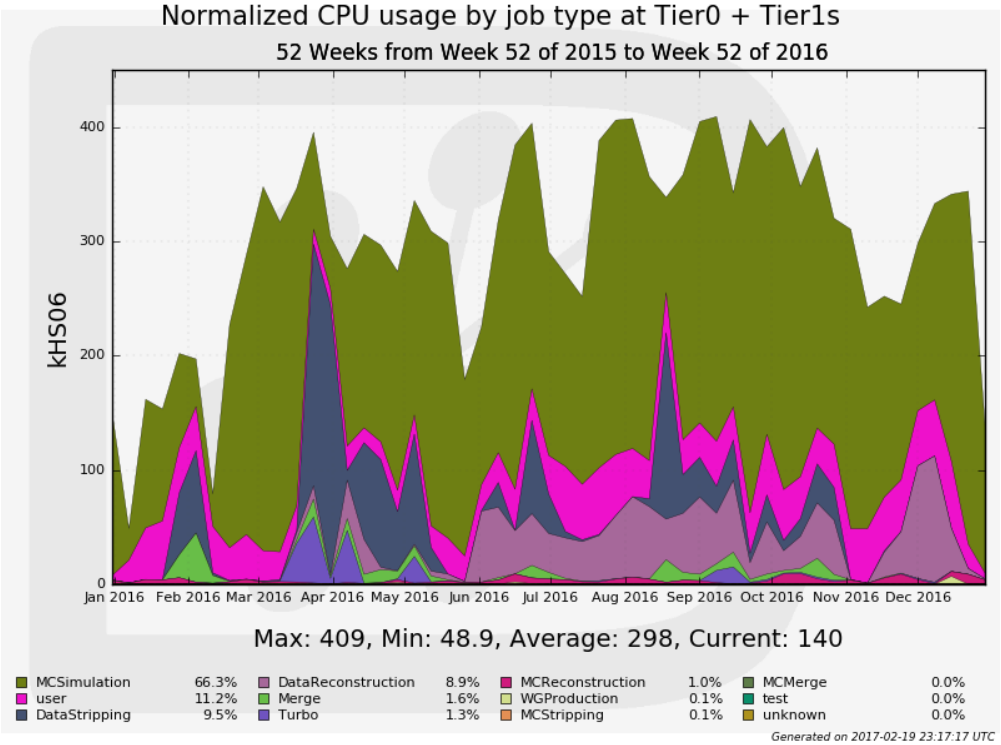


Fig. 3: Usage of LHCb resources at Tier-0 and Tier-1s during 2016. The plot shows the normalized CPU usage (kHS06) for the various activities.

Disk (PB)	CERN	Tier1s	CNAF	GRIDKA	IN2P3	PIC	RAL	RRCKI	SARA
LHCb accounting	**4.40**	**13.51**	**2.68**	**1.92**	**1.88**	**0.84**	**3.61**	**1.11**	**1.46**
SLS T0D1 used	4.59	12.99	2.71	1.87	1.88	0.84	3.09	1.11	1.47
SLS T0D1 free	0.47	1.75	0.32	0.27	0.23	0.08	0.45	0.15	0.25
SLS T1D0 (used+free)[1]	1.25	1.13	0.45	0.14	0.03	0.01	0.38	0.09	0.03
SLS T0D1+T1D0 total[1]	**6.32**	**15.86**	**3.48**	**2.28**	**2.14**	**0.94**	**3.92**	**1.35**	**1.75**
Pledge '16	***7.60***	***15.90***	***3.15***	***2.50***	***2.23***	***0.97***	***4.05***	***1.26***	***1.74***

Tab. 3: Situation of disk storage resource usage as of December 31st 2016, available and installed capacity, and 2016 pledge. The contribution of each Tier1 site is also reported.[2]

Table 3 shows the situation of disk storage resources at the Tier-0 and Tier-1s at the end of December 2016. The real data DST space increased in the first part of the year due to a stripping cycle, then from June onwards due to the 2016 data taking. Stripping was stopped in July because of its high throughput, and resumed in the second half of August, with a sudden increase in disk occupancy due to the backlog accumulated in the meantime. Following periodical analyses of dataset popularity, it was possible to recover about 3.5PB of disk space throughout 2016 by removing old datasets. The disk space available at CNAF at the end of 2016 is 3.48PB, 10% above the pledge of 3.15PB.

[1] This total includes disk visible in SRM for tape cache. It does not include invisible disk pools for dCache (stage and read pools).

In summary, the usage of computing resources in 2016 has been quite smooth for LHCb.

Simulation is 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.

The LHC live time expected in 2016 was significantly higher than initially expected. Also, the stripping throughput and the TURBO event size are larger than initially planned. Measures have been taken in order to reduce both to acceptable levels. This, paired with more aggressive changes in the computing model, such as the reduction of the number of datasets copies, the parking of a fraction of TURBO data and, last but not least, the removal of unused datasets following a data popularity analysis, allow LHCb to stay within the available disk resources in 2016.

The INFN Tier1 at CNAF provided large and reliable amounts of computing resources for LHCb, resulting in the second most used Tier1 site.

4. Expected resource growth

In terms of CPU requirements, Tab. 4 presents for the different activities the CPU work estimates for 2017-2019. The last row shows the power averaged over the year required to provide this work.

CPU Work in WLCG year (kHS06.years)	**2017**	**2018**	**2019**
Prompt Reconstruction	42	49	0
First pass Stripping	17	20	0
Full Restripping	38	0	61
Incremental (Re-)stripping	10	10	15
Processing of heavy ion collisions	4	38	0
Simulation	269	342	411
VoBoxes and other services	4	4	4
User Analysis	26	32	38
Total Work (kHS06.years)	**410**	**495**	**529**

Tab. 4: Estimated CPU work needed for the various LHCb activities in 2017-2019.

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. The CPU work required at CNAF would correspond to about 20% of the total CPU requested at Tier1 sites.

CPU Power (kHS06)	**2017**	**2018**	**2019**
Tier 0	67	81	86
Tier 1	207	253	271
Tier 2	116	141	152
Total WLCG	**390**	**475**	**509**
HLT farm	10	10	10
Yandex	10	10	10
Total non-WLCG	**20**	**20**	**20**
Grand total	**410**	**495**	**529**

Tab. 5: CPU power requested at the different Tiers in 2017-2019.

Tables 6 and 7 present, for the different data classes, the forecast total disk and tape space usage at the end of the years 2017-2019. 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. The storage resources required at CNAF would be about 20% of those requested for Tier1 sites. 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. The mitigation measures described above were put in place to keep the growth in the tape storage requirement to a manageable level.

Disk storage usage forecast (PB)	**2017**	**2018**	**2019**
Stripped Real Data	19.3	16.7	22.6
TURBO Data	3.6	5.3	5.3
Simulated Data	9.3	13.1	15.5
User Data	1.2	1.2	1.2
Heavy Ion Data	2.6	4.2	4.2
RAW and other buffers	1.1	1.2	1.2
Other	0.6	0.6	0.6
Total	**37.7**	**42.3**	**50.7**

Tab. 6: Breakdown of estimated disk storage usage for different categories of LHCb data

Tape storage usage forecast (PB)	**2017**	**2018**	**2019**
Raw Data	31.8	47.9	47.9
RDST	11.1	15.4	15.4
MDST.DST	6.4	6.8	6.8
Heavy Ion Data	2.3	3.7	3.7
Archive	17.0	24.1	31.3
Total	**68.6**	**97.9**	**105.1**

Tab. 7: Breakdown of estimated tape storage usage for different categories of LHCb data

Disk (PB)	**2017**	**2018**	**2019**
Tier0	10.9	12.0	14.6
Tier1	22.1	24.5	29.0
Tier2	4.7	5.8	7.1
Total	**37.7**	**42.3**	**50.7**

Tab. 8: LHCb disk request for each Tier level in 2017-2019. Countries hosting a Tier-1 can decide what is the most effective policy for allocating the total Tier-1+Tier-2 disk pledge.

Tape (PB)	**2017**	**2018**	**2019**
Tier0	25.2	36.4	37.7
Tier1	43.4	61.5	67.4
Total	**68.6**	**97.9**	**105.1**

Tab. 9: LHCb tape request for each Tier level in 2017-2019

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 2019. It has been shown that CNAF has been in 2016 the second most important LHCb computing centre in terms of CPU work 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. 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 LHCf experiment

L. Bonechi[1], O. Adriani[2,1], E. Berti [2,1], M. Bongi[2,1], R. D'Alessandro[2,1], A. Tiberio[2,1] and A. Tricomi[3,4] for the LHCf Collaboration

[1] INFN, Sesto Fiorentino - Florence, IT
[2] Department of Physics, University of Florence, Florence, IT
[3] Department of Physics, University of Catania, Catania, IT
[4] INFN, Catania, IT

E-mail: Lorenzo.Bonechi@fi.infn.it

Abstract. The LHCf experiment is dedicated to the measurement of very forward particle production in the high energy hadron-hadron collisions at LHC, with the aim of improving the cosmic-ray air shower developments models. Most of the simulations of particle collisions and detector response are produced exploiting the resources available at CNAF. The role of CNAF and the main recent results of the experiment are discussed in the following.

1. Introduction

The LHCf experiment is dedicated to the measurement of very forward particle production in the high energy hadron-hadron collisions at LHC. The main purpose of LHCf is improving the performance of the hadronic interaction models, that are one of the important ingredients of the simulations of the Extensive Air Showers (EAS) produced by primary cosmic rays. Since 2009 the LHCf detector has taken data in different configurations of the LHC: p-p collisions at center of mass energies of 900 GeV, 2.76 TeV, 7 TeV and 13 TeV, and p-Pb collisions at $\sqrt{s} = 5.02$ TeV and 8.16 TeV. The main results obtained in 2016 is shortly presented in the next paragraphs.

2. The LHCf detector

The LHCf detector [1, 2] is made of two independent electromagnetic calorimeters placed along the beam line at 140 m on both sides of the ATLAS Interaction Point, IP1. Each of the two detectors, called Arm1 and Arm2, contains two separate calorimeter towers allowing to optimize the reconstruction of neutral pion events decaying into couples of gamma rays. During data taking the LHCf detectors are installed in the so called r̈ecombination chambers; a place where the beam pipe of IP1 splits into two separate pipes, thus allowing small detectors to be inserted just on the interaction line (this position is shared with the ATLAS ZDC e.m. modules). For this reason the size of the calorimeter towers is very limited (few centimeters). Because of the performance needed to study very high energy particles with the requested precision to allow discriminating between different hadronic interaction models, careful simulations of particle collisions and detectors response are mandatory. In particular , due to the tiny transversal size of the detectors, large effects are observed due to e.m. shower leackage in and out of the calorimeter towers. Most of the simulations produced by the LHCf Collaboration for the study and calibration of the Arm2 detector have been run exploiting the resources made available at

CNAF. Many simulations are still on-going to be used for the analysis of the most recent data taken for p-p collisions at 13 TeV and p-Pb collisions at $\sqrt{s} = 8.16\,\text{TeV}$.

3. Results obtained in 2016

2016 was an year full of successfull data taking runs for the experiment. In June, the most important design goal of the experiment was achieved, collecting a statistically relevant data set for p-p collision at 13 TeV, the highest energy currently available at an accelerator. A dedicated run at low luminosity, mandatory to reduce radiation damage for the detector and to improve the quality of data at very high rapidity, has been performed in a real collaborative spirit with the ATLAS Collaboration, implementing a combined data taking that allows a simultaneous study of the forward (LHCf) and central (ATLAS) regions of the p-p interaction. The scope of this study had been demonstrated in the previous months by implementing dedicated simulations that had run on the PC farm at CNAF. At the end of the summer run the detector was removed from underground and brought to the SPS experimental area for a test with electron, proton and muon beams. The results of the test, combined with extensive simulations performed at CNAF, allowed to make a post-run calibration of the LHCf detectors. The first results of this run, concerning the study of photon production at extreme rapidity were the main topic of two PhD Theses [3, 4] and will be published soon. During 2016 the data analysis for neutral pions produced in different configurations of the LHC collisions, finalized in 2015, was finally published [5]. Between September and October 2016 the only Arm2 detector was re-installed in the LHC tunnel for a run with p-Pb collisions at $\sqrt{s} = 8.16\,\text{TeV}$. Also in this case the combined LHCf-ATLAS data taking strategy was followed and a common data set is now available to both the collaborations for future analysis and joint publications. Concerning the purpose of the LHCf Collaboration, the information on the activity in the central ATLAS detector, event by event, can be exploited to separate single diffractive (SD), double diffractive (DD) and ultra-peripheral collisions (UPC) from the non-diffractive (ND) collisions, thus providing an important missing information to the developers of hadronic interaction models. The evaluation of the scope of this measurement has been performed with dedicated simulations between January and March 2016 using the resources available at CNAF. The results of these simulations are reported in the Letter of Intent presented to the LHC Committee in March and successively approved by the LHC Research Board [6].

4. LHCf simulations and data processing

In 2016, the CNAF resources were mainly used by LHCf for mass production of MC simulations needed for the analysis of LHC data relative to proton-proton collisions at $\sqrt{s} = 13\,\text{TeV}$. Two kinds of simulations were needed: the first one was produced making use of the COSMOS and EPICS libraries, the second one making use of the CRMC toolkit. In both cases we used the most common generators employed in cosmic ray physics. For the second group only secondary particles produced by collisions were considered, whereas for the first group transport through the beam pipe and detector interaction were simulated as well. For this purpose, all this software was at first installed on the CNAF dedicated machine, then we performed some debug and finally we interactively run some test simulations. All over the year we continuously submitted simulation jobs on the CNAF PC farm, employing about 1887 HS06 of CPU time and using almost all the amount of disk space allocated to store the output (40 TB). In order to optimize the usage of resources, simulations production was shared between Italian and Japanese side of the collaboration. For this reason, the machine was used as well to transfer data from/to Japanese server.

In addition to simulations activity, CNAF resources were important for data analysis, both for experimental and simulation files. This work required to apply all reconstruction processes, from binary data up to a ROOT file containing all relevant quantities reconstructed from detector

information. For this purpose, LHCf analysis software was installed, debugged and continuously updated on the system. Because the reconstruction of a single file can take several hours and the number of files to be reconstructed is large, the usage of the queue dedicated to LHCf was necessary to accomplish this task. ROOT files were then transferred to local PCs in Firenze, in order to have more flexibility on the final analysis steps, that does not require long computing time.

References

[1] O. Adriani *et al.*, JINST **3**, S08006 (2008)
[2] O. Adriani *et al.*, JINST **5**, P01012 (2010)
[3] E. Berti, "Measurement of the energy spectra relative to neutrons produced at very small angle in $\sqrt{s}$=13 TeV proton-proton collisions using the LHCf Arm2 detector", March 2017, University of Florence, Florence, Italy
[4] A. Tiberio, "Study of the very forward electro-magnetic component produced in proton-proton collisions at $\sqrt{s}$=13 TeV with the LHCf experiment", March 2017, University of Florence, Florence, Italy
[5] O. Adriani *et al.*, Phys.Rev. D94 (2016) no.3, 032007, CERN-PH-EP-2015-201, DOI: 10.1103/PhysRevD.94.032007
[6] CERN-LHCC-2016-003 / LHCC-I-027, 01/03/2016

The NA62 experiment at CERN

A. Sergi, on behalf of the NA62 collaboration

E-mail: antonino.sergi@cern.ch

Abstract. The rare decays are theoretically clean processes excellent to make tests of new physics at the highest scale complementary to LHC. The NA62 experiment at CERN SPS aims to collect of the order of 100 events in two years of data taking, keeping the back- ground less than 20% of the signal.

1. Introduction

Among the flavour changing neutral current K and B decays, the $K \to \pi\nu\bar{\nu}$ decays play a key role in the search for new physics through the underlying mechanisms of flavour mixing. These decays are strongly suppressed in the SM (the highest CKM suppression), and are dominated by top-quark loop contributions. The SM branching ratios have been computed to high precision with respect to other loop-induced meson decays: $\mathrm{BR}(K^+ \to \pi^+\nu\bar{\nu}) = 8.22(75) \times 10^{-11}$ and $\mathrm{BR}(K_L \to \pi^0\nu\bar{\nu}) = 2.57(37) \times 10^{-11}$; the uncertainties are dominated by parametric ones, and the irreducible theoretical uncertainties are at a $\sim 1\%$ level [1]. The theoretical cleanness of these decays remains also in certain new physics scenarios. Experimentally, the $K^+ \to \pi^+\nu\bar{\nu}$ decay has been observed by the BNL E787/E949 experiments, and the measured branching ratio is $\left(1.73^{+1.15}_{-1.05}\right) \times 10^{-10}$ [2]. The achieved precision is inferior to that of the SM expectation.

The main goal of the NA62 experiment at CERN is the measurement of the $K^+ \to \pi^+\nu\bar{\nu}$ decay rate at the 10% precision level, which would constitute a significant test of the SM. The experiment is expected to collect about 100 signal events in two years of data taking, keeping the systematic uncertainties and backgrounds low. Assuming a 10% signal acceptance and the SM decay rate, the kaon flux should correspond to at least 10^{13} K^+ decays in the fiducial volume. In order to achieve a small systematic uncertainty, a rejection factor for generic kaon decays of the order of 10^{12} is required, and the background suppression factors need to be measured directly from the data. In order to achieve the required kaon intensity, signal acceptance and background suppression, most of the NA48/NA62 apparatus used until 2008 was replaced with new detectors. The CERN SPS extraction line used by the NA48 experiment is capable of delivering beam intensity sufficient for the NA62. Consequently the new setup is housed at the CERN North Area High Intensity Facility where the NA48 was located. The decay in flight technique will be used; optimisation of the signal acceptance drives the choice of a 75 GeV/c charged kaon beam with 1% momentum bite. The experimental setup includes a ~ 100 m long beam line to form the appropriate secondary beam, a ~ 80 m long evacuated decay volume, and a series of downstream detectors measuring the secondary particles from the K^+ decays in the fiducial decay volume.

The signal signature is one track in the final state matched to one K^+ track in the beam. The integrated rate upstream is about 800 MHz (only 6% of the beam particles are kaons, the others

being mostly π^+ and protons). The rate seen by the detector downstream is about 10 MHz, mainly due to K^+ decays. Timing and spatial information are required to match the upstream and downstream tracks. Backgrounds come from kaon decays with a single reconstructed track in the final state, including accidentally matched upstream and downstream tracks. The background suppression profits from the high kaon beam momentum. A variety of techniques are employed in combination in order to reach the required level of background rejection. They can be schematically divided into kinematic rejection, precise timing, highly efficient photon and muon veto systems, and precise particle identification systems to distinguish π^+, K^+ and positrons. The above requirements drove the design and the construction of the subdetector systems.

The main NA62 subdetectors are: a differential Cherenkov counter (CEDAR) on the beam line to identify the K^+ in the beam; a silicon pixel beam tracker; guard-ring counters surrounding the beam tracker to veto catastrophic interactions of particles; a downstream spectrometer composed of 4 straw chambers operating in vacuum; a RICH detector to identify pions and muons; a scintillator hodoscope; a muon veto detector. The photon veto detectors include a series of annular lead glass calorimeters surrounding the decay and detector volume, the NA48 LKr calorimeter, and two small angle calorimeters to provide hermetic coverage for photons emitted at close to zero angle to the beam. The design of the experimental apparatus and the R&D of the new subdetectors have been completed. The experiment started collecting physics data in 2015, and since 2016 is fully commissioned and in its production phase.

2. NA62 computing model and the role of CNAF

NA62 raw data consist in custom binary files, collecting data packets directly from the DAQ electronics, after a minimal overall formatting; there is a one to one correspondence between files and spills from the SPS. Data contains up to 16 different level-0 trigger streams, for a total maximum bandwidth of 1 MHz, which are filtered by software algorithms to reduce the output rate to less than 50kHz. Raw data is stored on CASTOR and promptly calibrated and reconstructed, on a scale of few hours, for data quality monitoring using the batch system at CERN and EOS. Near-line fast physics selection for data quality, off-line data processing and analysis is currently performed using only CERN computing facilities.

Currently NA62 exploits the GRID only for Monte Carlo productions, under the management of the UK GRID-PP collaboration members; in 2016 CNAF resources have been used as one fo the GRID sites that serve NA62VO.

References

[1] J. Brod, M. Gorbahn and E. Stamou, Phys. Rev. **D83**, 034030 (2011).
[2] A.V. Artamonov *et al.*, Phys. Rev. Lett. **101** (2008) 191802.

The PADME Experiment at INFN CNAF

E. Leonardi
INFN Sezione di Roma, P.le Aldo Moro, 2 – 00185 Roma, Italy

E-mail: emanuele.leonardi@roma1.infn.it

Abstract. The PADME experiment at the DAΦNE Beam-Test Facility (BTF) in Frascati is designed to detect dark photons produced in positron on fixed target annihilations decaying to dark matter ($e^+e^-\rightarrow\gamma A'$) by measuring the final state missing mass. The collaboration will complete the design and construction of the experiment by the end of 2017 and will start data taking in April 2018. This report, after a brief introduction of the experiment and its physics goals, describes the PADME data analysis model and its use of computing facilities at INFN CNAF.

1. Introduction

The long standing problem of reconciling the cosmological evidence of the existence of dark matter with the lack of any clear experimental observation of it, has recently revived the idea that the interaction of the new particles with the Standard Model (SM) gauge fields is not direct but occurs through "portals", connecting our world with new "secluded" or "hidden" sectors. One of the simplest models introduces a single U(1) symmetry, with its corresponding vector boson, called Dark Photon or A'. In the most general scenario, the existence of dark sector particles with a mass below that of A' is not excluded: in this case, so-called "invisible" decays of the A' are allowed. Moreover, given the small coupling of the A' to visible SM particles, which makes the visible rates suppressed by ε^2 (ε being the reduction factor of the coupling of the dark photon with respect to the electromagnetic one), it is not hard to realize a situation where the invisible decays dominate. There are several studies on the searches of A' decaying into dark sector particles, recently summarized in [1][2].

At the end of 2015 INFN formally approved a new experiment, PADME (Positron Annihilation into Dark Matter Experiment) [3][4], to search for invisible decays of the A'. Aim of the experiment is to detect the non-SM process $e^+e^-\rightarrow\gamma A'$, with A' undetected, by measuring the final state missing mass, using a 550 MeV positron beam from the improved Beam-Test Facility (BTF) of the DAΦNE Linac at the INFN Frascati National Laboratories (LNF) [5]. The collaboration will complete the design and construction of the experiment by the end of 2017 and will collect $O(10^{13})$ positrons on target in two years starting in April 2018, with the goal of reaching a $\varepsilon\sim10^{-3}$ sensitivity up to a dark photon mass of $M_{A'} \sim 24$ MeV/c^2.

The experiment, shown in figure 1, is composed of a thin active diamond target, to measure the average position and the intensity of the positrons during a single beam pulse; a set of charged particle veto detectors immersed in the field of a 0.5 Tesla dipole magnet to detect positrons losing their energy due to Bremsstrahlung radiation; and a calorimeter made of BGO crystals, to measure/veto final state photons. As the rate of Bremsstrahlung photons in its central region is too high, the calorimeter has a hole covered by a faster photon detector, the small angle calorimeter (SAC). Finally, a silicon pixel detector measures the time and spatial distributions of the outgoing beam. The

apparatus is inserted into a vacuum chamber, to minimize unwanted interactions of primary and secondary particles that might generate extra photons. The maximum repetition rate of the beam pulses from the DAΦNE Linac is 50 Hz.

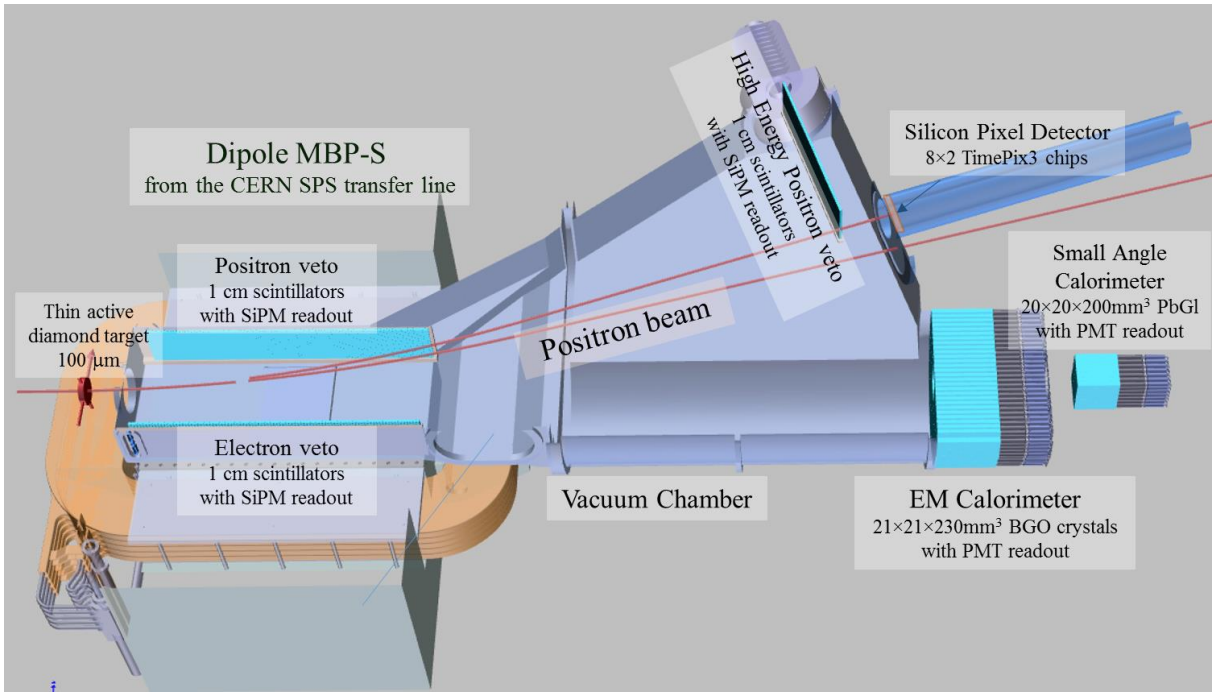


Figure 1. The PADME experiment as seen from above. The positron beam travels from left to right.

2. Data analysis model

The expected total data rate in output from the on-line system is of the order of 10 MB/s, corresponding to roughly 300 TB of data for the expected 10^{13} e^+ on target full statistics. This estimate assumes that all ADC samples from each non-zero suppressed channel are written to the final raw data file. Table 1 shows the contribution of each detector to the final data set.

Table 1. Data rate output from the PADME DAQ system.

Detector	ADC channels	DAQ output MB/s	Raw data GB/d	Total RAW data TB	RAW event size kB/event
Target	32	1.26	109	40	24.6
Calorimeter	616	2.70	234	85	52.8
SAC	49	1.08	93	34	21.1
E veto	96	0.60	51	19	11.6
P veto	96	1.31	113	41	25.6
HEP veto	32	1.15	100	36	22.5
Pixel	Special	1.50	130	47	29.3
Total	921	9.60	830	303	187.5

To optimize the experimental layout in terms of signal acceptance and background rejection, the full experiment was modelled with the GEANT4 simulation package [6][7]. Production of MC events will start in 2017 to continue for the whole duration of the experiment, thus producing an amount of simulated data of the same order of the real data.

The simulation of MC events and the reconstruction and analysis of both real and simulated data will require a substantial amount of CPU power which will be provided by GRID-based resources. In

2017 INFN Committee 1 financed 1000 HEP-SPEC of CPU power for MC event production which are being used to create the initial kernel of the PADME Tier0 site at LNF.

3. CNAF resources for PADME

3.1. Tape Library

As all data produced by the experiment will need to be stored to tape, the collaboration decided to use the INFN CNAF tape library for long term storage of the data. This tape library guarantees all necessary resources, both in term of storage capacity, data access speed, and data conservation. In addition, CNAF and LNF are connected via a high speed link which allows the transfer of all data produced at the Tier0 in real time, thus reducing the requirements for local disk buffers.

At the end of 2016 a 100 TB tape pool was allocated for PADME. This pool was initially used to store all data collected during the 2015 and 2016 test beams, for a total of roughly 3 TB of data. In this occasion the data transfer proceeded manually via a dedicated intermediate disk buffer. Since the beginning of 2017 the PADME tape pool is accessible from all GRID nodes via a StoRM interface and in March 2017 the collaboration was able to successfully complete a test of the full MC data production chain, running jobs on the PADME Tier0 Worker Nodes (WN) at LNF and copying the simulated events files directly to the tape library at CNAF.

3.2. GRID services

A complete GRID-based computing model requires some basic services to guarantee the correct authentication and access to the distributed resources and the automatic distribution of all needed software packages to the sites where the jobs should run.

In June 2016 CNAF created the “vo.padme.org” virtual organization (VO) which is currently hosted on the voms2.cnaf.infn.it VOMS server. This VO was registered on the EGI infrastructure and is now used for authentication and role management on all GRID resources used by PADME.

In order to easily distribute all the software needed for MC production and data reconstruction and analysis to all WNs used by the collaboration, CNAF created a virtual CVMFS server, cvmfs-padme.cnaf.infn.it, hosting the /cvmfs/padme.infn.it area. Here PADME software managers can install all the experiment’s software packages which are then automatically made available to all GRID sites hosting PADME resources.

4. Conclusions

The PADME experiment will search for the dark photon A' with mass up to 24 MeV in the annihilation process $e^+e^- \rightarrow \gamma A'$, with A' undetected, using the Beam-Test Facility of the DAΦNE Linac at the INFN Frascati National Laboratories. Data taking will start in 2018 and will collect $O(10^{13})$ e^+ on target in 2 years. Starting at beginning of 2016, CNAF provided several important computing services to the PADME collaboration, namely access to resources on their tape library for long term data storage and the hosting of the VOMS and CVMFS services. In addition to this, CNAF personnel always provided us with technically savvy information and advice, effectively helping the PADME collaboration in the delicate phase of designing and implementing a computing model, for which we are truly grateful.

5. References

[1] Raggi M and Kozhuharov V 2015 Results and perspectives in dark photon physics *Riv. Nuovo Cimento* **38** (10) 449-505

[2] Alexander J *et al.* 2016 Dark Sectors 2016 workshop: community report *arXiv:1608.08632 [hep-ph]*

[3] Raggi M and Kozhuharov V 2014 Proposal to search for a dark photon in positron on target collisions at DAΦNE Linac *Adv. High Energy Phys.* **2014** 959802

[4] Raggi M, Kozhuharov V and Valente P 2015 The PADME experiment at LNF *EPJ Web of Conferences* **96** 01025
[5] Valente P *et al.* 2016 Linear accelerator test facility at LNF conceptual design report *arXiv:1603.05651 [physics.acc-ph]*
[6] Agostinelli S *et al.* 2003 Geant4 – A simulation toolkit *Nucl. Instrum. Methods A* **506** 250-303
[7] Leonardi E, Kozhuharov V, Raggi M and Valente P 2017 GEANT4-based full simulation of the PADME experiment at the DAΦNE BTF *Proceedings of the CHEP 2016 conference tbp in the Journal of Physics Conference Series (JPCS)*

The PAMELA experiment

A. Bruno, F. Cafagna
on behalf of the PAMELA collaboration
INFN and University of Bari, Bari, IT

Abstract. The PAMELA cosmic ray detector was launched on June 15^{th} 2006 on board the Russian Resurs-DK1 satellite, and during ten years of nearly continuous data-taking it has observed very interesting features in cosmic rays. 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 2016

The PAMELA satellite experiment has reported 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 [1]. 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). 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. In addition, features of the penumbra region around the geomagnetic cutoff were investigated in detail, including the variations of the geomagnetic shielding during magnetospheric storms [2].

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 [1].

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 [3].

Finally, PAMELA has measured the cosmic-ray hydrogen and helium (1H, 2H, 3He, 4He) isotopic composition [4]. The rare isotopes 2H and 3He in cosmic rays are believed to originate mainly from the interaction of high energy protons and helium with the galactic interstellar medium. The isotopic composition was measured between 100 and 1100 MeV/n for hydrogen and between 100 and 1400 MeV/n for helium isotopes using two different detector systems over the 23^{rd} solar minimum from July 2006 to December 2007.

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 (Fortran-77) based on numerical integration methods, 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] Bruno, A., et al., Geomagnetically trapped, albedo and solar energetic particles: Trajectory analysis and flux reconstruction with PAMELA, Advances in Space Reasearch (Available online), 2016; http://dx.doi.org/10.1016/j.asr.2016.06.042.

[2] Adriani, O., et al., PAMELA's measurements of geomagnetic cutoff variations during the 14 December 2006 storm, Space Weather, vol. 14, Issue 3, 210-220 (2016).

[3] Adriani, O., et al., Time dependence of the electron and positron components of the cosmic radiation measured by the PAMELA experiment between July 2006 and December 2015, Phys. Rev. Lett. 116, 241105 (2016).

[4] Adriani, O., et al., Measurements of cosmic-ray hydrogen and helium isotopes with the PAMELA experiment, ApJ vol. 818, issue 1, pag 68 (2016).

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. During last years, 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. The XENON100 experiment 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]. During 2011, while XENON100 was still taking data, the XENON1T project started. It is the largest dual phase (LXe/GXe) Xe-based detector ever realized and it is now acquiring data 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. Two meshes enclose the TPC: 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 difference 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 it. With respect to its predecessor, XENON1T uses a larger mount of LXe: about 3.3 tonnes 2 of which represent the sensitive volume available for the WIMP interactions. XENON1T 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 is required in order to choose the materials with the lowest contaminations, and a MC study, through simulations

with the GEANT4 toolkit, in order to optimize the detector design and to evaluate the expected background [2]. Due to the large amount of simulations required to perform that research, the GRID is the most appropriate facility to be used.

2. XENON1T

For what concern the data flow, the XENON1T experiment uses a DAQ machine hosted in the XENON1T service building underground to acquire data. The DAQ rate in DM mode is 1 TB/day, while in calibration mode it is larger: ∼7 TB/day. The raw data are copied into rucio, a data handling system. There are several rucio endpoints or rucio storage elements (RSE) around the world, including LNGS, NIKHEF, Lyon and Chicago. The raw data are replicated in at least two positions and there is a tape backup in Stockholm with 5.6 PB in total. When the data have to be processed, they are first copied onto Chicago storage then they are processed using the Open Science Grid. The processed data are then copied back to Chicago and become available for the analysis. In addition, for each user there is a home space of 100 GB available on a disk of 10TB. A dedicated server will take care of the data transfer to/from remote facilities. 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) and documents repository on SVN/GIT, the run database, the on-line monitoring web interface, the XENON wiki and GRID UI. CNAF resources have been extensively tested for data simulation (MC with GEANT4 software) using the GRID technology. During 2016 we used about 500 HS06 per day and we are currently using about 30 TB of simulated data for the optimization of the detector design and the background evaluation. It is foreseen to continue to use the GRID to produce simulated data and to store them in the related disk storages. Moreover, we are planning to use the EGI to process the raw data. That will certainly increase the need of CPU and disk space for 2017 with respect to 2016.

3. 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

Advanced Virgo computing at CNAF

C. Lazzaro[1], C. Palomba [2], M. Punturo[3], L. Rei[4], L. Salconi[5], on behalf of the Virgo collaboration

[1] INFN, Padova, IT
[2] INFN, Roma, IT
[3] INFN, Perugia, IT
[4] INFN, Genova, IT
[5] EGO-European Gravitational Observatory, Cascina, Pisa, IT

E-mail: michele.punturo@pg.infn.it

Abstract. Advanced Virgo is a gravitational wave (GW) interferometric detector realised near Pisa, Italy. It is a Michelson interferometer having 3 km long Fabry–Perot cavities in the arms; it is the largest GW detector in Europe, second only to the LIGO detectors (4 km long cavities) in US. Advanced Virgo is entering in function in 2017, after a 5 years long upgrade period, but the Virgo collaboration contributed to the detection of the GW signal emitted by the coalescence of a system of two black holes announced jointly by the LIGO scientific collaboration and by the Virgo Collaboration the 11th of February, 2016[1]. CNAF computing centre has been deeply involved in the data analysis of the first LIGO scientific run (O1).

1. Advanced Virgo computing model

1.1. Data production and data transfer

Advanced Virgo data acquisition system currently writes, in commissioning mode, about 36MB/s of data (so–called "bulk data "). This amount of data is larger with respect to the amount of data expected in science mode, that should be slighly less than 30MB/s (quite larger than the original specifications of 23MB/s). Commissiong bulk data are kept in a few–months circular buffer at the Virgo site and aren't currently transferred to remote repositories. A sub–set of data, named "trend-data ", containg a down-sampling of the bulk data, sizing about few GB/day are transferred periodically to the Virgo remote repositories (CNAF and CCIN2P3 in Lyon). In science mode, different data streams are produced by Advanced Virgo:

- bulk data (up to 36MB/s), stored in the circular buffer at the Virgo site and simultaneously transferred to CNAF and CCIN2P3 using a system developed by EGO IT department based on lcg–tools (toward CNAF)and iRODS (toward CCIN2P3); the measured maximum transfer rate of this system is about 65–70MB/s toward both CNAF and CCIN2P3;
- trend data (few GB/day), transferred periodically using the system above described;
- Virgo–RDS or Reduced Data Set (about 100GB/day), containing the main Virgo channels including the calibrated dark dringe. This set of data will be transferred from Virgo to CNAF and LIGO computing repositories using LDR (LIGO Data Replicator), a tool based on Grid–FTP (and using iRODS to CCIN2P3).

In addition to the above mentioned data fluxes, another stream is currently arriving to CNAF, the LIGO-RDS, containing the reduced set of data produced by the two LIGO detectors and analysed at CNAF.

1.2. Data Analysis at CNAF
The analysis of the LIGO and Virgo data is made jointly by the two collaborations. A GW source can be localised having at least three detectors simoultaneously detecting it and coherent analysis methods have been developed. In 2016 at CNAF only one analysis pipeline has been intensively executed, addressed to the continuous (CW) wave search, but other pipelines have been prepared to be able to run at CNAF via Open Science Grid (OSG).

1.2.1. CW pipeline CNAF has been in 2016 the main computing center for Virgo all-sky continuous wave (CW) searches. The search for this kind of signals, emitted by spinning neutron stars, covers a large portion of the source parameter space and consists of several steps organized in a hierarchical analysis pipeline. CNAF has been mainly used for the "incoherent" stage, based of a particular implementation of the Hough transform, which is the heaviest part of the analysis from a computational point of view. The code implementing the Hough transform has been written in such a way that the exploration of the parameter space can be split in several independent jobs, each covering a range of signal frequencies and a portion of the sky. This is an embarassingly parallel problem, very well suited to be run in a distributed computing environment. The analysis jobs have been run using the EGI (grid) middleware, with input and output files stored under Storm at CNAF, and registered in the LFC file catalogue. Candidate post-processing, consisting of clusterization, coincidences and ranking, and parts of the candidate follow-up analysis have been also carried on at CNAF. Typical Hough transform jobs needs about 1GB of memory, while for candidate post-processing a larger memory (4–6GB) has been requested at the level of JDL job submission scripts. Past year most of the resources have been used to complete an all-sky mock data challenge and to analyze Advanced LIGO O1 data. Overall, in 2016 about 15kHSE06 have been used at CNAF for CW searches, by running $O(10^5)$ jobs, with durations from a few hours to $\sim$2 days.

1.2.2. cWB via OSG The cWB2G is a pipeline for detection of gravitational wave transients without prior knowledge of the signal waveforms, based on a likelihood method.The algorithm performs a time frequency analysis of the data, using wavelet representation, and identifies the events by clustering time frequency pixels with significant excess coherent power. The likelihood statistic is built up as a coherent sum over the responses of different detectors, and estimates the total signal to noise ratio of the GW signal in the network. The pipeline divides the total analysis time into sub-periods to be analysed in parallel jobs, using Condor tools. Starting since 2015 it has been adapted to run on the Open Science Grid; to do that the pipeline has been modified to reproduce the cWB environment setup on the worker nodes, without the constrain to read the user home account during running.

1.3. 2017 outlook
In 2016 Advanced Virgo data analysis used fairly completely its computing power pledge (25 kHS06) at CNAF for the analysis of O1 data; at the end of the year it has been asked a 30% increase of the computing power for a limited time period for specific analysis needs. In the scenario of the computing centres serving in the world the LIGO and Virgo data analysis CNAF occupied, in 2016, the 4^{th}–5^{th} in terms of provided computing resources. For CNAF, in terms of computing resources, with only one data analysis pipeline in execution, Virgo has been the largest consumer after the LHC experiments. In 2017 the analysis of O2 data will require

additional computing resources also thanks to the expected Virgo data contribution. For this reason an increase of the resources devoted to Virgo as been requested at CNAF, achieving the threshold of 30 kHS06 (having planned a large contribution by other computing centres related to the other countries of the Virgo collaboration). Obviously also the usage of the storage will be increased, depending on the duration of the Virgo science run; current expectation will be of the order of 150–200 TB of additional data in 2017.

It is worth to mention that the Virgo software environment is going to be replicated at CNAF with the aim to allow commissioners to use CNAF to analyse data older than the lifetime of the data buffer at the Virgo site. This could propose to CNAF new challenges in providing computing services to Virgo.

References

[1] B P Abbot et al. Phys. Rev. Lett. 116, 061102 (2016)

The Tier 1 and Data center

The INFN Tier-1

L. dell'Agnello
INFN-CNAF, Bologna, IT
E-mail: luca.dellagnello@cnaf.infn.it

1. Introduction

Since 2003, CNAF hosts the Italian Tier-1 for the high-energy physics experiments at the Large Hadron Collider (LHC) in Geneva, ALICE, ATLAS, CMS, and LHCb, providing the resources, support and services needed for all the activities of data storage and distribution, data processing, Monte Carlo production and data analysis. Nowadays, besides the four LHC experiments, the INFN Tier-1 provides services and resources to 30 other scientific collaborations, including BELLE2 and several astro-particle experiments (Tab.1)[1]. Some of these collaborations have started using our computing center during 2016: Cupid, COSMO_WNEXT (this is a computing initiative common to Opera and Euclid experiments), LSPE, DAMPE, LHAASO, FAMU, and NEWCHIM.

Currently, the INFN Tier-1 hosts nearly 1,000 worker nodes for a total amount of about 21,500 computing slots and a power capacity of $\sim$ 204,000 HS06[1]. All the computing resources are centrally managed by a single batch systemand the resources are dynamically allocated according to a fair-share policy, which prevents resources underutilization and user starvation. Also a small ($\sim$ 33 TFlops) HPC cluster with nodes interconnected via Infiniband is available for special applications. CNAF also operates a large storage infrastructure based on industry standards: all disk servers and disk enclosures are interconnected through a dedicated Storage Area Network (SAN) and the data are hosted on several IBM GPFS[2] file systems, typically one per major experiment. These choices allowed the implementation of a completely redundant data access system. Currently there are $\sim$ 21.5 PB of net disk space. Also a tape library, interconnected via a dedicated SAN, is available with, currently, $\sim$ 45 PB of used tape space. The tape system is managed by IBM TSM[3], integrated with GPFS to build a Hierarchical Storage Manager system[4].The disk-servers are connected to the farm through multiples of 10 Gbps links: the aggregate available bandwidth between the farm and the storage is 100 GByte/s. Besides POSIX, the data can be accessed through standard protocols and interfaces, as defined by the WLCG/EGI projects (GridFTP and SRM, XRootD, and WebDAV).

The data center is interconnected to LHCOPN[2] and LHCONE[3] networks with a dedicated link (60 Gbps where 20 Gbps are reserved for LHC ONE) and has access to the General Internet with a 20 Gbps link. An upgrade of the link to LHCOPN and LHCONE to 2x100 Gbps is foreseen in 2017.

[1] CSN 1 is the INFN National Scientific Committee for High Energy Physics experiments at accelerators, CSN 2 is the committee for astro-particle experiments and CSN 3 is for Nuclear Physics experiments at accelerators.
[2] The Large Hadron Collider Optical Private Network (LHCOPN) is a private network interconnecting the LHC Tier-0 and Tier-1 sites. http://lhcopn.web.cern.ch/lhcopn/ for more details.
[3] The Large Hadron Collider Open Network Environment (LHCONE) is a private network interconnecting the LHC Tier-1, Tier-2 and Tier-3 sites. http://lhcone.web.cern.ch/ for more details.

Experiment	CPU (kHS06)	Disk (PB-N)	Tape (PB)
ALICE	29045	3885	5460
ATLAS	46800	4230	10440
CMS	48000	3960	12000
LHCb	28080	3150	7578
LHC Total	**151925**	**15225**	**35478**
Belle2	5000	150	0
CDF	1300	344	4000
KLOE	0	33	1575
LCHf	2000	70	0
NA62	3000	250	0
LHCb Tier2	15840	0	0
CSN 1 Total	**27140**	**847**	**5575**
AMS	9800	1790	510
ARGO	200	320	1000
Auger	3090	615	0
Borexino	1000	144	9
COSMO_WNEXT	1000	42	0
CTA	3700	296	120
CUORE	1400	162	0
Cupid	100	5	5
DAMPE	380	24	0
DarkSide	4000	760	300
Fermi	900	15	40
GERDA	40	25	20
ICARUS	0	0	330
JUNO	1000	130	0
KM3NeT	0	200	200
LHAASO	300	60	0
LSPE	1000	7	0
MAGIC	296	65	150
Opera	200	15	15
PAMELA	650	90	140
Virgo	25000	592	1368
XENON	700	60	0
CSN 2 Total	**54756**	**5417**	**4207**
Agata/GAMMA	0	0	660
FAMU	250	0	0
NEWCHIM/FARCOS	0	0	150
CSN 3 Total	**250**	**0**	**810**
Grand Total	**234071**	**21489**	**46070**
Installed	**195059**	**21489**	**46070**

Table 1: Pledged and installed resources at INFN Tier-1 in 2016 (for the CPU power an overlap factor is applied)

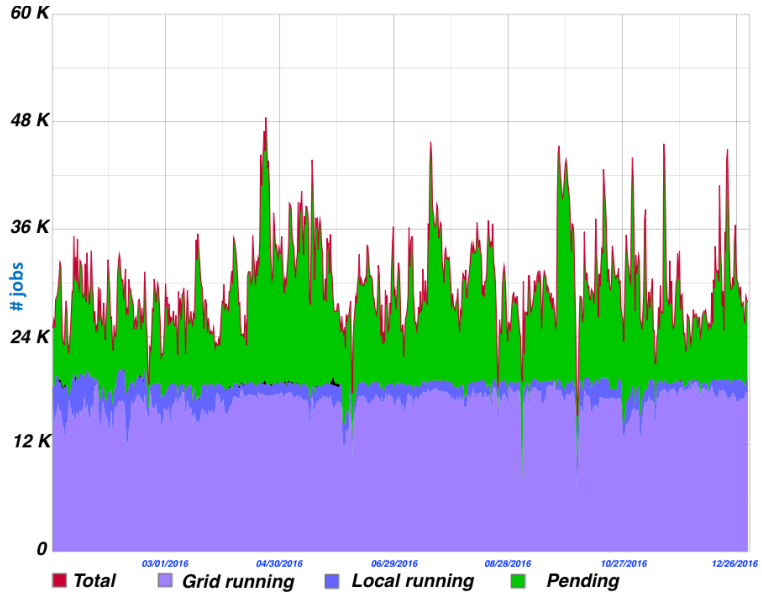


Figure 1: The INFN Tier-1 farm usage in 2016

More details on the activity in the year 2016 can be found in the next Chapters.

1.1. Organization

The Tier-1 staff 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 the users)

At the end of 2016, the staff is composed by about 20 people (including post docs) with a sensible decrease respect to the previous years. To be noticed that one person in the Network unit is affiliated to the GARR Consortium.

2. References

[1] HEPiX Benchmarking Working Group. "HEP-SPEC06 Benchmark" http://w3.hepix.org/benchmarks/doku.php
[2] Schmuck, Frank B., and Roger L. Haskin. "GPFS: A Shared-Disk File System for Large Computing Clusters." FAST. Vol. 2. No. 19. 2002.
[3] Brooks, Charlotte, et al. "IBM tivoli storage management concepts." IBM Redbooks, June (2006).
[4] Ricci, P. P., Bonacorsi, D., Cavalli, A., Dell'Agnello, L., Gregori, D., Prosperini, A., ... & Vagnoni, V. (2012). The Grid Enabled Mass Storage System (GEMSS): the Storage and Data management system used at the INFN Tier1 at CNAF. In Journal of Physics: Conference Series (Vol. 396, No. 4, p. 042051). IOP Publishing.

The INFN-Tier-1: the computing farm

A. Chierici, S. Dal Pra, S. Virgilio, A. Falabella and D. Michelotto
INFN-CNAF, Bologna, IT
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 2016 remained the same as last year, since the 2016 tender suffered some delays that prevented us from being able to host the nodes before the end of the year. To summarize we provided 184.000 HS06 internally and an extra of 20.000 HS06 provided by Bari ReCaS [1] nodes, for a total of 204.000 HS06. The farm is composed mainly by supermicro twin square nodes, with AMD CPUs, then we have Lenovo Blades with Intel v3 CPUs and some legacy ASUS nodes with old Intel E5 CPUs. ASUS nodes will be decommissioned as soon as we will have the new tender installed. Even if half of the farm currently is out of vendor hardware support, we did not experience significant number of hardware failures and were able to provide a constant CPU power along all the year. Anyway modern computing technologies made tremendous improvements and the computing power that was possible to achieve with a rack of machines may be easily be contained in less than a rack, allowing for a significant save in cooling and power costs. For this reason in 2017 we are going to decommission 5 full racks that will be substituted by 2016 tender, providing in 4 racks, a double amount of HS06.

2.1. 2016 tender

The 2016 CPU tender was awarded to Huawei, with a 4U enclosure, model number X6800, each containing 8 nodes, model number XH620v3. The nodes host 2 Intel Xeon E5-2618L v4, the latest generation Intel CPU, integrating 10 cores each. This CPU has a lower power consumption compared to standard 10 cores v4 CPUs, and this will allow us to save 5120 W when the nodes will run at full power. The nodes mount 128GB ram, this means we will be able to provide 3GB per virtual core (2x 10 cores plus hyperthreading). The network connection is a standard 1Gbps, we think that next year we will require 10Gbps, since this speed is becoming common in commodity hardware. The number of nodes we have acquired is 255, and the whole supply will be hosted in just 4 racks.

2.2. New virtualization infrastructure

At the end of 2015 we acquired some new hardware in order to install a new virtualization infrastructure and decommission the old one. The new infrastructure is composed of a Lenovo

blade, hosting 14 blades with 2x Intel E5-2630v3, 128GB ram and 10Gbit connection and an equallogic storage, with 20TB of raid6 disk space. The hardware configuration has been implemented in order to guarantee the maximum fault tolerance: every part of this infrastructure is redundant and in case of any major hardware failure, the support in guaranteed by a next-business-day contract. To support this hardware we have chosen Ovirt [2] software layer, a free version of RedHat Enterprise Virtualization solution, that is already well known among Tier-1 staff and has proven to be reliable and feature-rich. On this new infrastructure, we not only were able to migrate all the virtual machines running on the previous one, but we could also host the provisioning cluster, that was running on less reliable hardware. This service consolidation has allowed us to simplify the management of both the infrastructures, increasing fault tolerance and performance at the same time.

2.3. Quattor Phase-out

During 2016 we gained more and more experience using the new provisioning tools (puppet and foreman) and were able to decommission the legacy tool we used for years, called quattor. Right now we only use quattor on some legacy nodes that will be decommissioned during 2017 and for which a new instantiation is not foreseen or has already been configured.

3. References

[1] ReCaS Bari webpage: https://www.recas-bari.it
[2] Ovirt webpage: https://ovirt.org

Protecting the batch cluster from short job flooding

S. Dal Pra
INFN-CNAF, Bologna, IT
E-mail: stefano.dalpra@cnaf.infn.it

Abstract.
The INFN Tier-1 centre provides computing resources to several competing users and groups. Access to these resources is arbitrated by a batch system according to a fair share policy which aim at granting an average amount of runtime proportional to fixed quotas of the total computing power of the cluster. The fairshare algorithm behaves well and is effective on most cases, however there are special circumstances where it can induce a sensible reduction (10% or greater) of the available computing centre, such as that of users continuously submitting short jobs. Work has been done to study this effect and implement a corrective component for the fair share algorithm. This have been adopted at CNAF, exhibiting good results. In the following, drawbacks of the default fair share algorithm and the implementation of a corrective plugin are explained, and results are presented.

1. Short job flooding

The Tier-1 computing centre always works on saturation, meaning that every free computing resource is expected to get a new job as soon as possible after a previous one ends, and there always are pending jobs, queued in the wait of their turn to be dispatched by the batch system.

Most users make use of automated submitting tools, which try to always keep a number of pending jobs greater than a custom threshold. This ensures that when more resources become free at the site, the user does not run out of jobs to be dispatched. However, when most of the jobs submitted are short, the submission rate must increase, attempting to keep the number of pending jobs higher than the threshold. The fair share algorithm adopted by the LSF batch system, on the other hand, is designed to ensure that each submitter can get enough runtime over time according to their share. The fair share works by continuously updating a *dynamic priority* index for each user: it is increased for those having less runtime than expected, and reduced it for those running more. The combination of these two behaviours, of the submitter tool, and of the fair share can produce an undesidered effect at the batch cluster, that is a "short job flooding", when a user having a high dynamic priority keep submitting only short jobs. The batch system dispatches many of these jobs, but they finish to soon to sum up enough runtime to actually reduce the user priority, and more jobs will be dispatched at the next dispatching time. The overall result is a sort of race between submiter and dispatcher, with the cluster flooded by short jobs. When this happens the computing power of the center is severely reduced, as detailed in the remainder.

1.1. Short jobs and Unusable cputime

If we consider the time intervals δ_k between two consecutive jobs on a computing slot, and their number N over a time window of duration $T = t - t_0$, we can get an estimation of the average number of unusable slots in the cluster over that period:

$$U(t) = \frac{1}{T} \sum_{k=1}^{N} \delta_k \tag{1}$$

.

Of course $U(t)$ grows with N, which in turn is bigger when many short jobs are flowing through the cluster, and with the turnover time δ, whose expected value is measured to be $21 < E[\delta] < 26$ sec, as described in the next section.

2. Turnover time estimation

The average turnover time has been estimated by measuring the time for a new job to start on a full Worker Node after a previous one has finished. There always are jobs waiting to run, hence the measured time actually is a suitable "turnover time" sample. A statistic of these values have been collected from the accounting database over a period of ~ 200 days for different WN models. Results are reported in Fig. 1.

2.1. Fairshare

The algorithm implementing fairshare on LSF works by assigning a *hierarchical dynamic priority* [1] to groups and users. Pending jobs of users with higher DP are dispatched first. The priority U_{prio} of each user (and group) is continuously updated by the following formula:

$$U_{prio} = \frac{U_{share}}{\varepsilon\,\mathrm{CPT} + \alpha\,\mathrm{WCT} + \beta(1 + \mathrm{SLOTS}) + \gamma\,\mathrm{ADJUST}} \tag{2}$$

where CPT and WCT are the overall amount of cputime and WallClockTime collected by the user over a configurable *Fair Share Window* [2], SLOTS is the number of its running jobs at the time of evaluation, and ADJUST is a customizable value, which defaults to zero.

The positive weighting factors $\varepsilon, \alpha, \beta, \gamma$ are set to emphasize WCT over CPT ($0 < \varepsilon \ll \alpha$), so that the dynamic priority is driven by usage of slots over time, no matter how much CPU is being used during that time.

This mechanism ensures to steadily working users a share of the whole computing power proportional to their quota (thus preventing underutilization) whilst providing highest priority to the inactive ones, so that they quickly get resources when restarting their activity.

2.2. Limits of the fairshare formula

The fairshare formula is effective on most cases, however it has poor performances on some usage patterns. For example, let us consider a user starting to submit a continuous flow of short jobs, with a duration of a few seconds each. At first the user has a high DP, thus many of its pending jobs are dispatched; however, they finish very soon. When the formula is evaluated again, SLOTS is likely too small to have impact, and the cumulated WCT is negligible with respect to that of longer jobs. The U_{prio} does not decrease, and the dispatching rate at the next round remains high.

2.3. Automatic submitters and short job flooding

The problem with the dynamic priority described above would have little impact, if the user only had a reasonably limited set of short jobs to be submitted. Things can be worse however,

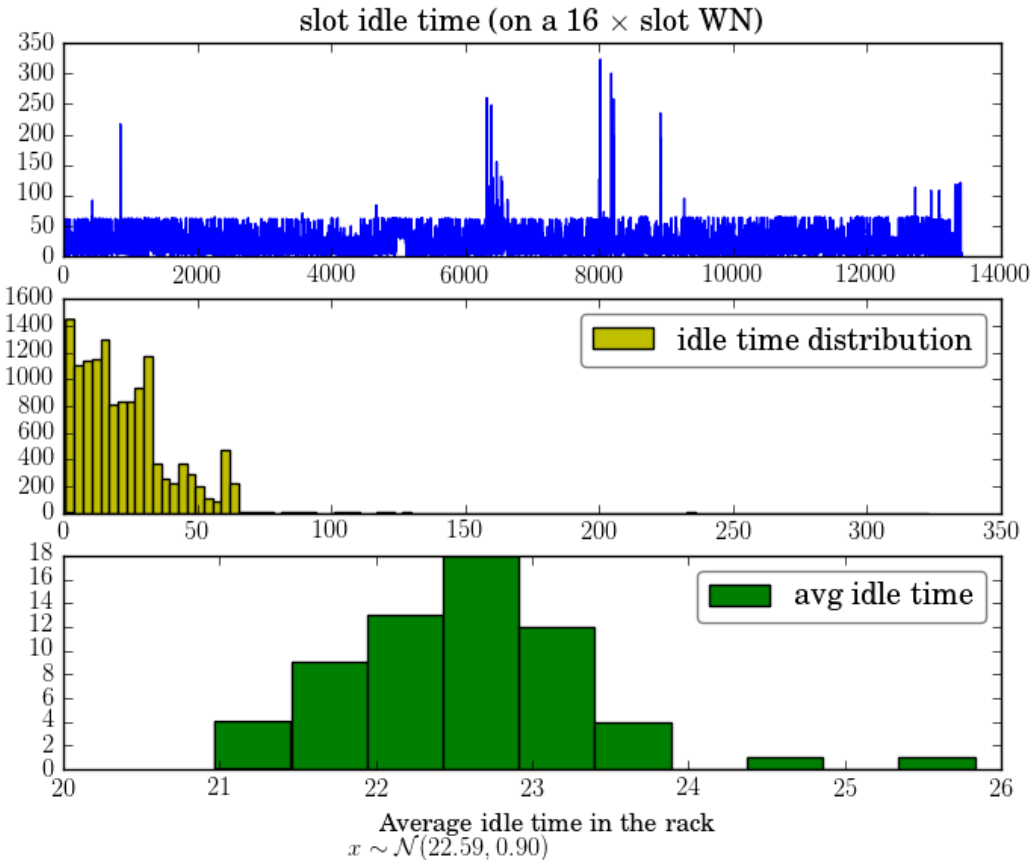


Fig. 1: Turnover time estimation for one model of WN. The upper chart shows the turnover time w for single slot becoming free on a full WN. The data are collected over a period of more than 200 consecutive days. The frequency distribution shows $0 < \delta < 60\,\text{sec}$ and $E[\delta] \simeq 22\,\text{sec}$. The same estimation for all the WNs of the same model yelds $\delta \sim \mathcal{N}(22.6, 0.9)$. The average turnover time for all the WN models gives $21 < E[\delta] < 26$ seconds, with a standard deviation $\sigma_\delta \cong 25\,\text{sec}$.

when the submitter makes use of their own automatic submitter agent. These tools tipically keep submitting new jobs until more than a certain amount is pending. When most of these jobs are sistematically short, a race between submitter and dispatcher begins, producing a "short job flooding" at the cluster. Not only the computing power loss $U(t)$ due to the high number of turnover times grows: the batch system itself can suffer from overload issues becoming less reactive, which further increases the average turn over times itself.

Fig. 2 represents the total number of running jobs at the Tier–1 site during a short job flooding event happened on September 2016. One user starts his activity through a tool that keeps submitting short jobs (few seconds each) until at least 300 are pending. Because of the steadily high value of the dynamic priority of the user, his jobs are dispatched at high rate and new ones are sent as quickly by the submitter. Slots become free in the cluster very soon as most jobs are short. As a result many slots become unavailable to all the active users during the short job flooding. The loss is estimated in the order of two thousand slots, which is $\sim 10\%$ of the total computing power of the centre. An upward peak can be observed just after banning

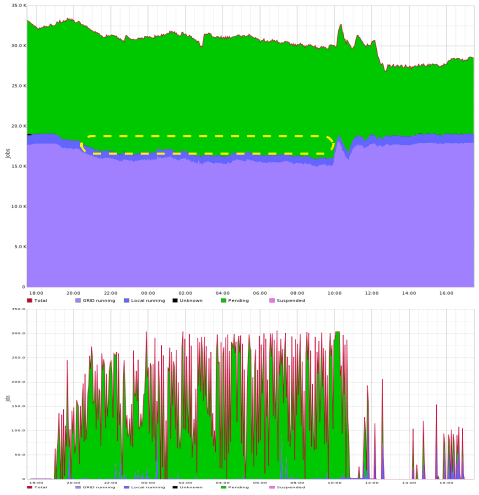


Fig. 2: Unusable slots during a short job flooding (dashed area, Sep. 7, 2016)

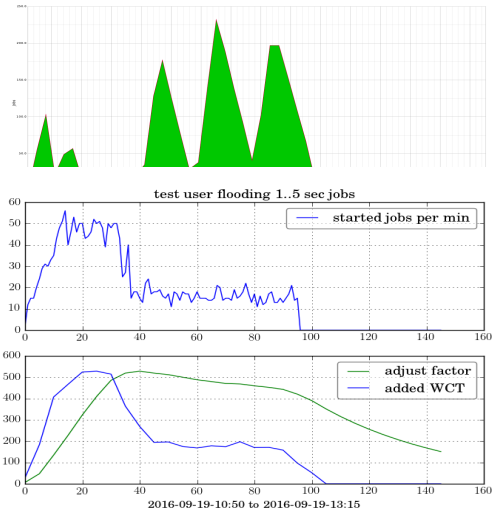


Fig. 3: Testing fairshareadjust with a controlled short job flooding

the user. After activating the mitigating solution described later and re–enabling the user, the workload of the cluster normalizes to a smoother and stable level.

3. Preventing short job flooding

3.1. At user side

One simple way to prevent flooding would require users to adapt their computing model, by putting together several executions into a single job submission, or by limiting the maximum submission rate. Frequently, new individual users adopt a variant of a submitter script similar to the following:

```
while True:
   numj = count_my_jobs()
   if numj < 300: submit_one_job()
   else: sleep(1)
```

which repeatedly submit one job at a time while there are less than a fixed amount, and then sleeps one second before repeating. When short jobs are submitted through such a simple tool, job flooding happens. Effort have been put at INFN Tier-1 to spread awareness through the local user communities, encourage them to aggregate multiple executions in a single job, and to provide template scripts for safer submitters.

3.2. At Batch System side

Taking appropriate measures at submitter side can prevent issues, however a batch cluster cannot only rely on user behaviour and should better be robust against ill–conditioned usage patterns. To achieve this, a customization of the fairshare formula has been implemented.

4. Customizing fairshare

The fairshare formula (2) can be customized by setting the ADJUST to add a "missing WCT quota". This is based on the simple idea of adding a "runtime penalty" to short jobs and treat them "as if" they had run a minimum fixed time. Then, the Dynamic Priority U_{prio} would decrease accordingly. The overall effect would be to act like an automatic rate self–limiter.

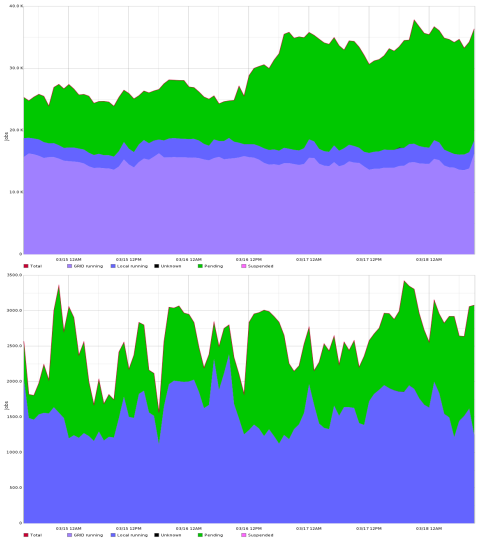


Fig. 4: March, 2016. Running Jobs on the production cluster (upper) during a campaign of short jobs of several days from one user group (lower). The overall level of running jobs is clearly irregular.

Fig. 5: An episode of short jobs (systematically failing because of thecnical problems affecting two user groups). The adjust factor prevents short job flooding. The level of running jobs remaining higher and regular, the batch system does not suffer overloading.

4.1. Implementation

With LSF, the ADJUST from (2) factor is the return value of a `fairshareadjust C` function, which can be customized and activated by an administrator. Particular care must be taken in writing this piece of code, because U_{prio} is updated at every scheduling cycle, for each user and group known to the batch system, thus the function is evaluated very frequently. For this reason the `fairshareadjust` function should be as fast as possible. Moreover, the information needed to compute the adjust factor (recently finished jobs per user and their duration) are not directly available from inside the function. Because of this, needed parameters are retrieved and updated externally every three minutes. The number $N_s(t)$ of short jobs finished over the last interval per user and their runtime penalty $T_s(t)$ are computed by a python script. The Adjust factor for each active user is then updated by an autoregressive filter on the previous values:

$$A(t) \leftarrow \lambda A(t-1) + (1-\lambda)T_s(t); \qquad \lambda = 0.9 \tag{3}$$

These data are saved as a `C` structure on a ramdisk filesystem. Doing so, the `fairshareadjust` function only has to load a small datafile from the ramdisk into an ordered lookup table and return the ADJUST value as the result of a binary search. To further speed up the search process, an `adler32` hash is used in place of the real username. This way the binary search is performed comparing integers instead of strings.

4.2. Effect of fairshareadjust

To evaluate the effect of the ADJUST factor from (2) a burst of short jobs have been submitted by a user having high nominal priority U_{share}, using a simple submitter script similar to the one described in 3.1. Initially, jobs are dispatched at high rate (Fig. 3). The adjust factor gradually grows, thus reducing the dynamic priority of the user. After a while, the number of dispatched jobs per minute fall by more than 50%, thus actually preventing a short job flooding. After the

submission flow ends, the ADJUST factor gradually decays to zero. The decay time and the reactivity are controlled by the λ coefficient and can be set in the python code implementing the autoregressive filter (3).

The overall effect with and without fairshareadjust can be appreciated by comparing Fig. 5 and Fig. 3 with Fig. 2 and Fig. 4.

Fig. 5 reports an episode of short jobs systematically short because of thecnical problems affecting two user groups (middle and bottom chart) happened on September 29. The adjust factor prevents short job flooding by reducing the DP of the affected users. The number of running jobs remains higher and smoother if compared with Fig. 4, where the effect of a short job campaign is represented before enforcing countermeasures.

5. Conclusions

A sustained rate of short jobs has negative impact on the computing power of a batch system cluster by increasing the average number of unavailable slots (1). When a fairshare policy based on cumulated runtime is adopted by the batch system, and simple automatic submitter tools are adopted by users, short job flooding can happen, causing a severe computing power loss at the centre. This can be prevented by adapting the fairshare formula (2) to add fictious extra runtime to short jobs. This solution proved to be effective and is currently adopted at CNAF on the Tier–1 cluster. Eventhougth the implementation is specific to LSF, the general principle should apply to other fairshare based batch systems too.

References

[1] IBM 2014 Dynamic user priority `https://www.ibm.com/support/knowledgecenter/SSETD4_9.1.3/lsf_admin/dynamic_user_priority_lsf.html` [Online]

[2] Vasupongayya S 2009 *proceeding of the 2009 WASET International Conference on High Performance Computing, Venice, Italy*

Data management and storage systems

A. Cavalli, D. Cesini, E. Fattibene, A. Prosperini and V. Sapunenko
INFN-CNAF, Bologna, IT
E-mail: vladimir.sapunenko@cnaf.infn.it

1. Introduction

The Storage and Data Management Group is responsible for the management of the hardware infrastructure for the Storage Area Network (SAN), the storage systems, the Mass Storage System (tape library, Tivoli Storage Manager (TSM) servers, Hierarchical Storage Manager - HSM - servers), the disk servers and servers the Data Management services are running on. As to related services the group supports GEMSS[1], the Grid Enabled Mass Storage System, StoRM[2], the in-house developed version of the Storage Resource Manager (SRM), GridFTP, XrootD and the WLCG File Transfer Service (FTS). Additional responsibilities include the installation and management of Oracle based database services and the backup and recovery infrastructure, based on IBM Spectrum Protect (formerly Tivoli Storage Manager) software. Further this group is working on Long Term Data Preservation (LTDP) and a comprehensive monitoring and alarming infrastructure based on Nagios. The provided performance of the storage services is well aligned with the requirements as stated by the large user groups. The group is actively engaged in working in several areas to provide cost-effective and high performance storage solutions to fulfill ever-increasing requirements.

2. System provisioning

During this year our provisioning system based on Quattor has been migrated to the central CNAF provisioning infrastructure based on the open-source softwares Foreman and Puppet[3]. The work consisted in:

- Configuration of Foreman to properly classify nodes to install (both bare metal and virtual machines)
- Creation of Puppet classes to configure different services
- Creation of Foreman profiles for each host to install
- Re-installation of almost all hosts previously managed by Quattor (at the time of this report 40 hosts are not yet managed through the CNAF provisioning infrastructure)
- Installation of almost all the new nodes via Foreman

At the end of the year, 53 Puppet modules have been developed, 36 host groups have been defined in Foreman, 170 hosts result managed via this infrastructure.

3. Monitoring

A lot of work has been carried on to move monitoring data to the CNAF monitoring infrastructure based on Sensu and InfluxDB technologies. Even if other tools are still used to perform monitoring operations, such as Lemon (developed at CERN and no longer maintained)

and a system based on Graphite database and ad-hoc web pages, the target is to completely migrate to the new monitoring system. Lemon sensors, that have been developed ad-hoc to measure specific metrics such as the number of recall and migrations from/to the tape system, the data throughput from the GPFS servers, and some metrics related to the StoRM services have been re-written to be compliant with the Sensu/InfluxDB system. Other Lemon sensors used to collect common system metrics were replaced by Sensu community plugins. Monitoring data are saved in InfluxDB database. We have realized a set of dashboard using the open-source software Grafana, to easily show monitoring information at CNAF operators and users. At the end of this activity, all Lemon web views and plots were rebuilt in Grafana. Moreover different dashboards have been realized on the basis of data stored in Graphite database. For further details see the Chapter "CNAF Monitoring system".

4. Database management
Expertise in the implementation and management of Oracle Databases is today still part of the Storage department commitments. After the substantial reduction of the ATLAS and LHCb Oracle instances during the past years, in 2015 CNAF has completed the migration and started to provide the ATLAS Muon Calibration Database service, that was previously hosted and administered by the INFN-ROMA IT staff. In 2016 we had the confirmation that the work done for this Calibration Database was good, since the experiment personnel is using it, and the scripts deployed to ensure the replication of all the data to a central Database at CERN are doing their duty. The Oracle RAC Database installation currently in production is described as follows:

- 3 RAC nodes running Oracle 11.2.0.4 on a RHEL6 x86_64 platform;
- 2 cluster Databases, managed with transparent cluster services and SCAN listeners;
- External EMC2 storage connected via FC hba (redundant links), managed by Oracle through the ASM features;
- Double (redundant) private vlan for cluster interconnection;
- RMAN backup on separate disk space and on tape library.

This setup hosts also the Database for the legacy Lemon monitoring system, that is going to be replaced by a new monitoring system, hence it will be dismissed at the end of the substitution process. And for completeness of the report, other minimal-impact Databases are hosted:

- a small DB used by developers of the VOMS software;
- a backend Database for the FTS (File Transfer Service), that is for a pilot or not yet in production service.

The Oracle setup of the Tier-1 infrastructure includes and benefits also of a Oracle Enterprise Manager Cloud Control 12c installation, that considerably facilitates all the management tasks, and is shared and managed together with another CNAF unit that make use of Oracle installations.

5. Backup and recovery service
The CNAF backup and recovery service, operated by Tier-1 Storage department, as been reconfigured to improve data safety and restore speed. For further details see Chapter "CNAF backup system".

Figure 1: The DDN SFA12K storage system

6. Disk-based (online) Storage

The total amount disk-based storage installed at CNAF and available to end users is $\sim$ 21.5 PB. All this space is being managed by Spectrum Scale (GPFS) parallel file system and subdivided in 5 major clusters with a mean size of a single file system of 3 to 4 PB. After the storage server consolidation campaign carried on during previous year we have replaced all 1Gbe-based I/O servers with 10Gbe ones significantly reducing number of servers and increasing available bandwidth. This time, the consolidation went into direction to replace 10Gbe servers with 40 Gbit servers by bonding together 4 10Gbe interfaces on each server. To stay in line with increased front-end bandwidth we decided to give a try to a different interconnection protocol for back-end connectivity - FDR InfiniBand. The new storage infrastructure was installed and went into production at the beginning of the 2016. Apart of 10PB of usable disk space, it includes two Mellanox IB switches (36 FDR ports each) and 16 I/O servers equipped with Mellanox dual-port FDR HBA. For the IB interconnection we have decided to replicate already proven fully-redundant SAN-like configuration with two independent switched fabrics, so each storage controller can be accessed from any server via both IB switches preserving maximum bandwidth that any server can provide even in case of a fabric failure. Considering fully redundant interconnection, every server in normal conditions provides access to $\sim$ 700 TB of disk space, and in case of some server failure could serve up to 4 time more space with the only limiting factor on the 40 Gbit front-end interconnection.

The new storage systems (DDN SFA12K), whose picture is shown in Fig. 1, came with new HDD technology: Helium-filled 8TB Near-Line SAS with 4K sector size (4KN). Since 4K sector size is not widely supported, and in particular GPFS v.3.5 does not support them at all, we were forced to upgrade GPFS to v.4.1, recreate new file systems and copy about 8 PB of data. Another important improvement was to move metadata for all GPFS file systems from rotating

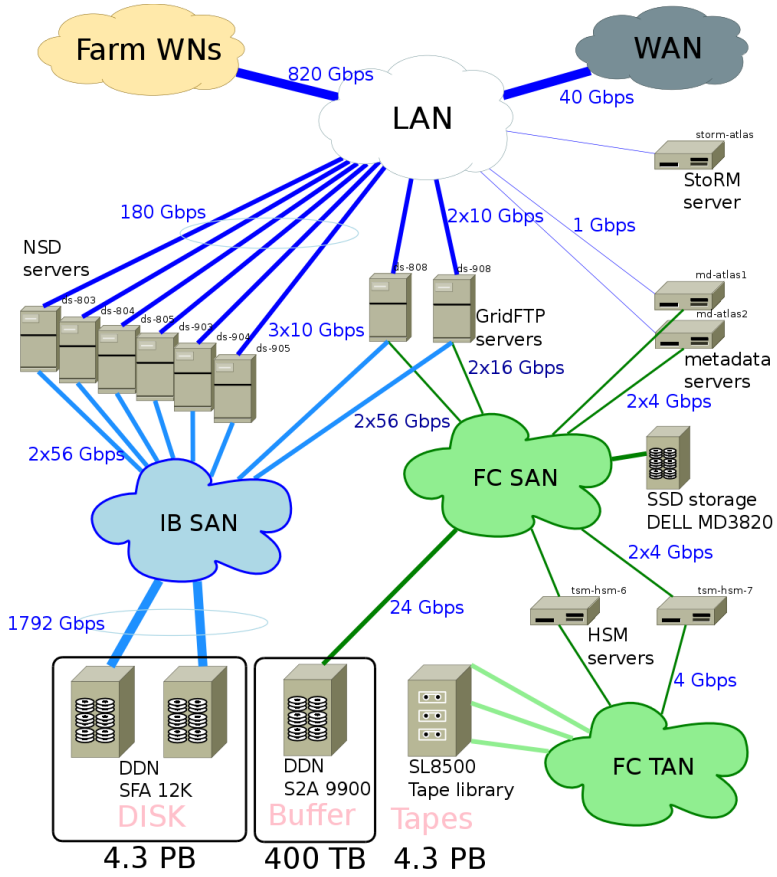


Figure 2: ATLAS exploitation of CNAF storage systems

disks (SAS) to SSD-based storage. For this purpose we have procured and installed two dual-controller Dell MD3820f storage arrays equipped with 24x600GB MLC SSD. On Each MD3820f we have created 12 LU of 2 SSD each in RAID1 (mirrored) configuration. For the metadata of each file system we allocated one or two LU on each storage system and imposed replication level of 2 to have 1 full copy of all metadata on each SSD storage array.

At the end, all HW dedicated for a typical GPFS cluster (ATLAS in this case) has been reduced to:

- 8 I/O servers;
- 2 metadata server;
- 1 HSM (+1 warm spare)
- 1 data management server (VM)

Fig. 2 shows ATLAS data at CNAF (4.3 PB) physically located storage systems shared with other experiments.

7. Tape-based (near-line) Storage

CNAF tape system has been consolidated during 2016. For each big experiment running at CNAF, HSM servers have previously configured as a couple of servers in active-active mode. To decrease TSM license costs we moved from active-active to active-standby configuration, such as a server is running to accomplish migration and recall activities between tape media and buffer

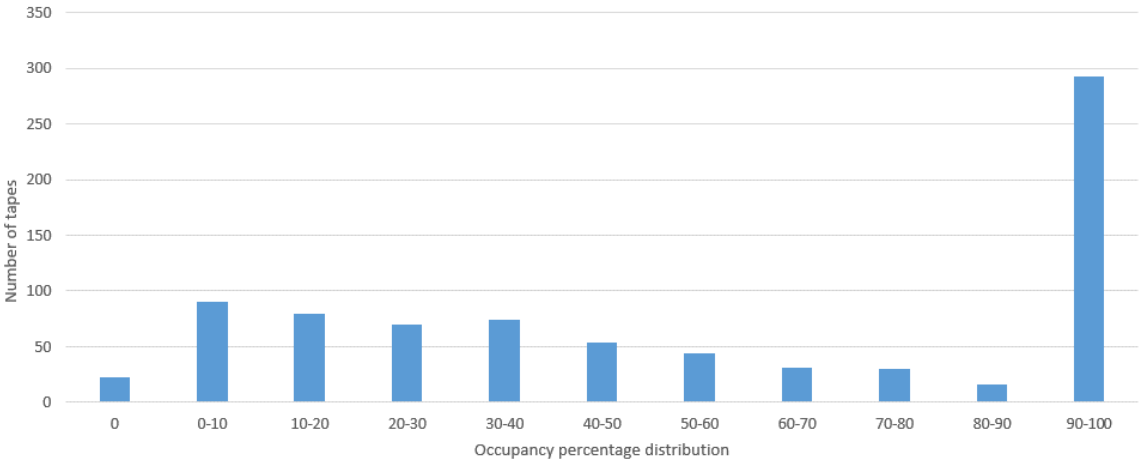


Figure 3: CMS distribution of occupancy percentage after CMS cancellation campaign

disk and another is started in case of outage of the primary instance. Each HSM server is capable of handling 800 MB/s in and out of tapes (simultaneously). Theoretical limit is defined by a single FC8 connection to the Tape Area Network, but some inefficiency in migration candidate selection prevents us from reaching 800 MB/s in mean daily transfer rate, that results arriving to 600 MB/s. A bunch of 750 new Oracle-Storagetek T2 tape cartridges have been installed in the SL8500 tape library, to reach a total amount of 5200 cartridges available for experiment data (corresponding to 44,5 PB of space). During 2016 a total of 16 PB of additional space has been occupied on tape, with the following distribution among LHC VOs and no-LHC aggregated VOs:

- ALICE: 4 PB
- ATLAS: 1,9 PB
- CMS: 2,1 PB
- LHCb: 2,6 PB
- no-LHC VOs: 5,4 PB

The main data cancellation campaign was carried out by CMS experiment, that freed 3,7 PB of space, corresponding to 1.1 million files. At the end of this cancellation activity (end of September), a total of 22 tapes were completely emptied. Fig. 3 shows the percentage of occupancy of the tape belonging to the two main CMS tape storage pools (MC08 and RECO).

A total of 785 tapes resulted filling with a percentage of occupancy less than 90%. A repack activity has been accomplished until mid-December to free all these 785 tapes, moving around 2 PB of data, for a net earnings of 550 tapes. In order to investigate the high failure rate in T2 tapes over the last two years (17 read failures over a mean number of 3000 tapes installed), we sent some cartridges to Oracle laboratory in Colorado (USA). The response confirmed our feeling that dust particles affect our tapes. In order to solve this problem an activity on a dust sensor development has been planned.

8. Data Management

Remote data access supported by CNAF are:

- SRM (StoRM)
- GridFTP

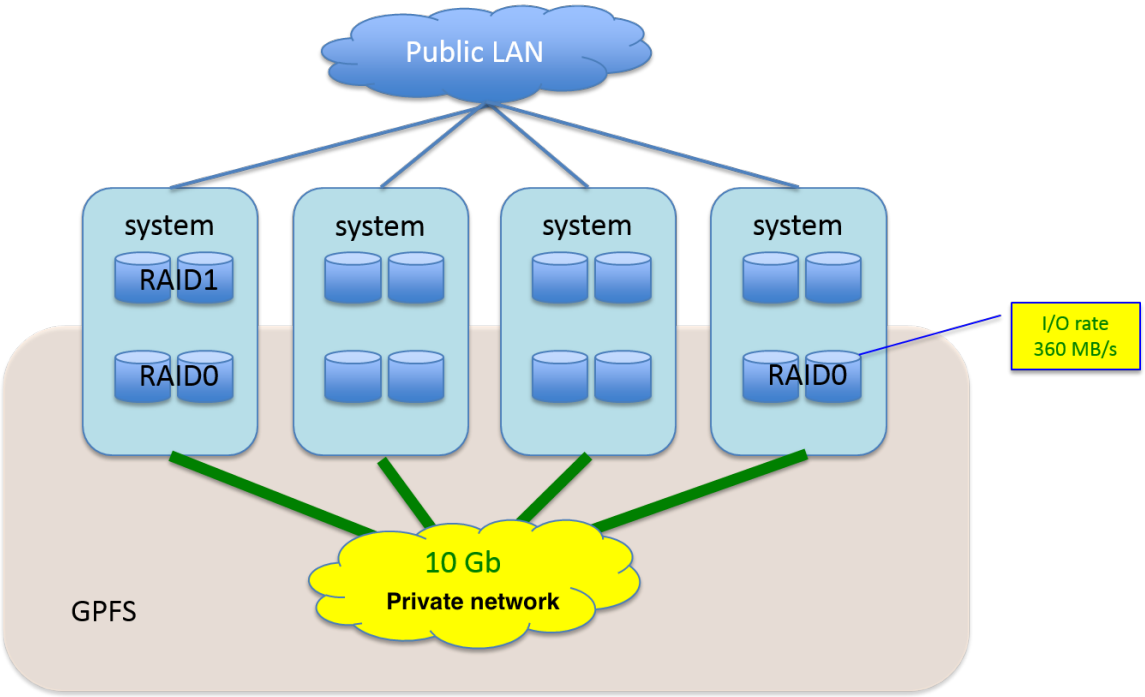


Figure 4: Layout of the StoRM cluster

- XrootD
- WebDAV

Since SRM is a pure management service and dont need to do intensive I/O with data (apart of metadata) it has been separated in a dedicated pool. All other services, having necessity to move data were compacted in dedicated I/O servers, 2 or 4 for each experiment. Virtualization of data management servers (StoRM FrontEnd and BackEnd) within single VM for each major experiment permitted to consolidate HW and increase availability of services. Virtualization infrastructure for StoRM services located on a GPFS FPO (File Placement Optimizer) Cluster of 5 KVM Hypervisors. The particularity of this cluster is that the shared filesystem based on local disks of 4 servers (so called"Shared Nothing Cluster") configured in 3-way redundancy (3 copy of each file or data object). Each HV apart of public IP on dedicated 1Gbe or 10 Gbe interface, has also a dedicated 10Gbe connection for the cluster interconnection which makes live migration and data replication faster than access to local disks (Fig. 4).

9. Long Term Data Preservation[1]

In the scope of CDF Long Term Data Preservation (LTDP) project, during 2016 we completed Run2 data copy to the CNAF storage systems. A contact has been established with the CDF Oracle administrators at Fermilab in order to synchronize the Run2 database with a copy at CNAF. This work will be carried on during 2017. Simultaneously a work has been started to preserve CDF Run1 data, at the moment stored on about 4000 old tape cartridges that need specific hardware to be read. This hardware has been bought and a functionality test is scheduled for the first months of 2017.

[1] This activity has been carried out in collaboration with S. Dal Pra, M. Pezzi and P.P. Ricci

10. Transparent Remote Data Access

AFM (CNAF-Bari, and CNAF-ASI) The goal is to provide to jobs running on remote site transparent access to Tier-1 data with the same namespace and using the same protocols as in case of local access. This has been archived using Active File Management (AFM) feature of GPFS and is being used by two remote sites at $\sim$ 600 km distance.

11. References

[1] Ricci, Pier Paolo et al., The Grid Enabled Mass Storage System (GEMSS): the Storage and Data management system used at the INFN Tier1 at CNAF *J. Phys.: Conf. Ser.* **396** 042051 - IOP Publishing (2012)

[2] Carbone, A., dell'Agnello, L., Forti, A., Ghiselli, A., Lanciotti, E., Magnoni, L., ... & Zappi, R. (2007, December). Performance studies of the StoRM storage resource manager. In e-Science and Grid Computing, IEEE International Conference on (pp. 423-430). IEEE.

[3] CNAF Provisioning system on CNAF Annual Report 2015

CNAF backup system

A. Cavalli, E. Fattibene, A. Prosperini and V. Sapunenko
INFN-CNAF, Bologna, IT
E-mail: enrico.fattibene@cnaf.infn.it

Abstract. A backup system is crucial in a datacenter running critical IT services and resources for a large set of users. This paper presents the work carried out during 2016 to re-configure the CNAF backup system, exploiting the latest version of IBM Spectrum Protect backup and archive software, disk based storage and tape systems. Here we describe the system components, together with details about the data workflow. The operation activity is illustrated with particular attention on problem resolution procedures and restore tests.

1. Introduction

In a complex data center as CNAF a set of services are considered as essential, and must be protected against service and data loss. National Services, configurations of Tier-1 systems, centralized log servers, components of IT services and mail servers are some of the most critical ones, whose data have to be saved on different places and possibly on different media. In case of data loss on such important services, the data center has to be equipped of a robust and resilient backup system, to guarantee the restore of data as soon as possible. In order to achieve this goal, the INFN Tier-1 Storage team has implemented its backup service based on IBM Spectrum Protect [1], formerly Tivoli Storage Manager (TSM), a closed source software designed by IBM, one of the leader data protection solutions available. Data backed up through Spectrum Protect are stored on different media, such as disk and tape storage pools. Disk storage pools are provided by a DDN storage system served by a GPFS filesystem. Tape storage pools are composed by tape resources offered by an Oracle-StorageTek SL8500 tape library, used also for storing HEP and multidisciplinary experiment data, served by 9 Oracle T10000C tape drives (2 of them dedicated to Spectrum Protect server database backup on tape).

2. Backup system configuration

During 2016 CNAF backup system has been re-configured in order to better take advantage of Spectrum Protect software and CNAF storage resources. The latest version (7.1.5) of Spectrum Protect software has been installed on a new host equipped with an Intel Xeon E5-2640-v2 8-core processor and 32GB RAM. This setup allows to exploit both Incremental backup and Archive features available with Spectrum Protect. In the scope of Incremental backup the most recent version of an object saved on the system is defined Active. This version stands on the system for an unterminated period of time. When an object is modified or deleted from its original filesystem, at the subsequent backup the new version becomes Active and the modified version (or the deleted one) becomes Inactive. Inactive data are stored on the system for a period of time specified as retention for each client node by Spectrum Protect server administrators. We have

a set of client nodes under Incremental backup with retention of 30 days and another set with a retention of one year. In both cases backups are scheduled on a daily basis. Schedules are set on the server and automatically start at different hours during the day in order to minimize the interferences with the services running on the client nodes. Besides this, thanks to the Archive feature, its possible to store data on another dedicated storage pool, typically for longer-term needs. This activity is scheduled on the client node. Archived data remain available for a period of time set as retention (at time of writing we are taking advantage of the Archive feature for one clients data that needs to be preserved for a period of 5 years). The setup is composed by:

- a Spectrum Protect server version 7.1.5;
- 16 client nodes (TSM / Spectrum Protect versions 6 or 7) running on physical and virtual nodes (hosted by oVirt and VMWare virtualization infrastructures) distributed as follows:
  - 12 standard clients on Linux (CentOS or Scientific Linux);
  - 2 Oracle database clients (Tivoli Data Protection for Oracle Database - TDPO) on RedHat Enterprise Linux;
  - 2 standard clients running Windows Server 2012;
- a 64TB disk area (4 Logical Unit Numbers (LUN) of 16TB each) on DDN storage system, dedicated to Active data;
- 9 Oracle-StorageTek T10000C (7 for data and 2 for Spectrum Protect database backup);
- 40 T2 tapes with capacity of 5,4TB each (35 tapes for data and 5 for Spectrum Protect database backup).

Fig.1 shows the backup system workflow. Multiple copies of data are stored on different media in order to minimize the risk of complete data loss. In the scope of Incremental backup, when data are sent form clients to the server, they are simultaneously written on disk and tape systems. Full lines represent data flow for every client nodes; dotted lines are depicted to show additional data flow for selected clients (second copy of Inactive data or Archive copies). We preserve Active data of each client node with a copy on disk and a second copy on tape. The first copy on disk has been arranged to minimize the reactivation time in case of complete restore for a given client node. This is especially needed when you have several clients with an order of 10 million files. Inactive data are stored only on tape. In particular, there is at least a copy of Inactive data for all client nodes; a second copy on tape (together with a third copy of Active data) is provided only for selected clients. This additional copy is not necessary in some cases when data are already duplicated outside from CNAF backup system, that can occur both locally on the CNAF systems or even at other (remote) INFN locations. The Archive feature is used to store data on tape (2 copies) for a long period of time, regardless of their presence, modification or deletion on the client nodes filesystem. At the time of this report, the backup system is holding roughly the following amount of data:

- 16 TB / 72 million files of Active data on disk;
- 75 TB / 145 million files of Inactive and copy data on tape;
- 2 TB of Archive data on tape.

2.1. Notification emails

Each CNAF department having at least a client node under backup receives via email a daily report on the nodes managed by the department itself. This report is scheduled through the Spectrum Protect Operation Center, a monitoring and administration tool released with the Spectrum Protect suite. The report is composed by the results of 3 queries to the server database:

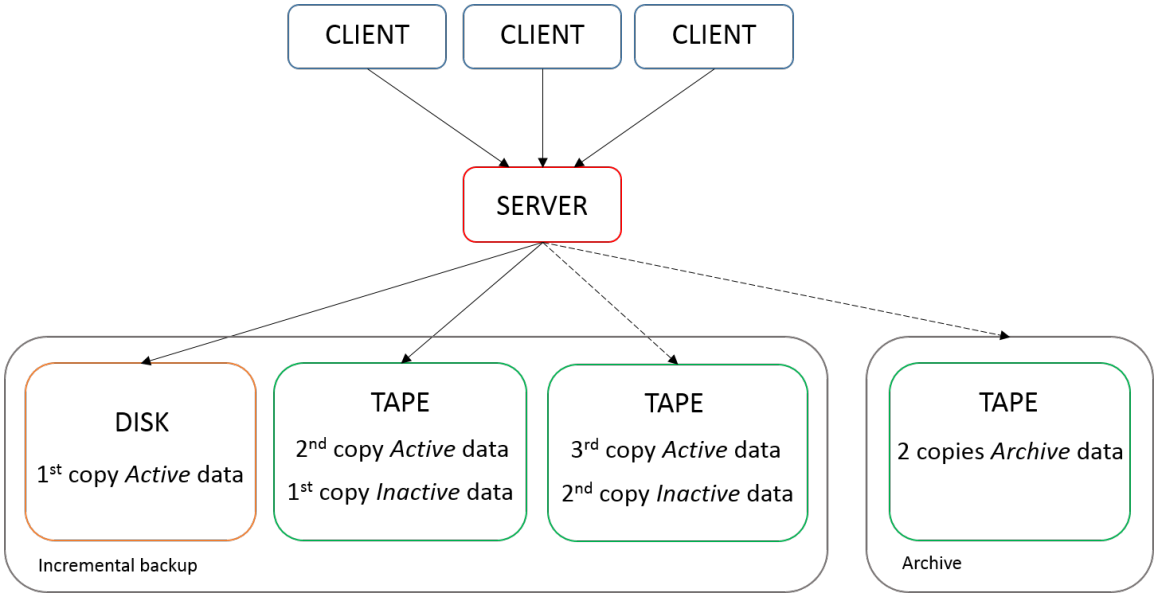


Figure 1: CNAF backup schema

- scheduled backups in the last 24 hours: schedule time, real start time, end time, final status - Completed, Failed, Future, Missed, Started, Restarted and the output that gives information about the presence of warnings or errors);
- space occupation and number of files per client on each storage pool (disk or tape);
- space occupation and number of files (all copies on every storage pools) per filespace.

The aim of this report is to minimize errors and warnings. Each client administrator is in charge to analyze the report and take the suitable actions to solve the problems, supported by the backup system administrators.

2.2. Restore tests

The maximum time requested to restore data of a compromised service depends on the level of criticality of data backed up. However, in case of outage of a service, data have to be restored in the minimum amount of time, in order to minimize the downtime. To test the correct operation of the backup system and to minimize the restore time, we consider to run a couple of complete data restore for each client node on a yearly basis. This test consists in a restore of all Active data from disk storage pool and a sample of Inactive data from tape. These tests are planned together by backup system administrators and client operators.

3. References

[1] http://www-03.ibm.com/systems/storage/spectrum/protect/

Dataclient: a simple interface to an X.509-based data service

V. Ciaschini
INFN-CNAF, Bologna, IT

Abstract. There is a disconnect between user preferences and site preferences when using data management tools in a grid-based data center: the tool of choice on the server side is globus's GridFTP, which requires client certificates, while users find certificates and all their management (request, renewal, VOs, etc...) complicated and would prefer not to have to deal with them.

Dataclient is a way to let the two sides meet in the middle. It manages all the X509 certificates and VO infrastructure automatically, without user interaction at all, while at the same time allowing them to connect to the site's preferred backend.

1. Introduction

Dataclient is a tool that aims to address one of he issues reported by users, i.e. certificate and proxy management is difficult and complex, while not requiring the data center to migrate its services to versions that no longer require X.509 certificates. It does so by completely hiding the complexity by moving it from the user to the tool itself. It does so by presenting a simple, password-based interface whose operation is already well understood, and keeping some of the advantages of proxies, like single sign-on.

2. Operation

Dataclient's operation is simple. At first invocation, it takes a little time to setup the client side of the X.509 infrastructure and then asks the user to authenticate with its CNAF credentials and experiment. If authentication succeeds than a client certificate with full VOMS[1] credential for the experiment is created and from that moment on the user can use the dataclient executable as a synonym of 'globus-url-copy', passing the same options, but without any need of managing a certificate. The user is not required to reauthenticate each time unless:

- the user's password at CNAF has changed. (Note that CNAF policies force at least one password change each year)
- the user's membership has been revoked.

If the password has merely changed that at the next execution the user will be asked to reauthenticate. However, if his membership had been revoked, he will be informed of this fact but will *not* be able to reauthenticate until his membership has been reinstated. In both cases his existing certificate will be revoked immediately so that he will not be able to extract it and use it anyway.

3. Architecture

Dataclient is formed by two main components, a client that will be run by the user and a server that is at CNAF.

The server verifies the successfully verifies the user authentication using PAM, and if successful act as the issuer of the certificate that will be granted to a user, and at the same time takes an accurate catalog of all files that have been successfully transferred along with their md5 hash. It also keeps track of files that are in transit, and considers the transfers failed if it does not complete inside of a set amount of time.

The server keeps track of what certificate has been granted to which users, and acts as a Certificate Revocation List (CRL) publisher and, limited to its gridmap interface, as a voms-admin. The credentials with which the user has authenticated are destroy immediately after authentication, regardless of failure or success.

The client at first start downloads into the .duc sub-directory of the user's home a full set of Certification Authorities (CAs), both the IGTF ones and the ones used for the online CA, along with the full libraries and cli clients of Globus. Afterwards, it copies itself in the .duc directory.

From now on, calling dataclient regardless of its path will always redirect to the copy in the $HOME/.duc directory. This way, dataclient can easily update itself, and indeed any request sent to the server contain its version number. If the server knows a subsequent version than the user's request is paused momentarily while the client updates itself by updating its version in $HOME/.duc and then restarts.

After an eventual update, the user must authenticate. The client asks the user its username, password and experiment. The password is then salted, encrypted with a public key, and then sent, along with username and experiment, to the server on a TLS connection established with a different public key. The server decrypts the password, removes the salt, and then asks PAM if the credentials are valid. Subsequently, its copy of the credentials is destroyed and, if the authentication succeeds, a full VOMS-compliant proxy for the experiment is created and returned to the user.

After this, the user has all he needs to be able to successfully use globus commands. Indeed, from this point on, dataclient becomes a synonym for 'globus-url-copy' and supports all its options. There are also two small extensions:

(i) dataclient detects automatically if the source address is a local directory, and in case adjusts the syntax to transfer the whole directory, while by default 'globus-url-copy' would fail with a syntax error and,

(ii) it tries to interpret the error messages from the gridftp server and print something more user friendly.

In details, when one tries to transfer files with dataclient the following happens.

(i) The full list of files that would be transferred is determined, and this list is sent to the server which marks them as 'in transfer.'

(ii) globus-url-copy is called to transfer the files.

(iii) globus-url-copy is called a second time to verify that the transfer was successful. This is done by globus-url-copy itself by comparing the md5 hash on both ends. If the result is a success, the server is contacted and the transfer marked as completed, otherwise it is marked as aborted.

If the process is interrupted at any spot for any reason, the next time dataclient is called it will detect this fact and mark all transfers as failed. A transfer is also marked as failed ex officio if a notice of completion does not arrives within 24 hours of its start.

Trying to overwrite files that are being transferred is detected and denied.

It is also possible to request from the server the full list of all files successfully transferred using this facilities, along with their md5 hash.

For this to work the gridftp server must be able to see a gridmap-file containing the users of the experiment, and it must have fresh CRLs for the online CA to which they belong. Which in turn it means that the list of users must be published. To solve these issues, the dataclient server also acts as CRL provider and, limited to the creation of gridmap-files, as a VOMS Admin, thus providing all that is needed.

Finally, detection of when an account's password has changed is done by having a publisher on the Kerberos side, reachable ONLY from the dataclient server, which publishes the last date in which an account's password was changed. The dataclient server consults this periodically and, if a password has been changed after the date in which it issued a certificate to the corresponding user, the certificate is revoked, forcing the user to authenticate again.

All in all, a typycal usage would look like this:

```
$ dataclient
This is the first time you used this program.
Please wait a bit while a one-time initial setup proceeds
to initialize the system.

You are either a first time user, or your local setup was corrupted
and had to be recreated from scratch.
Because of this, you are required to reauthenticate.

Please use the username and password with which you registered at CNAF.
The experiment name is only required if you are registered as a member of more
than one experiment, and can be left empty otherwise.
Username: xxxxxx
Password:
Experiment:

$ dataclient file:///some/path gsiftp:///some/other/path
```

where the first command would only be necessary the first time or when the password associated with the username changes.

4. Early Adopters

Dataclient is currently under evaluation by representatives of some experiments.

5. Security Considerations

Even though they are well protected when in transit, and the server is careful never to save them anywhere, a malicious administrator could modify the server to log them.

This problem can be solved operationally simply by imposing that the server should be administered by the Kerberos administrator. This effectively removes the risk since Kerberos's administrators can already login as any of the Kerberos user without necessity of knowing the passwords, and therefore have nothing to gain from doing so and are already trusted not to (ab)use this privilege.

6. Thanks

I would like to thank the whole of the User Support service at CNAF, whose people not only provided the initial request for the creation of this server, but were also invaluable in dedicating

many hours for testing the final product and providing first-line support to those who are evaluating it.

References

[1] R. Alfieri,R. Cecchini,V. Ciaschini,L. dell'Agnello,A. Frohner,K. Lorentey,F. Spataro, From gridmap-file to VOMS: managing authorization in a Grid Environment, Future Generation Computer Systems, Vol. 21 issue 4, Pages 549-558.

The INFN Tier-1: Network

S. Zani[1], L. Chiarelli[2] and D. De Girolamo[1]
[1] INFN-CNAF, Bologna, IT
[2] GARR Consortium, Roma, IT
E-mail: stefano.zani@cnaf.infn.it

1. Introduction

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

2. Wide Area Network

Inside CNAF data center is also hosted the main PoP of the GARR network, one of the first Nodes of the recent GARR-X evolution based on a fully managed dark fibre infrastructure. CNAF is connected to the WAN via GARR/GEANT essentially with two physical links (Fig. 1):

- General IP link (GPN), which has been upgraded to 20 Gb/s, via GARR[1] and GEANT[2];
- The link to WLCG destinations, which has been upgraded to 60 Gb/s. This link is shared between the LHCOPN[3] for the Tier-0 $\leftrightarrow$ Tier-1 and Tier-1 $\leftrightarrow$ Tier-1 traffic and LHCONE[4] used for traffic with most of the Tier-2 and Tier-3 sites.

While GPN is accessible from the whole CNAF, LHCOPN and LHCONE are dedicated to the Tier-1 nodes.

Also, a 2x10 Gb/s L3 VPN with the Bari-ReCaS data center has been configured to allow the extension of the Tier-1 (see Chapter "The INFN Tier-1 towards LHC Run 3").

2.1. Network capacity usage

During 2016, network capacity usage has been growing and on both LHCOPN and GPN (Figg. 2 and 3).

For this reason, the Tier-1 link to LHCOPN and LHCONE has been upgraded from 4x10 Gb/s to 6x10 Gb/s and the OPN link (dedicated connectivity from Milan to CERN) has been upgraded from 2x10 Gb/s to 4x10 Gb/s. In fact, the traffic on LHCOPN was limited to 2x10 Gb/s: it is clearly visible (Fig. 4) the input traffic peak due to the upgrade to 4x10 Gb/s done on July 27th.

The Tier-1 link can be upgraded at any time to 8x10 Gb/s or more, while GARR has in its roadmap a backbone upgrade in order to support 100 Gb/s links between GARR POP at CNAF and GARR POP in Milan (GEANT peering). We foresee to be able to upgrade to 100 Gb/s our network infrastructure during the second half of 2017.

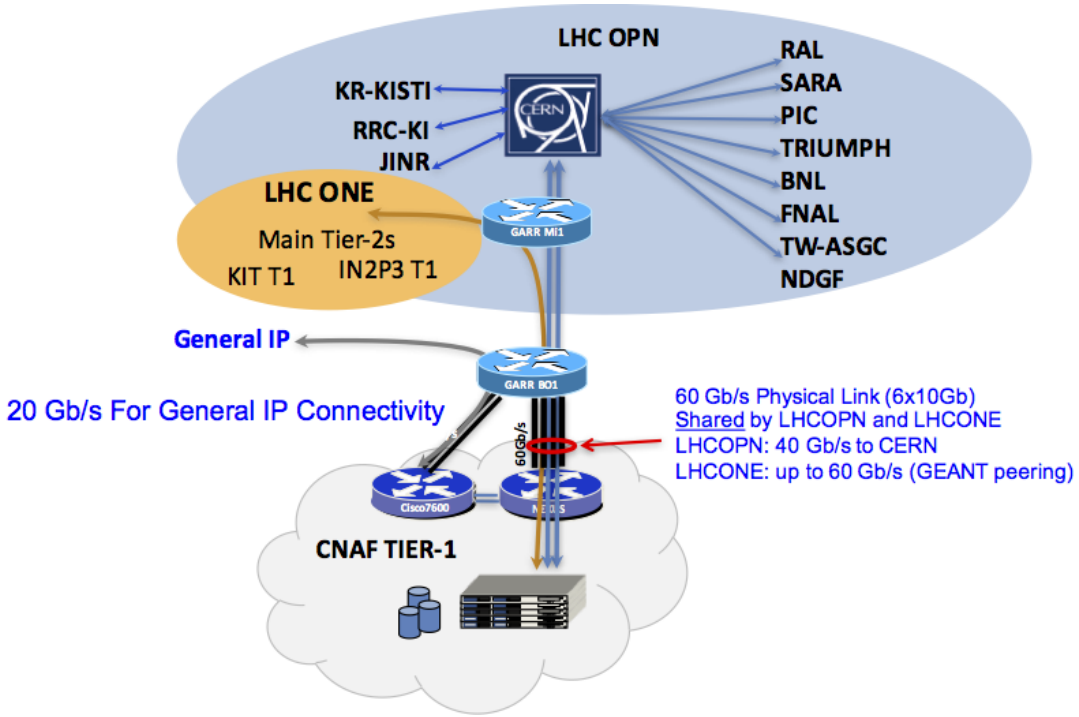


Figure 1: INFN CNAF WAN connection schema

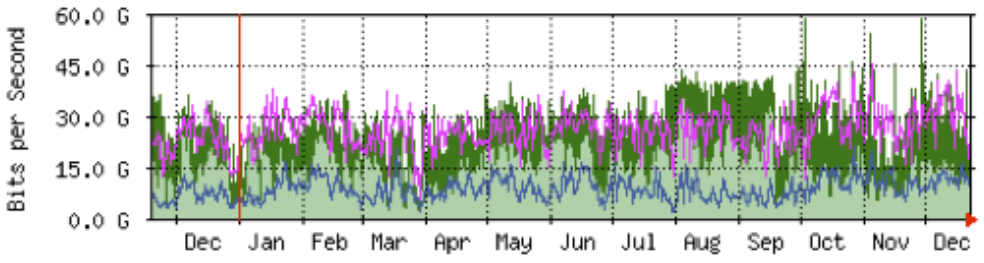


Figure 2: Aggregate traffic on LHCOPN and LHCONE networks

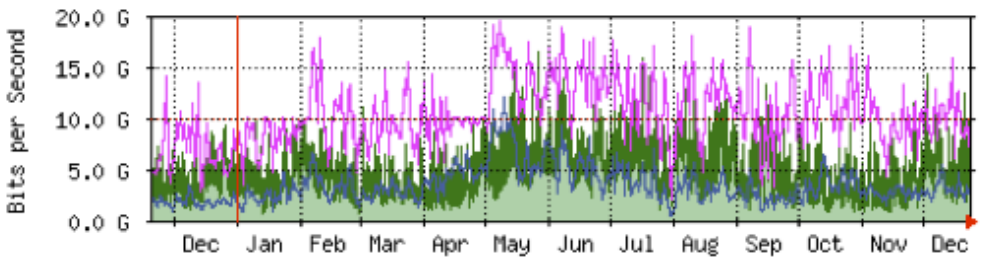


Figure 3: General IP Link usage

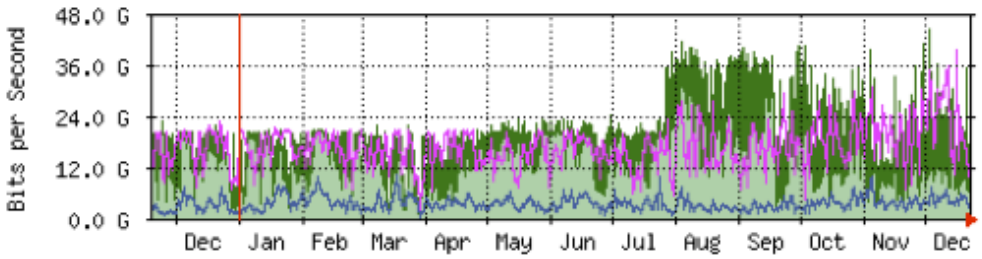


Figure 4: Traffic on LHC OPN network

3. Local Area Network

The Tier-1 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 switches (Top Of the Rack) with Gigabit Ethernet interfaces for the Worker Nodes of the farm and 10Gb Ethernet interfaces used as uplinks to the core switch. Disk-servers and gridftp servers are directly connected to the core switch at 10Gb/s. General Internet access, local connections to the offices and INFN national services provided by CNAF are managed by another network infrastructure based on a 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 Tier-1 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.

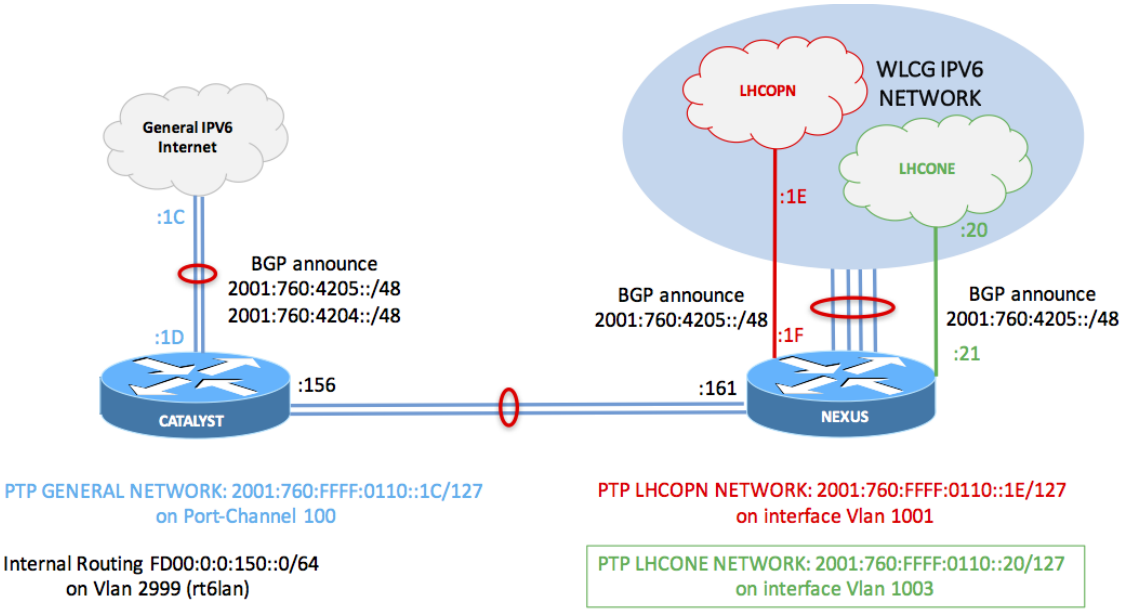


Figure 5: IPv6 CNAF implementation schema

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 has been implemented in March

2015 on the LHCOPN/ONE and on the General IP Network (Fig. 5).

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

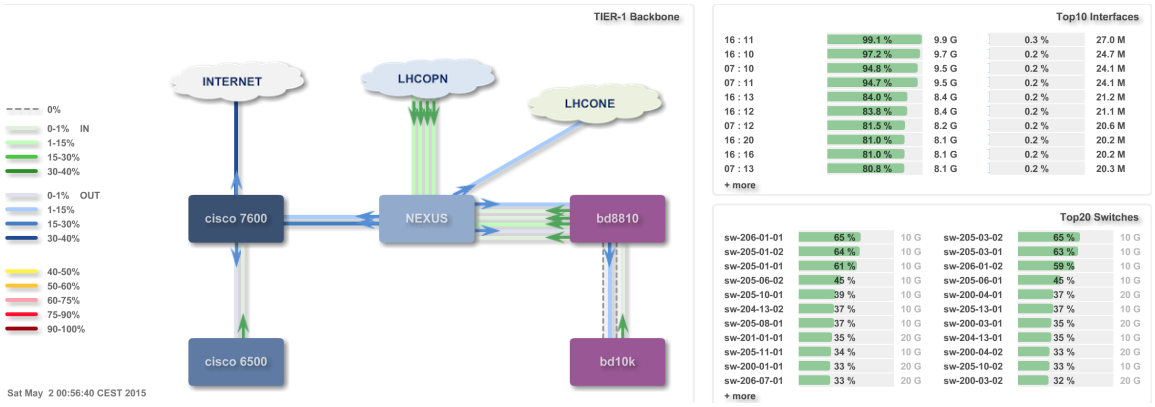


Figure 6: Net-board screenshot

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. The Net-board integrates MRTG, NetFlow Analyser and Nagios with some scripts and web applications to give a complete view of the network usage and of possible problems (Fig. 6). The alarm system is based on Nagios. The network security policies are mainly implemented as hardware based ACLs on the access router and on the core switches (with a dedicated ASICS on the devices). The network group, in coordination with GARR-CERT and EGI-CSIRT, also takes care of security incidents at CNAF (both for compromised systems or credential and known vulnerability of software and grid middleware) cooperating with the involved parties.

6. OpenFlow SDN application on KM3Net DAQ

The CNAF Network Group has supported KM3Net experiment in the implementation of part of the Data Acquisition Network. The experiment detector is composed by several Digital Optical Modules (DOMs). Every DOM (Fig. 7) has a Central Logical Board (Fig. 8) implementing a standard Ethernet interconnection port at 1 Gb/s. These links carry asymmetric traffic (Fig. 9). On one hand, data is transmitted from each DOM (on the TX fibre) to the onshore acquisition calculators, the DataQueue, through an aggregation switch (DOM Front End Switch, DFES) and the Star Center Switch Fabric (SCSF). On the other hand, each DOM receives (on the RX fibre) the synchronization signals, the address from the DHCP server and the slow control data, through two dedicated "White Rabbit" switches[1]. An intermediate switch, the Slow Control Base Data (SCBD), is the bridge between the two branches of this network. As first step we have had to modify the firmware of the "White Rabbit" switches in order to bypass the default per forwarding table not suitable for our case.

[1] A White Rabbit Switch (WRS) provides precision timing and high accuracy synchronization in an Ethernet-based network. See http://www.ohwr.org/projects/white-rabbit/wiki/Switch for more details.

Figure 7: DOM (Digital Optical Module)

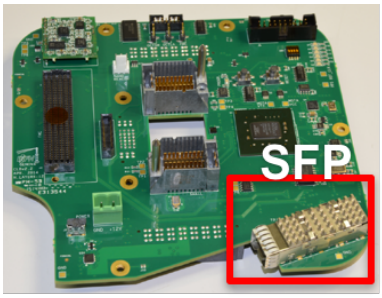


Figure 8: CLB (Central Logical Board)

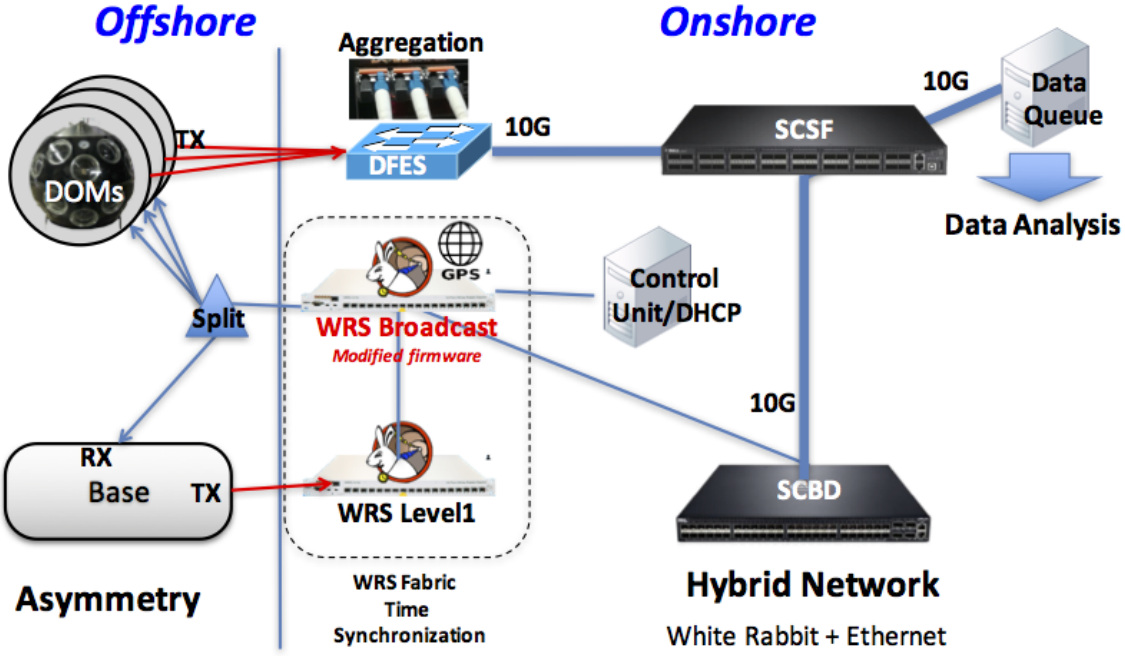


Figure 9: Legacy configuration (no SDN)

To allow the correct forwarding of the packets from the DOMs, through the various switches, to the various components (i.e. data to the DataQueue, controls to the SCBD etc...), we have chosen to adopt a SDN approach.

To implement this specifc configuration, OpenFlow 1.3 has been used on an Open Daylight SDN controller and DELL S-series Switches. With the adoption of the SDN approach it has been possible to connect all the switches in the most effective way and taking care of the only necessary traffic patterns. Simply implementing a MAC address routing based on the injection of a very limited numbers of rules (Tab. 1).

The effect of the SDN approach is clearly evident from the comparison of Fig. 11 with Fig. 12, showing the traffic to the SCBD in the two cases:

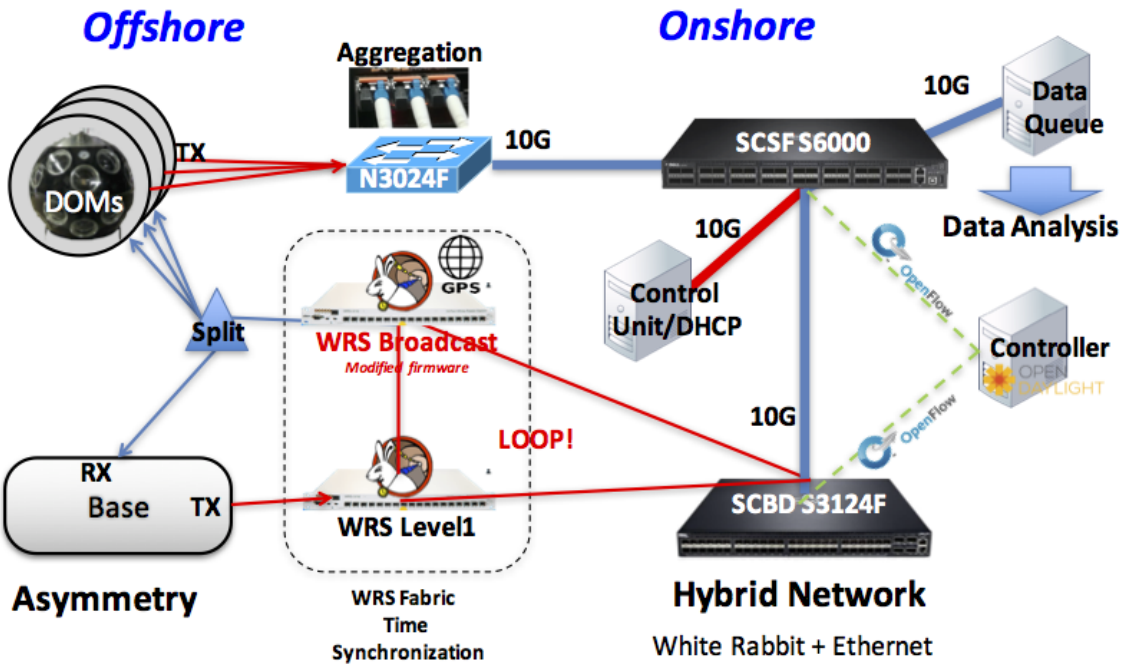


Figure 10: SDN configuration

Rule	Source	Destination	Action	Description
SCSF-1	*	FF:FF:FF:FF:FF:FF	Out to CU (DHCP server)	Broadcast to CU
SCSF-2	08:00:30:00:00:00/ff:ff:ff:00:00:00	MAC[CU]	Out to CU	All DOMs to CU
SCSF-3	08:00:30:00:00:00/ff:ff:ff:00:00:00	MAC[DataQueue]	Out to DataQueue	All DOMs data to DQ
SCSF-4	MAC[CU]	08:00:30:00:00:00/ff:ff:ff:00:00:00	Out to SCBD	All controls to SBCD
SCBD-1	08:00:30:00:00:00/ff:ff:ff:00:00:00	*	Out to uplink to SCSF	All Base data to SCSF
SCBD-2	Any from uplink to SCSF	08:00:30:00:00:00/ff:ff:ff:00:00:00	Out to WRS-broadcast	From SCSF to WRSB

Table 1: SDN rules description

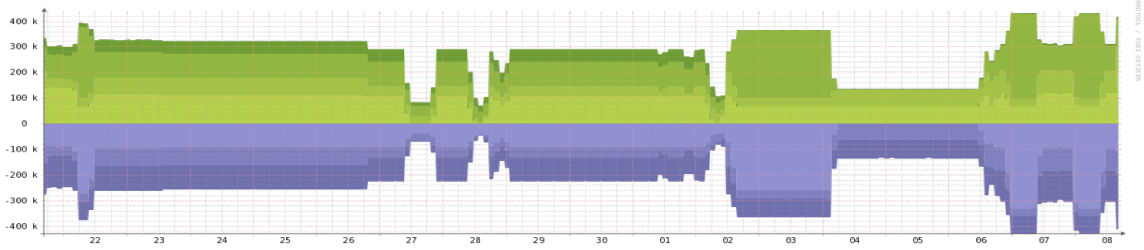


Figure 11: Traffic to the SCBD (Legacy Network Interconnection)

- improved stability of the data flow to the Slow Control Base Data
- White Rabbit Switch Broadcast more reliable
- Better uptime of the system (no more reload needed)

This technical advice job done by CNAF is being used in the KM3Net production site of

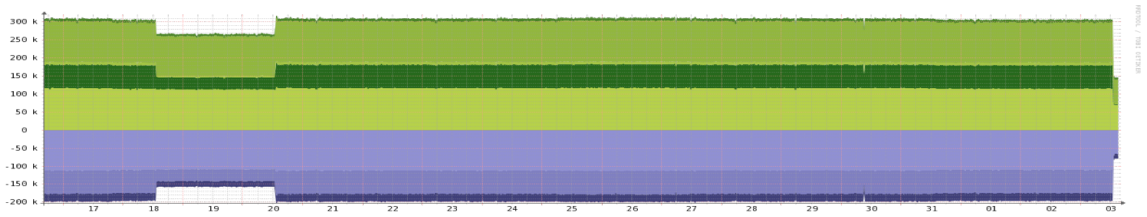


Figure 12: Traffic to the SCBD (SDN configured)

Portopalo di Capopassero (SR) but it has to be consolidated with the development of a user-friendly configuration interface for the injection of new rules or the modification of the existing ones with Experiment afferent staff in order to give continuity on the management and evolution of the DAQ network infrastructure.

7. References

[1] http://www.garr.it/
[2] https://www.geant.net/Pages/default.aspx
[3] http://lhcopn.web.cern.ch/lhcopn/
[4] http://lhcone.web.cern.ch/

The INFN Tier-1 towards LHC Run 3

A Cavalli[1], L Chiarelli[2], A Chierici[1], D Cesini[1], V Ciaschini[1], S Dal Pra[1], L dell'Agnello[1], D De Girolamo[1], A Falabella[1], E Fattibene[1], G Maron[1,3], A Prosperini[1], V Sapunenko[1], S Virgilio[1] and S Zani[1]

[1] INFN-CNAF, Bologna, IT
[2] GARR Consortium, Roma, IT
[3] INFN-LNL, Legnaro, IT

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

1. Introduction

Currently, the INFN Tier-1 hosts nearly 1,000 worker nodes (WNs) for a total amount of about 21,500 computing slots (equivalent to power capacity of 200k HS06), $\sim$ 22 PB of net disk space and a tape library with, currently, $\sim$ 45 PB of used tape space. These resources are available to $\sim$ 30 scientific collaborations, besides the four LHC experiments. The computing resources are used all the time at CNAF with a large amount of waiting jobs ($\sim$ 50% of the running jobs).[1] This is paired by a good efficiency in data access and hence in job processing. In the next years, a huge increase of resource request is foreseen, primarily due to the WLCG experiments, and a non-negligible contribution will come from astro-particle ones, noticeably CTA[2]. By the end of 2023, i.e. at the end of LHC Run3, the amount of needed CPU power at INFN Tier-1 will be a almost a factor 5 larger than the quantity currently installed; roughly, the same scaling factor applies for disk and tape storage (Fig.1 and Tab.1).

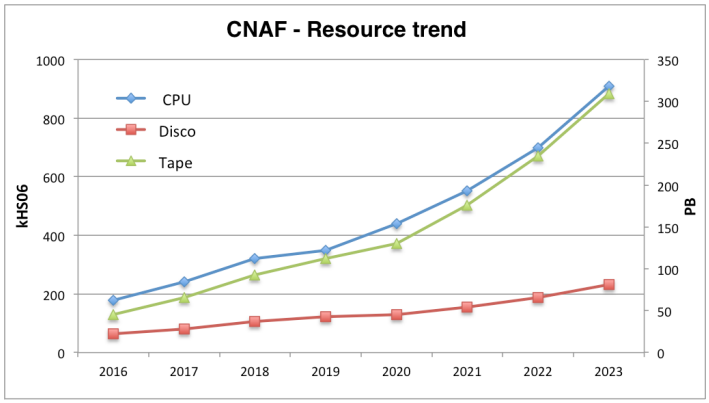


Figure 1: The foreseen trend of resources at INFN Tier-1 up to 2023

[1] The irregular profile visible on the first three months is due to "short jobs flooding" whose negative effect has been addressed and prevented as described in [1].
[2] Cherenkov Telescope Array https://www.cta-observatory.org

	CPU (kHS06)	Disk (PB-N)	Tape (PB)
2016	178	22	45
2017	242	28	66
2018	321	37	93
2019	349	43	112
2020	439	46	130
2021	552	54	176
2022	698	65	235
2023	910	81	309

Table 1: Forecast of installed resources at INFN Tier-1

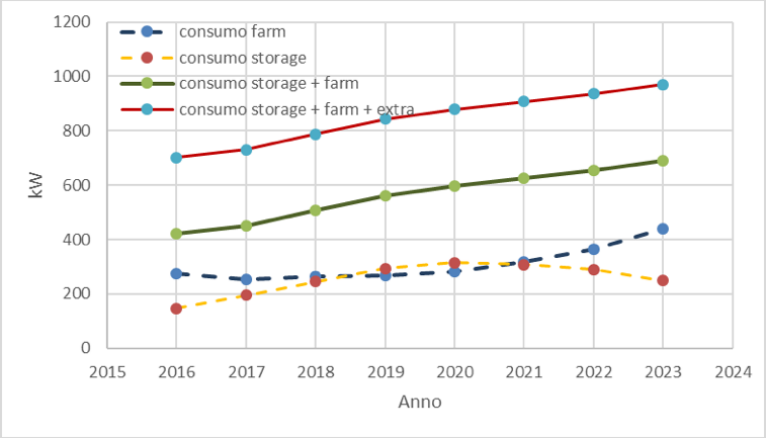


Figure 2: Extimated total power consumption (farm + storage)

2. Towards 2023

Taking into account the foreecast of the needed resources up to end of LHC Run 3 and the technological trend, we have extimated, year by year, the total power consumption (Fig. 2) and the occupied space in the data center (Fig. 3). Both the foreseen footage and the power consumption, are compatible with our data center infrastructure. In particular, while the maximum available electrical power is $3.8MW$, we have extimated a peak value of $\sim 1MW$ of load due to the IT components corresponding to a grand total of $\sim 1.6MW$[3]. The only necessary interventions on the infrastructure, not directly related to the increase of resources, are the extraordinary maintenance of the power continuity system and the replacement of the chillers.

In any case, while we have started preparing the upgrade of the chillers (see the Chapter "Tier-1 chiller upgrade plan"), we are also testing the usage of remote farms as a transparent (and possibly dynamic) extension of the Tier-1. In this way we will also be able to cope with unplanned requests of CPU resources, i.e. to cover peaks via the so-called cloud-bursting. The tests are performed according to two different (but complementary) scenarios:

[3] Currently we register a $PUE \sim 1.6$ and we know that this should be lower with newer chillers and with higher IT load.

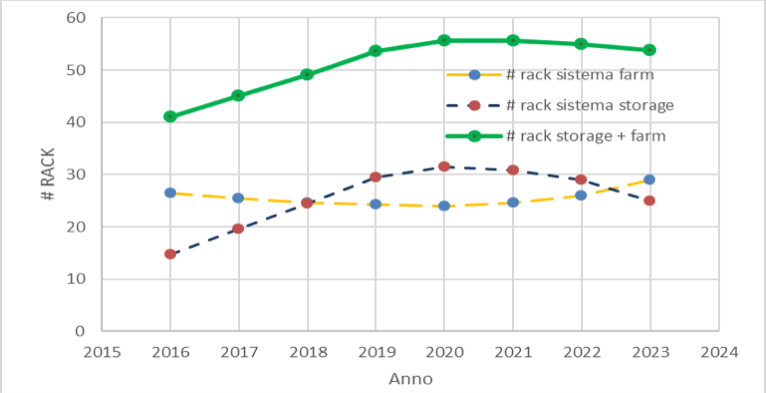


Figure 3: Extimated total space usage (farm + storage)

- Opportunistic computing on commercial Cloud;
- Static allocation of remote resources.

The first case has been addressed through tests with the Italian Cloud provider Aruba [2] while for the second one, a test has been performed during the 2016 extending INFN Tier-1 farm to Bari-ReCaS data center for the WLCG experiments. INFN Tier-1 is also participating to the European Pre Commercial Procurement project HelixNebula Science Cloud (see Chapter "Helix Nebula Science Cloud Pre-Commercial Procurement"). In the following sections, the setup and the results of a pre-production extension of the INFN-Tier-1 to the Bari-ReCaS data center will be presented.

3. Bari-ReCaS

The data center Bari-ReCaS is a common effort of INFN and Università degli Studi di Bari "Aldo Moro". The distance between Bari-ReCaS and CNAF is $\sim$ 600 km with a Round-Trip-Time of $\sim$ 10 ms (Fig. 4). Part of the resources ($\sim$ 25k HS06, 1.1 PB of disk) are allocated to CMS and Alice Tier-2. Following an agreement with Bari-ReCaS, 48 servers, for a total amount of $\sim$ 20k HS06 of CPU power, were allocated to the INFN Tier-1 as extra-pledge resources for WLCG experiments. These servers provide $\sim$ 13% of additional computing power to WLCG experiments at the INFN Tier-1.

The primary goal of our test was to establish an extension to the INFN Tier-1 which would be transparent to users, meaning that the entry point for jobs would remain CNAF, while they could be dispatched to Bari-ReCaS nodes. A similar case is the one between CERN and Wigner[4] with the notable difference being the absence of storage in Bari-ReCaS (excluding a small cache for LSF and transient data).

Figure 4: Relative positions of CNAF and Bari-ReCaS

3.1. Network setup

The first step was the setup of a dedicated connection between INFN Tier-1 and Bari-ReCaS data center. The network link was dimensioned allowing $\sim$ 1 MB/s per core (on the Tier-1 LAN this value is slightly higher, 5

[4] http://wigner.mta.hu/wignerdc/index_en.php for more details.

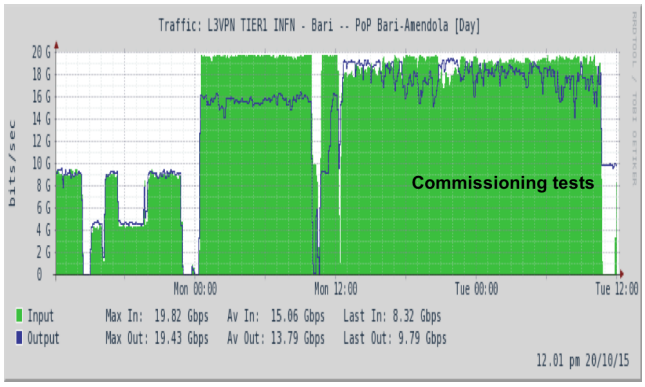


Figure 5: Commissioning of the VPN

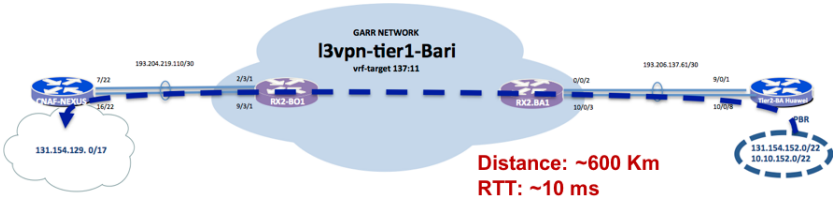


Figure 6: Layout of the VPN

MB/s per core). Hence for 2,000 cores a 2x10 Gbps link has been established (in Fig.5 the commissioning of the link performed during the last quarter of 2015). This link is configured as a Level 3 VPN and two /22 CNAF subnets (one for the public IP addresses of WNS and the other one for their management interfaces) are being routed over this link and assigned to Bari-ReCaS nodes (obviously only a small fraction of these addresses is currently used). Even if connected to the common core switch in Bari-ReCaS, these nodes are isolated from the other ones in the data center: they access the WAN, including LHCONE and LHCOPN, through CNAF (see the layout of the connection in Fig.6).

3.2. Farm setup

WNs have been configured as part of CNAF LSF cluster and the entry points for the users are the standard Computing Elements of INFN Tier-1. A cache for the shared file-system for LSF has been configured. Some auxiliary services (namely the CVMFS Squid servers for software distribution and the Frontier Squid servers for the condition databases of ATLAS and CMS) have been replicated in Bari-ReCaS. Using different "views" in the DNS zone, WNs in Bari-ReCaS are redirected to the local instances instead of the ones at CNAF. For all other services, including Kerberos for authentication, LDAP for authorization puppet/foreman fir the installation and configuration, the WNs access the instances at CNAF.

3.3. Data access

To match the requirement of transparent use of remote resources, jobs need to access data the same way as they do at CNAF. Online storage data at CNAF are managed through GPFS file-systems; remote mounting from disk servers at CNAF on the WNs in Bari-ReCaS of these file-systems is infeasible because of the excessive Round Trip Time ($\sim$ 10 ms) of the network connection. Some of the experiments can use Xrootd to remotely access their data as a default protocol (i.e Alice) or as a fallback one (i.e. CMS). However this is not a general enough solution,

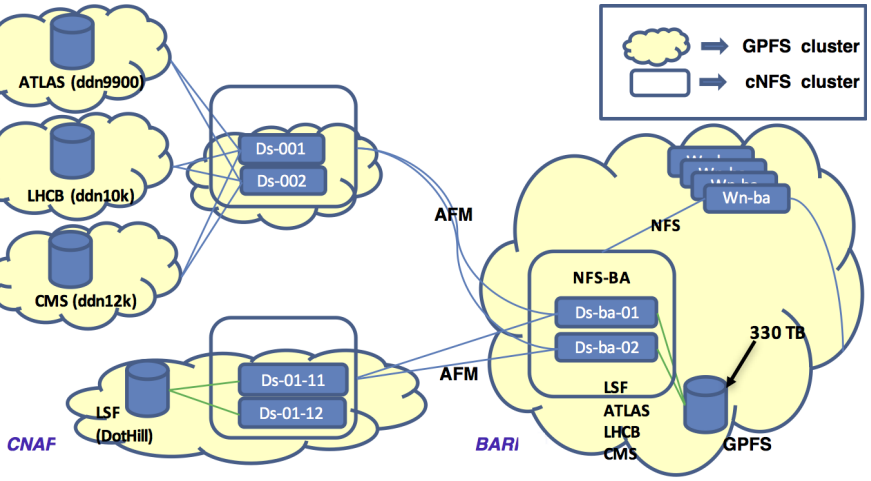


Figure 7: Layout of cache system in Bari-ReCaS

especially considering non–WLCG experiments.

The natural choice to implement the Posix access for WNs in Bari-ReCaS was to exploit a native extension of GPFS, AFM[3, 4], capable to implement a caching system (Fig. 7). Two 10 Gbps servers with $\sim$ 330 TB-N of disk (initially 100 TB-N only, exhibiting however too poor performances) have been allocated in Bari-ReCaS for AFM; the cache is configured in read–only mode to improve performances, as the output is written directly to CNAF using StoRM[5]. Alice data are not cached since their jobs only use remote access via Xrootd. One more small AFM cache, decoupled from the one for the data, has been configured for the LSF shared file-system.

Not surprisingly, the local cache access results to be critical and it is the potential bottleneck for the I/O intensive jobs. With the first incarnation of the cache (100 TB of net disk space) we observed a maximum usable bandwidth of $\sim 1\ GB/s$ for reading or writing exclusive access, while concurrent read/write accesses caused the degradation of the performances down to 100 MB/s. In this configuration, the space allocated to each experiment ($\simeq 30\ TB - N$), can be largely insufficient: while Atlas and LHCb use only 10% of the disk space, CMS can fill it in 12 hours. In this situation, the cache acts like a pass-through buffer introducing a delay, thus causing a degradation of the efficiency for all the jobs (Fig.8). In such a case, a better option resulted to be disabling the access to the cache for CMS jobs and relaying on Xrootd fallback.

We could increase the performances of the cache enlarging its size (up to 330 TB of net disk space) and with an ad hoc tuning of GPFS/AFM to take into account the specific hardware characteristics.

Anyway, even if we have performed several tunings of the cache system, we have found some limiting factors in the current infrastructure: the size of the cache (it can contain up to 3% of the amount of data stored at CNAF), the maximum fill-in and fill-out speed (1 GB/s r/w aggregated) and obviously the pattern of data access which is not tunable at our side.

4. Results

The configuration of the farm and the caching system in Bari-ReCaS is stable since June 2016: since then the farm has been extensively and steadely used (Fig. 9) with swinging results. Obtaining a reliable comparison between CNAF and Bari-ReCaS is quite difficult because the workload on the WNs of the two partitions is different: CNAF nodes are concurrently used by several VOs, with usually different mix, while only WLCG VOs can use WNs at Bari-

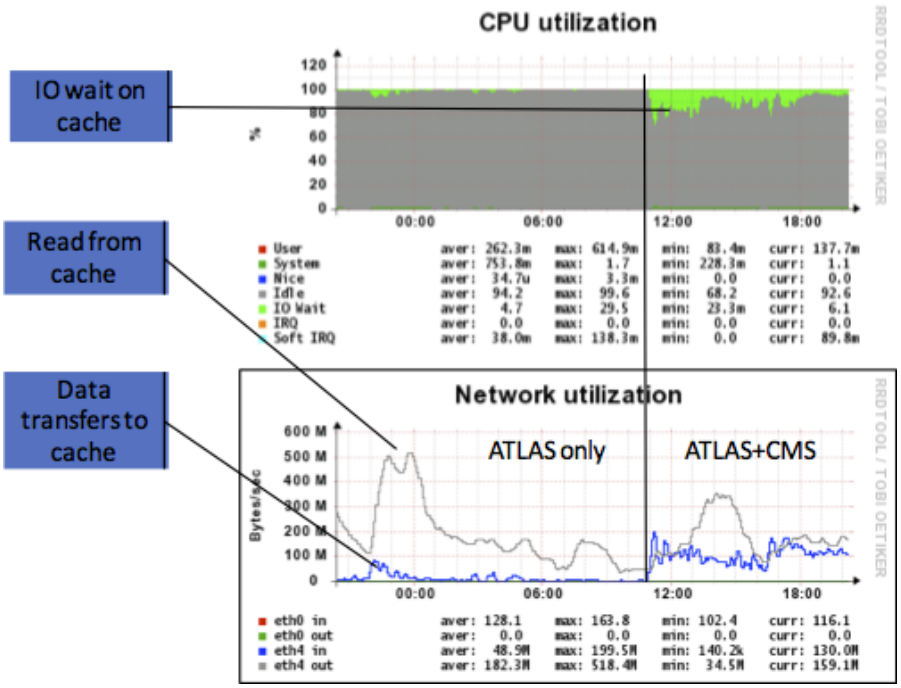


Figure 8: Saturation of the cache system due to CMS jobs

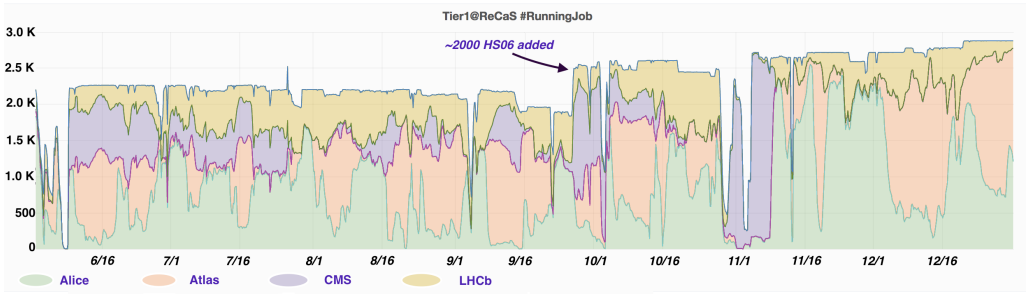


Figure 9: Farm usage in Bari-ReCaS in the second half of 2016

ReCaS. Hence the efficiency of jobs[5] can be penalized, expecially at CNAF, from occasional CPU overload due to jobs of other experiments running on the machine, resulting in a temporarily higher efficiency for Bari-ReCaS jobs.

On the long term, we have anyway observed that efficiency is better for I/O demanding jobs running at CNAF, noticeably for Atlas and CMS. The main reason is the low speed of the cache: in certain cases we have observed the data flowing continuously from CNAF to the cache and then to the WNs, with the cache introducing a delay. As an attempt to overcome the problems with the cache, we plan to test a new and more performant storage system during the first semester of 2017. In the meanwhile, we have mitigated the issue by defining an ad-hoc queue for Atlas in LSF to dispatch low-I/O jobs only to Bari-ReCaS WNs; moreover we have had at times to inhibit the access to CMS jobs. Noticeably, LHCb jobs show a high efficiency also in Bari-ReCaS. On the other hand, Alice jobs, that do not use the cache, have an efficiency

[5] The efficiency of a job is defined as the ratio between the used CPU and Wall-clock time.

Exp	CNAF	Bari-ReCaS
Alice	0.87	0,87
Atlas	0.81	0.81
CMS	0.75	0.67
LHCb	0.93	0.95

Table 2: Overall comparison of jobs efficiency (second half of 2016)

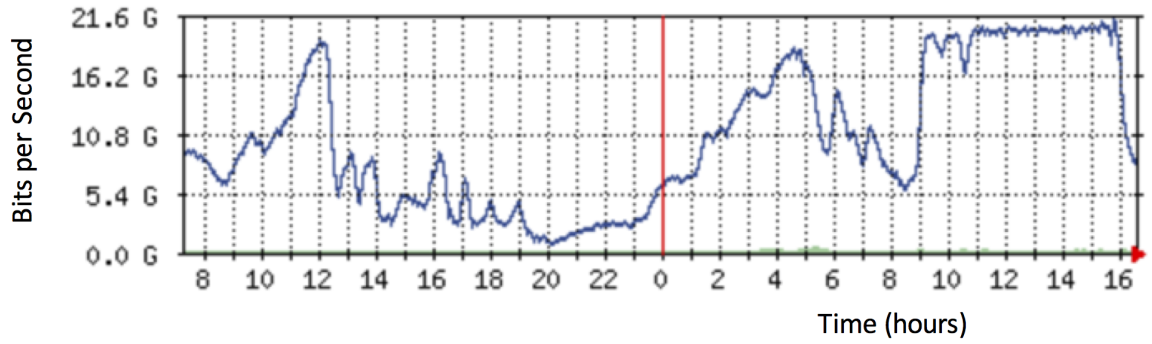


Figure 10: Saturation of VPN due to Alice jobs

comparable to that of jobs running at CNAF (the difference is minimal); indeed remote data access could be a viable option, probably having a dedicated channel with a bandwidth larger than 20 Gbps. In any case, this method could not be extended to non-LHC experiments which need Posix access to the data.
Given these constrains for the use of Bari-ReCaS partition, we have, with the exception of CMS, comparable efficiency with CNAF (Tab.2).
Furthermore, with the current configuration, the remote access exploited by Alice jobs can occasionally saturate the VPN for some hours, hence interferring with the cache system (see for example Fig.10). For these reasons, we are also considering to increase the bandwidth of the VPN: this will be possible starting from the second half of 2017.

5. References

[1] See Chapter "Protecting the batch cluster from short job flooding"
[2] Dal Pra, S., Ciaschini, V., dellAgnello, L., Chierici, A., Di Girolamo, D., Sapunenko, V., ... & Italiano, A. (2016, March). Elastic CNAF DataCenter extension via opportunistic resources. In Proceedings of the International Symposium on Grids and Clouds (ISGC) 2016. 13-18 March 2016. Academia Sinica, Taipei, Taiwan. Online at http://pos. sissa. it/cgi-bin/reader/conf. cgi? confid= 270, id. 31.
[3] Quintero, D., Ceron, R., Dhandapani, M., da Silva, R. G., Ghosal, A., Hu, V., ... & Velica, S. (2013). IBM Technical Computing Clouds. IBM Redbooks.
[4] Sapunenko, V., D'Urso, D., dell'Agnello, L., Vagnoni, V., & Duranti, M. (2015). An integrated solution for remote data access. In Journal of Physics: Conference Series (Vol. 664, No. 4, p. 042047). IOP Publishing.
[5] Carbone, A., dell'Agnello, L., Forti, A., Ghiselli, A., Lanciotti, E., Magnoni, L., ... & Zappi, R. (2007, December). Performance studies of the StoRM storage resource manager. In e-Science and Grid Computing, IEEE International Conference on (pp. 423-430). IEEE.

Helix Nebula Science Cloud Pre-Commercial Procurement

D. Cesini, A. Chierici and L. dell'Agnello
INFN-CNAF, Bologna, IT
E-mail: luca.dellagnello@cnaf.infn.it

1. Introduction

Helix Nebula[1] is a new, pioneering partnership between big science and big business in Europe that is charting the course towards the sustainable provision of cloud computing - the Science Cloud. The partnership brings together leading IT providers and three of Europe's leading research centres, CERN, EMBL and ESA (see Figure 1) in order to provide computing capacity and services that elastically meet big science's growing demand for computing power. In September 2015 a proposal for a "Pre-Commercial Procurement" in phase 8 of Horizon 2020 ICT call was approved. This PCP will allow to implement a prototype of an hybrid cloud with commercial cloud providers. INFN is one of the procurer and the main activity has been carried-out by CNAF.

Figure 1. HNSciCloud Procurers

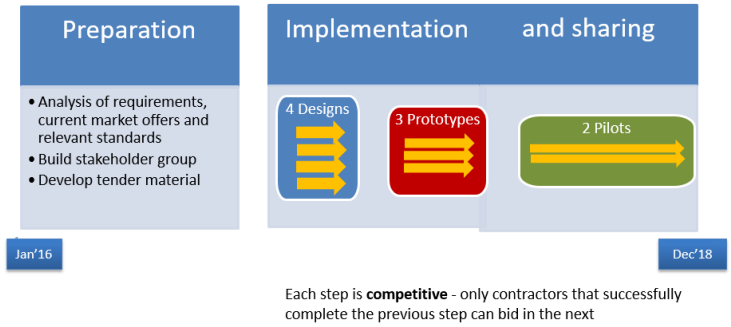


Figure 2. Helix Nebula Science Cloud PCP project phases

2. First year of PCP

During this first year two main activities were completed. First of all a tender was prepared in order to identify commercial cloud providers interested in participating in the project. Among the many participants 4 were selected, showing enough knowledge and the ability to answer all the requirements of the tender. After that, the Design Phase started, were the 4 selected bidders were asked to provide a detailed documentation with technical details about how to implement all the required features included in the tender (see Figure 2 for details on project phases). Shortly after the successful project review in Brussels on 8 December 2016, the 4 consortia contracted for the design phase gave an update on their progress during a series of telcos: these updates indicated good progress is being made, with the contractors being confident to deliver their design by the end of December. The updates also raised questions about the provisions for transparent data access as well as federated identity management. These points have been followed-up by via weekly telcos with the contractors. The deadline for the contractors to submit their designs and associated deliverables is the 30 January 2017.

3. References

[1] Helix Nebula Science Cloud webpage: http://www.helix-nebula.eu

The INFN Tier-1: the Facility Management group

M. Onofri, M. Donatelli, A. Mazza and P. Ricci
INFN-CNAF, Bologna, IT
E-mail: pierpaolo.ricci@cnaf.infn.it

1. Introduction

The data center Facility Management group takes care of the management and support of technological infrastructures of CNAF, firstly the power distribution and cooling systems of the data center.

The group is also responsible for:

- Management of control systems of CNAF;
- Management of general services (including phone, power, cooling systems for CNAF offices);
- Management of warehouse at CNAF;
- Delivery of all hardware resources for CNAF;
- Procurement of supplies for IT at CNAF.

Besides the standard duties, during 2016 the group has been working on two projects:

- the migration to SBO, a new Building Management System to replace TAC Vista in use since 2008 and now being phased out from the producer;
- the study of the upgrade of chillers.

2. Migration to the new Building Management System

The INFN Tier-1 data center is composed by two different main rooms containing IT resources and four additional locations that hosts the necessary technology infrastructures providing the electrical power and cooling to the facility. The power supply and continuity are ensured by a dedicated room with three 15,000 to 400 V transformers in a separate part of the principal building and two redundant 1.4MW diesel rotary uninterruptible power supplies. The cooling is provided by six free cooling chillers of 320 kW each with a N+2 redundancy configuration. Clearly, considering the complex physical distribution of the technical plants, a detailed Building Management System (BMS) was designed and implemented as part of the original project in order to monitor and collect all the necessary information and for providing alarms in case of malfunctions or major failures. After almost 10 years of service, a revision of the BMS system was somewhat necessary. In addition, the increasing cost of electrical power is nowadays a strong motivation for improving the energy efficiency of the infrastructure. Therefore, the exact calculation of the power usage effectiveness (PUE) metric has become one of the most important factors when aiming for the optimization of a modern data center. For these reasons, an evolution of the BMS system was designed using the Schneider StruxureWare infrastructure hardware and software products (SBO). This solution proves to be a natural and flexible development of the previous TAC Vista software with advantages in the ease of use and the possibility to customize

the data collection and the graphical interfaces display. Moreover, the addition of protocols like open standard Web services gives the possibility to communicate with the BMS from custom user application and permits the exchange of data and information through the Web between different third-party systems. Specific Web services SOAP requests has been implemented in our Tier-1 monitoring system in order to collect historical trends of power demands and calculate the partial PUE (pPUE) of a specific part of the infrastructure. This would help in the identification of spots that may need further energy optimization. The StruxureWare system maintains compatibility with standard protocols like Modbus as well as native LonWorks, making possible reusing the existing network between physical locations as well as a considerable number of programmable controller and I/O modules that interact with the facility. The high increase of detailed statistical information about power consumption and the HVAC (heat, ventilation and air conditioning) parameters could prove to be a very valuable strategic choice for improving the overall PUE. This will bring remarkable benefits for the overall management costs, despite the limits of the non-optimal actual location of the facility, and it will help us in the process of making a more energy efficient data center that embraces the concept of green IT. The migration from the TAC Vista to the SBO software of the whole infrastructure was easily split into three phases over an eight-week period. The software relies on two separated servers for providing the core of the software management, the web user interface server and the long-term archiving backend:

- Enterprise software server (ES): it runs the core software services for the management, configuration and backup operation of the system.
- Report server: it is used for archiving the long-term trends of the collected variables and it also includes advanced reporting options. It uses a Microsoft SQL Server database for storing all the information.

The two servers run on virtual Microsoft Windows machines and are used for managing the real engines of the systems. The real engines are actually three Schneider Automation Server (AS) devices which are located in three strategic physical location of our buildings (i.e. the Transformers Room, the Chiller Room and the Power Room). The SBO software is fully compatible with the Modbus protocol (TCP/IP and serial) which was heavily present in the previous TAC Vista implementation and it is also fully compatible with the TAC Vista Lonworks network, giving us the possibility to reuse the existing cabling and a great part of the hardware boxes. Furthermore, the web GUI user interface just needs a standard browser for working (HTML5 compatible) so it is easily accessible from mobile device. Compared to the old TAC Vista system the SBO architecture has increased the number of collected metrics and the overall archiving duration (thanks to the separated Report server and database layout). At present over 2500 metrics are being collected with a 15 minutes granularity which gives us the opportunity to store over 10 years of trend history. The optimization of variables collection has also been implemented in order to reduce load (e.g. a power switch condition is logged only when a change occurs) and an intuitive system GUI has been directly implemented in the graphical schemas, providing a simple and fast access to the end user. In addition to the standard variable metrics, the Power Usage Effectiveness (PUE) is calculated and collected. The PUE is a measure of how efficiently a data center uses energy; specifically, how much energy is used by the computing equipment in contrast to cooling and other overhead. We have also introduced the partialPUE (pPUE) metric in order to monitor the power demand of a specific area of the infrastructure and try to optimize it. For example, and indicator of the rotary diesel power continuity energy loss can be calculated with the following formula:

$$POWER\ CONTINUITY\ pPUE = \frac{UPS_{Energy} + IT_{Energy}}{IT_{Energy}}$$

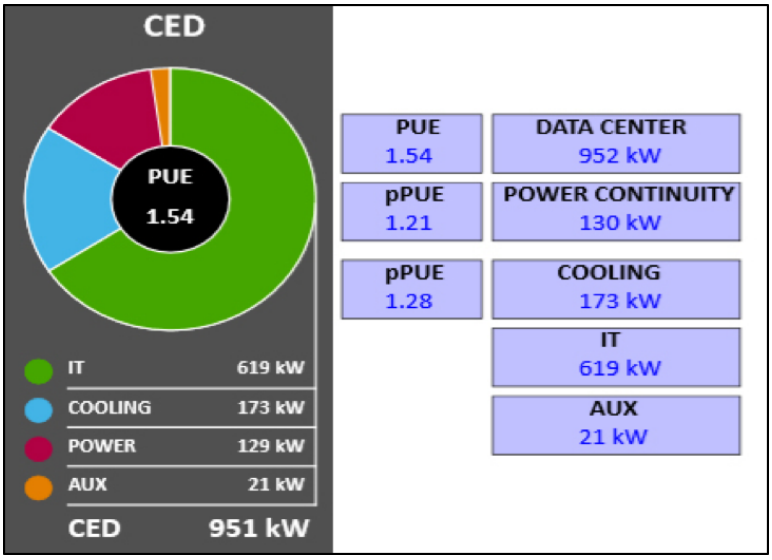


Figure 1: The total power and PUE graphical page.

where UPS_{Energy} is the loss due to the UPS rotary and the IT_{Energy} is the total amount of energy provided to the IT equipment in the 2 data center rooms. In Fig.1 an extract of the PUE report layout designed in the SBO interface is reported. On the left side of the figure a pie chart reports the ratio between the principal power distribution areas of our center: IT, COOLING (including chillers, pumps and air-handler supply), POWER (including UPS and distribution losses) and AUX (all the auxiliary electrical system that are not specifically IT equipment). The pPUE values and the total power pie chart give an immediate visual overview of the power distribution of our facility and all the relevant value can be easily extracted.

We can outline that the choice of migrating our BMS system to the Schneider SBO software has proven to be successful since it has showed a good reliability and great compatibility with the previous hardware and software installation. Our PUE and pPUE analysis clearly shows that we need some improvement. The seasonal PUE value of about 1.5 is a clear indication of the low energy-cost effectiveness. The pPUE analysis has shown that a more consistent improvement would be gained increasing the chillers efficiency and therefore a new project concerning a chiller technology refresh has already been developed and will be fulfilled in the next months. Also, a finer granularity of rack power consumption measurement could help the optimization of electric and cooling power distribution and for this reason an increase in the number of metered PDUs with TCP Modbus support will be considered among the future hardware installations. Eventually the SBO compatibility with open standards like Web Services improved our BMS integration and “open-mindedness”, and the alignment of communication protocols to not-proprietary ones, in particular Modbus, will permit the integration of different platforms (e.g. Arduino custom sensor probe that are currently tested) under a single BMS system.

Tier-1 chiller upgrade plan

A. Mazza
INFN-CNAF, Bologna, IT
E-mail: andrea.mazza@cnaf.infn.it

1. Tier-1 chiller upgrade plan

The technical and economic feasibility study prepared by CNAF aims to a gradual and progressive renewal of the refrigerating plant with the integration of new air-cooled chillers. The temporal distribution of installations and any future disposal chillers has been thought to minimize adverse effects on the activity of Tier 1, since the operation of the data center must be guaranteed 24/7. This particular solution relies on a new approach that breaks up with the past choices: it has to be adapted to the background and not vice versa.
In this affirmation, the word "environment" is considered in these terms.

- Location: the shape of the building, the university context, the existing infrastructure and the air quality and climate are all constrains prompting to a compact, low-noise, low-weight solution.
- Environment: the concern for low environmental impact and high efficiency solutions.
- Sustainability: the reduction of the life cycle costs.
- Uncertain load trend: the implementation of a modular approach and efficient solution at part-load conditions.

To meet all these requirements, the chosen solution includes custom chillers with these three main features:

- oil-free centrifugal and variable speed compressors;
- low environmental impact refrigerant;
- chiller with wide remote condensing unit (split system).

Referring to the nominal data of the new chillers and the load and the outdoor temperature trends of the last three years, the actual energy breakdown is shown in Fig. 1 whilst the attended one is shown in Fig. 2.

On the one hand, the choice of custom chillers implies the need of more engineering and cooperation between the design studio and the chiller supplier along with related extra costs. On the other hand, this solution will preserve the same maintenance and management efforts that are normally required for a more common solution, as in the case of single-block, air-cooled chillers. This happens because the custom chiller has actually the same components of a single-block chiller but the condenser and the evaporator are split and disposed in different locations, even if, of course, they are linked with long gas pipes. The minimum savings expected from this upgrade are roughly $130,000.00$ € per year, which corresponds to a payback period of about 3 years.

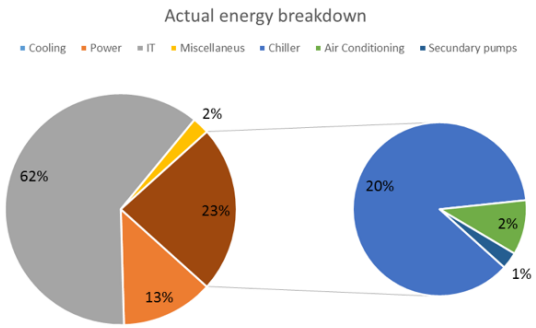


Figure 1: Retrieved data
(IT load = 670 kW and PUE = 1.63)

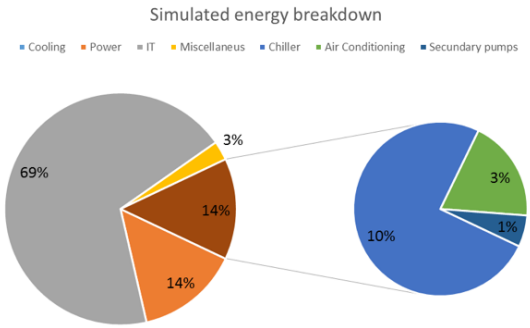


Figure 2: Simulate scenario
(IT load = 670 kW and PUE = 1.45)

National ICT Services Virtualization Infrastructure

S. Antonelli, L. Chiarelli, D. De Girolamo, S. Longo, M. Pezzi, F. Rosso, S. Zani

INFN-CNAF, Bologna, IT

E-mail: Stefano.Antonelli@cnaf.infn.it, Lorenzo.Chiarelli@cnaf.infn.it, Donato.Degirolamo@cnaf.infn.it, Stefano.Longo@cnaf.infn.it, Michele.Pezzi@cnaf.infn.it, Felice.Rosso@cnaf.infn.it, Stefano.Zani@cnaf.infn.it

Abstract. In this report we describe the new Highly Available Virtualization Infrastructure deployed during 2016 by National ICT Services. The attention will be focused particularly on the production part, that is the portion of the infrastructure employed to provide services INFN-wide. Designing the system we have focused our attention in particular on reliability and availability: in this report we will point out those aspect and the solution adopted for Storage, Networking and Virtualization.

1. Introduction

The infrastructure deployed at CNAF by National ICT Services is designed to assure services availability even in the case of a failure in its components, replicating more critical hardware parts and using software and protocols able to proactively manage the infrastructure, when a failure appear.

2. Storage Area Network

Both data and applications are stored inside a SAN implemented with DELL EqualLogic series PS6000 systems, platform chosen for its reliability and scalability characteristics. At present three appliances are running in a single group, that is a set of systems acting as a single SAN where archiving space and I/O load is balanced collaboratively among servers, furthermore implementing tiering among the available disk technologies.

Currently the setup of EqualLogic systems is as follows:

1 x PS6110E with two controllers, 2x10GbE connectivity and 24x2TB NL-SAS (7200 RPM) Hard Drives (for a net total of 33,7 TB in RAID6 setup with one hot spare)

1 x PS6210E with two controllers, 4x10GbE connectivity and 24x4TB NL-SAS (7200 RPM) Hard Drives (for a net total of 68,2 TB in RAID6 setup with one hot spare)

1 x PS6210XS with two controllers, 4x10GbE connectivity and 7x400GB SSD HardDisk plus 17x600GB SAS (10000 RPM) Hard Drives (for a net total of 9,1 TB in RAID6 *Accelerate* setup with one hot spare)

Every appliance's storage is configured to implement a RAID6 array, providing resiliency for two disk faults without any discontinuity in service execution; moreover using one hot spare for each EqualLogic we are able to guarantee that in case of an HD fault, the array reconstruction

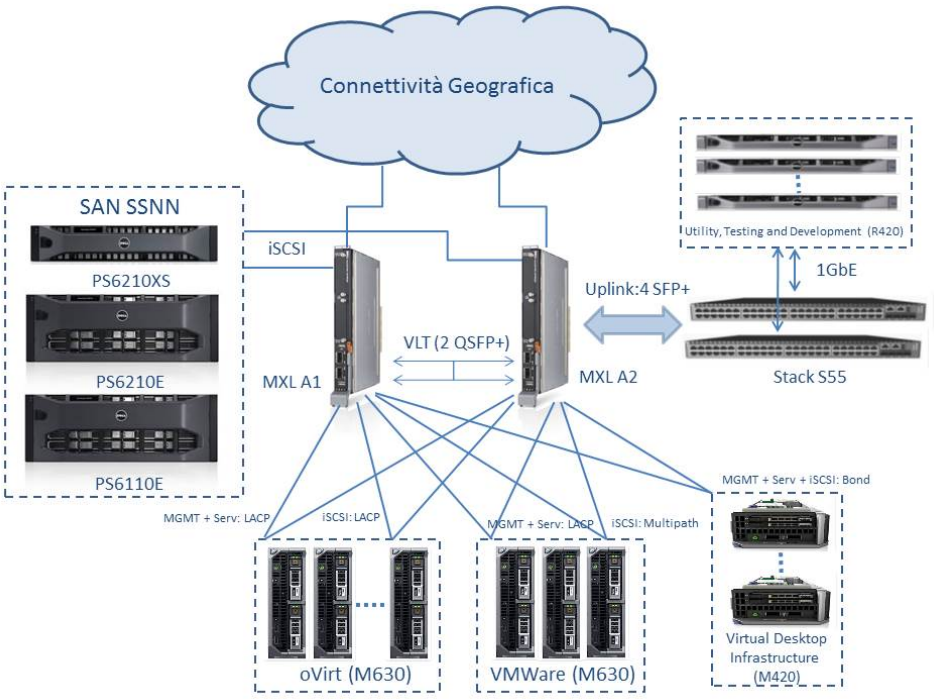


Figure 1. National ICT Services Infrastructure Scheme

starts immediatly. EqualLogic systems are redundant in all the critical parts, in particular they are configured with double power supply and with two controllers with vertical failover (active + standby setup with synchronized cache and the ability of each controller to employ network connectivity from its partner in presence of a failure).

We would like to point out that in addition to resiliency, this kind of storage was also chosen for the horizontal scaling characteristics: it is possible to have up to 16 EqualLogic PS6000 systems in a single group, mixing any type of appliance and hard drive available on the market. EqualLogic systems are also able to provide interesting features regarding data security, providing in particular the manual/automatic management of snapshot policies and the synchronous/asynchronous replications of volumes on remote partners.

3. Network

Having in mind High Availability and Reliability (HA), the network was designed to assure the presence of two distinct paths to any critical resource (storage, hypervisors, etc.). The core of the network infrastructure is made up with two Force10 MXL switches, embedded in a DELL M1000e blade enclosure hosting also the production servers, profiting in this way of the redundant power supplyes, cooling and management provided by the enclosure. The setup chosen for production provides 32 internal ports at 10GbE, 4 external QSFP+ ports at 40GbE and 4 external SFP+ ports at 10GbE for each switch. Network paths redundancy is implemented via VLT (Virtual Link Trunking) protocol, using two of the available QSFP+ ports on each switch. VLT protocol allows to setup network paths inside a *domain* (a set of switches implementing VLT) dinamically according to link status, similarly to Spanning Tree algorithms. As in a stack, VLT let switches inside a single domain to behave like a single device, allowing the setup of aggregated paths (LACP - 802.3ad) terminated on ports belonging to different switches. Unlike a stack however, VLT maintain network apparatus indipendency, allowing to set offline a portion of the VLT

domain for maintenance (e.g. firmware updates, etc.) without affecting network connectivity, clearly as long as inside the domain there is still an online path connecting source and destination of a communication.

For devices directly connected to the network core, redundancy is implemented by doubling network connections and by connecting paths from a single system to both switches. As example, for SAN connectivity each EqualLogic has the first controller connected to switch A1 and the second one to switch A2, employing external QSFP+ ports. A similar approach was chosen for hypervisors (servers running services) connecting NIC ports to both switches and setting up aggregated links via LACP. Finally each MXL switch is provided with an optical link for geographic connectivity.

4. Hypervisors

Not only to provide a bettere level of managent but also for providing operating economy, we have opted to run almost all the applications managed by National ICT Services on virtual systems. From an hardware point of view, the virtualization infrastructure hosting production services is made with DELL PowerEdge M630 servers, two ways half size blades equiped with Intel Xeon E5-2650/E5-2660 v3 CPU, 128GB or 256GB of ECC RAM, two 300GB SAS (10000 RPM) hard drives configured in RAID1 via an hardware controller and four 10GbE ports (with hardware offload of most common protocols, in particular TCP, iSCSI and FCoE). For each blade the four network ports are connected half on switch A1 and half on switch A2, with LACP protocol to double available network paths.

M630 Virtualization systems are managed with two different platforms. The biggest one (at present 7 active M630 plus 3 standby ones) is managed via oVirt, the open source version of Red Hat Enterprise Virtualization Manager. Services virtualization is performed by linux operating systems - in particular CentOS 7 - using KVM and QEMU libraries. Servers are then managed via oVirt, realizing in this way an HA infrastructure. Alongside this first infrastructure we also run a second production one managed by VMWare vSphere 6.5 Essential Plus, for applications requiring a certified platform. Both oVirt and VMWare datacenters are employed to provide Highly Available applications: management software monitors continuosly both hypervisors and services; in case of an incident, management software interacts with the infrastructure assuring service quasi-continuity (e.g. in case of an hypervisor fault, management software reschedules any service running on the faulty device on a different, available and healthy server). oVirt infrastructure is also setup to manage proactively the infrastructure to gain optimal service performance, migrating VM from heavy loaded servers to free hypervisors. The two infrastructure managers are also used to collect system metrics and to facilitate daily services and systems management.

National ICT Services also uses several ancillary systems as services and development systems (e.g.: monitoring, backup, etc.). Tipically these systems employ part of the production infrastructure, in particular netwoking and storage, but they are placed apart from the production infrastructure on different network apparatus (Force10 S55 in fig. 1); in this way a fault in any auxiliary system will not spread in the production environment.

The described virtualization infrastructure allows the provisioning of Highly Available services, being able to resist to common hardware problems (i.e. fault of a single component of the infrastructure or several hypervisors). During its operations in 2016, the infrastructure has performed pretty well: in the year there was only one unexpected down of a national service, fault that had a recovery time of more or less one business day.

The INFN Information System

S. Bovina, M. Canaparo, E. Capannini, F. Capannini, S. Cattabriga, C. Galli, G. Guizzunti, S. Longo

INFN CNAF, Bologna, IT

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's 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 set up in 2001 with the purpose of digitizing and managing all the administrative and accounting processes of the INFN Institute, and of carrying 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 of developing, maintaining and coordinating 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. Infrastructure
In 2016 the infrastructure-related activity was composed of various tasks that can be summarized as follows. Firstly, the migration to a new hardware infrastructure, secondly the setup of some new services to improve our development process and software release workflow, thirdly the setup of a "Yum" and "Maven" repository (using Artifactory) and finally the setup of a playbook to automate and standardize common tasks using Rundeck.

More in detail we worked on:

- The consolidation of the monitoring system, obtained by the creation of specific dashboard with a focus on java applications and databases;
- The production of new enclosure based on DELL m1000e;
- The storage system reorganization;
- The migration of production applications on the new hardware infrastructure.

3. Time and attendance system improvements
Most part of 2016 was dedicated to the porting of the application to the new infrastructure. This task involved a lot of effort which was spent in various tasks such as the adaptation of the code to the new environment (new Apache-Tomcat, JDK, operating system version), the testing of all the functionalities as used by the administrative staff, the removal of the software packages not any more used, the new installation of the Oracle DB from 10g version to 11.2 and so on.

In addition to the migration, some software development activities have been conducted in 2016. We included in the system the opportunity, given to all the employees, to specify to have worked from home (teleworking). This new feature has become necessary given the new regulations that involved the Italian Public Administration. Furthermore, some regulations changed for what concerns the use of the parental leave. Thus, the system was modified in order to adopt the new legislation. In particular, it has been given the opportunity to the employees to specify the parental leave of a half workday duration, making its use more flexible.

3.1. Migration
In 2016 our effort was mostly spent in porting the whole application, included the database, from the old to the new infrastructure.

The migration has been divided into 2 steps:

- Upgrade of the Oracle DB from 10g version (on RedHat 5 operating system) to 11.2 (on Oracle Linux 6 operation system);
- Porting of application from Java 6 (on RedHat 5 operating system) to Java 7 (on CentOS 6 operating system).

The migration task has been particularly challenging mainly because the management of the old installation was in charge of an external vendor and did not follow any best practice or standard rules. In addition to that, we had to deal with a complex application part composed of several and dependent projects as shown in Table 1.

Table 1. Attendance system numbers

Resource	Amount
Project	77 (29 unused)
Jar	42
War	6
Applet	1
Repository	1

From a careful analysis we have identified the following weaknesses in the system:

- Application not designed for parallelization;
- Hard-coded configuration parameters;
- Wrong FQDN of the application (sysinfo-12 instead of presenze.infn.it);
- "Home made" SSL verification and wrong certificates management;
- No automated tests;
- Poor knowledge of the server setup (made by an external vendor several years ago);
- No use of system management tool (e.g. Puppet);
- No dependencies management system, and no Jar repository;
- Not standardized deploy workflow;
- Mandatory configuration changes after software releases;
- Confused organization of applet Jars;
- Missing version number in jar files;
- Monolithic repository and build process entrusted to the developers PC;
- Public application managed with the same ACL of backoffice application.

Once installed the new server, we decided to put it in production following a smooth procedure: we set up two application servers in parallel, working on the same database; by exploiting ACL, we gradually opened the access to the new server to the INFN branches. This way, we had the chance to have the application tested initially by a small sample of administrative staff and we managed to fix the various problems encountered without affecting all the INFN branches.
In order to succeed with this process we had to solve some issues: we had to introduce a "lock" system to manage concurrent process; we had to move all the hard coded parameters from the Java code to configuration files and finally, we had to change the FQDN of the application to presenze.infn.it.

The migration of the system also involved the introduction or upgrade of some of the technologies employed:

- Recompilation of the attendance system with Java 7;
- Migration from one monolithic SVN project, to several git projects;
- Creation of a build script, used by all projects;
- Activation of Continuous Integration (CI) on every project, automation of build and package processes (on Docker);
- Migration of the build system from Ant to Gradle;
- Addition of a dependency management system and Jar repo using Artifactory;

- Modification of the type of release artifacts from war to rpm;
- Complete revision the setup of the application and its configuration with Puppet;
- Separation of the backoffice and frontoffice context with different ACL sets;
- Implementation of a release workflow through Rundeck, Puppet and Gitlab-CI;
- Introduction of a load balancer.

4. Vamweb upgrade and access management system

In 2016, the INFN's access management system, consisting of a proprietary PHP application called "VAMWeb" that is centrally installed at CNAF, Bologna, was updated to the latest release, called "Vam4". This update, given the level of criticality, required several days of testing on a pre-production server. All the Entrance-Point configurations have been migrated to the new system and the Oracle Database, used by the system to read and write information, was updated from version 10g to 11gR2 and patched to be suitable for the new version of the Vam software. The update of the access management system has required a new installation of the software on a new virtual machine server, that was installed with a new operating system version and libraries.

In 2016, two new INFN locations have been configured in the access management system:

- One at CNAF, which has purchased and installed various hardware devices (Entrance Points) for access control, in particular to control the meeting rooms, the CED room and the Tier1 Computing Centre;
- One at Legnaro National Laboratories (LNL), that has installed numerous Entrance Points, for the access control to the dining hall, the internal library and the accelerator.

5. Protocol system migration

In 2016, we introduced the new INFN's protocol system, more pertinent to the current laws and developed with modern technology. Furthermore, it was necessary to save and make available all the data processed and stored by the old protocol system (Webrainbow), thus we exported from it all the data, metadata and attachments in excel files. All these files have been stored on the Alfresco document management system. The Alfresco configuration management guarantees that the access to the data files depends on the role of the user in order to ensure data confidentiality.

6. Oracle EBS improvements

6.1. Oracle EBS developments

In 2016, several developments were conducted, to improve the usability and functionality of the INFN ERP system, in particular we:

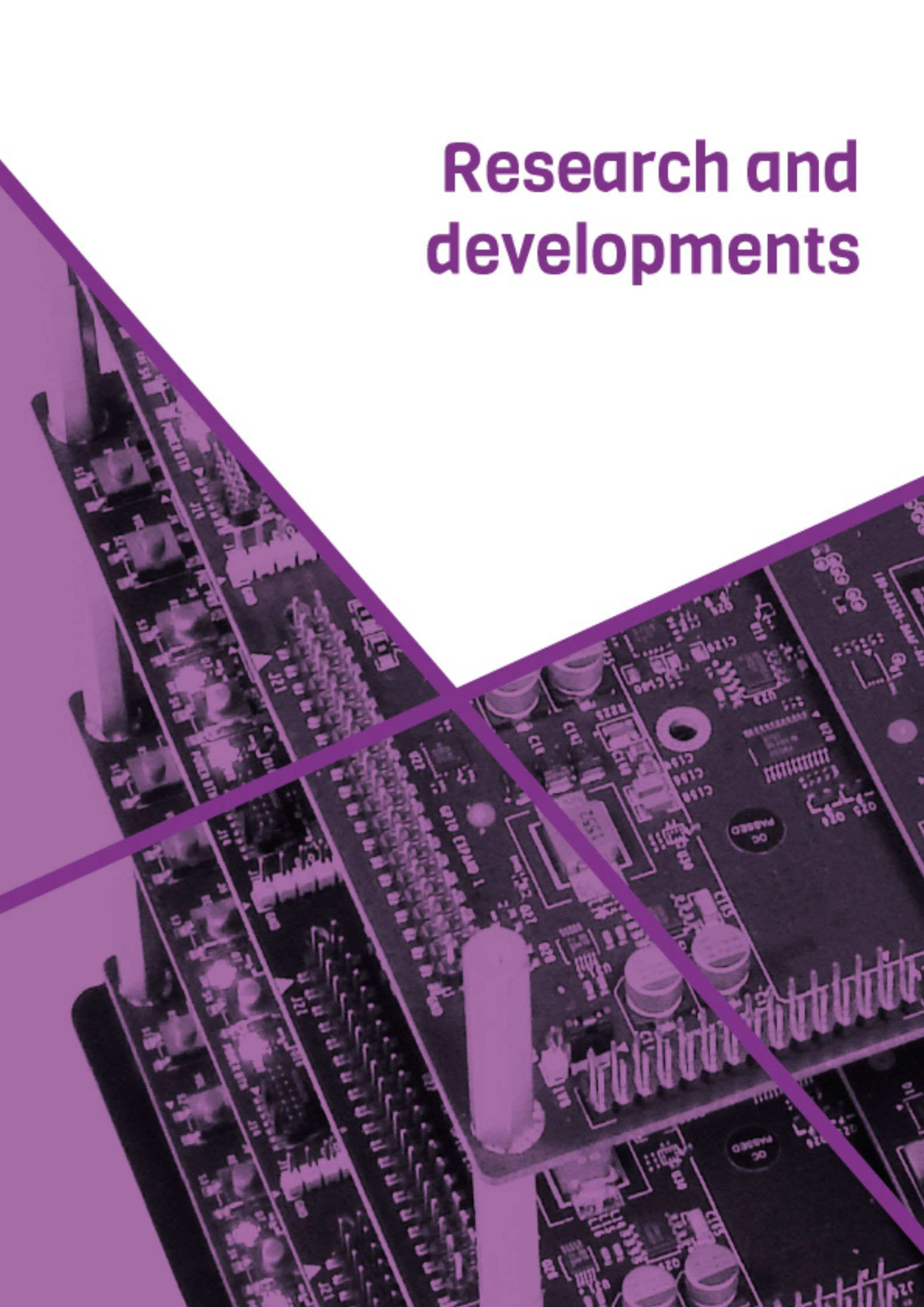
- Modified the procedure of communication with the bank, following the introduction of new "Piano dei conti integrato";
- Re-designed the system for creating and sending regularization movement (REG) flows;
- Created new form "Gestione Impegni/Accertamenti" (under menu: "Finanziaria Nativa");
- Implemented a procedure to import receipt and provisional data from the MIF logs;
- Created ad-hoc Oracle report for the calculation of invoice payment indices;
- Introduced several improvements to the "Anagrafica Fornitori" form and Oracle invoice registration system.

6.2. Oracle EBS Monitoring
Besides the standard EBS tools for monitoring, in 2016 some other PL/SQL tools were developed, registered and scheduled in the db to send notification via email in case of error.

6.3. HR improvements
In 2016, we worked on some improvements and bug fixing of the import mechanism implemented in 2015.

7. Disaster Recovery
Concerning some kind of personal data, a policy retention of 5 years has been established instead of the usual one of 30 days. The backup files are verified by scripts which execute the data restore, both local and remote, on a dedicated partition; the final result is compared with the original data. Both the data and database backups are periodically checked by restoring the service in the remote site.

Research and developments



Cloud@CNAF - maintenance and operation

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

[1]INFN CNAF, Bologna, IT
[2]IFIN - "Horia Hulubei", Bucharest - Magurele, RO

E-mail: cristina.aiftimiei@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. 2016 was for the Cloud@CNAF IaaS (Infrastructure as a Service) based on OpenStack [1], a free and open-source cloud-computing software platform, a period of consolidation and improvement from the point of view of stability and reliability, in which the activities were focused on the support of the operation of the infrastructure, definition of monitoring checks, dashboards and notifications, integration in the general CNAF provisioning and monitoring system. During this period the number of supported users and uses cases has increased, and as a consequence, the infrastructure saw a grouth of the resources allocated. Work on the preparation of a parallel cloud infrastructure also started, infrastructure dedicated to the testing of OpenStack upgrade procedures, usually a complex and difficult task. This paper presents the activity carried out throughout the year in the directions mentioned and gives also a vision on the future evolution foreseen for the cloud infrastructure at CNAF.

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 and staff members
  - Provisioning of VM for CNAF staff members
  - Tutorial and courses
- National and international projects
  - Providing VMs for experiments hosted at CNAF, like CMS, ATLAS, EEE
  - testbeds for testing the services developed by projects like the OpenCityPlatform & INDIGO-DataCloud

The infrastructure made available is based on OpenStack, version Juno, with all the services deployed using a High-Availability (HA) setup or in a clustered manner (for ex. for the DBs used). During 2016 the infrastructure has been enhanced, by adding new resources, compute and network, and its operation has been improved and guaranteed by adding the monitoring part, improving the support, automating the maintenance activities.

2. IaaS enhancements

During 2016 a number of improvements where done at the infrastructure level, like:

- the number of the computing resources has been increased, growing from 7 hypervisors to 22, for a total of 352 CPU + 1.3TB RAM, so that the computing capacity offered to users, taking in consideration also the default overcommitting ratio for the CPU and RAM calculation, reached 5632 VCPU and 1,95TB VRAM, allowing the provisioning of a number of VM between 500 (if the RAM limit is considered) and 2000 (if the CPU limit is considered).
- the networking layer of the SDDS infrastructure, where Cloud@CNAF is hosted, has been upgraded, by replacing the core swith, a Black Diamond 8810 Extreme Network switch, with a new and high performance, low-latency switch 48 porte x 10Gb Brocade VDX 6740

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

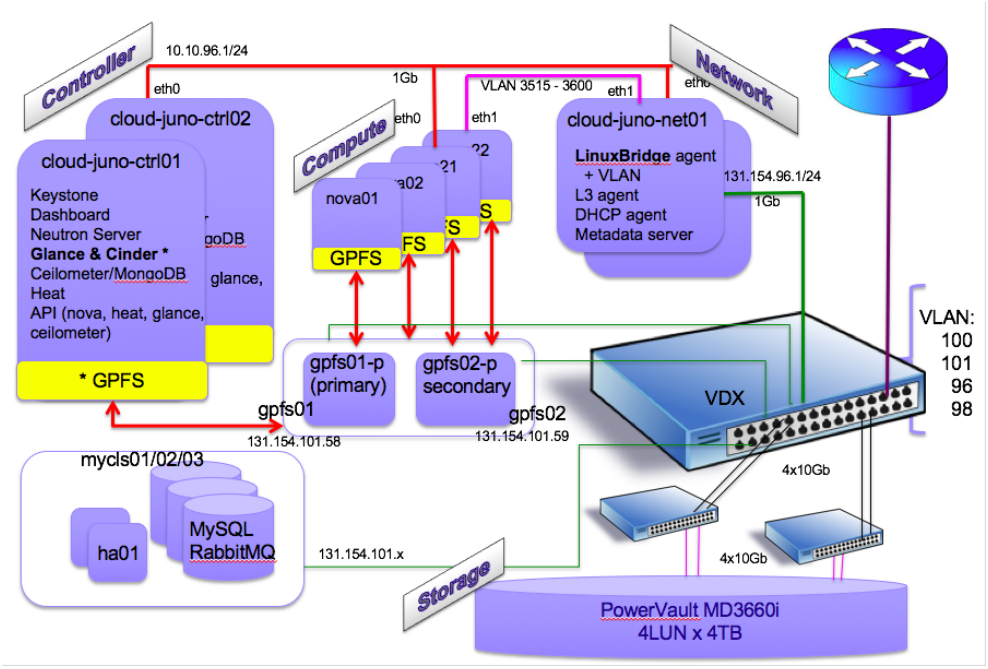


Figure 1. Cloud@CNAF v. Juno

After the upgrade of the networking part, at the beginning of the year our attention concentrated on the management of the infrastructure:

- integration into the CNAF Provisioning framework, Foreman [5], as can be seen in Fig 2
- integration into the CNAF Monitoring & Notification framework, Sensu [6], InfluxDB [8] and Grafana [7], as can be seen in Fig 3
- automation of operations activities - by installing and configuring Rundeck [2], as can be seen in Fig 4
- deployment of Rally [9], the OpenStack automatic testsuite, in order to test the status of the infrastructure after every major change

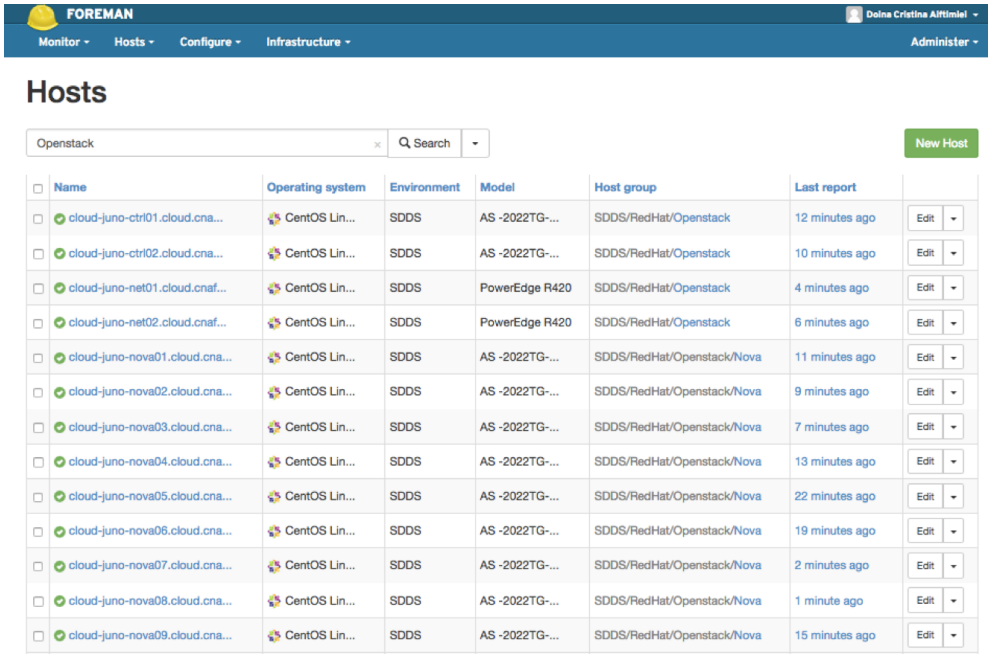


Figure 2. Foreman for Cloud@CNAF

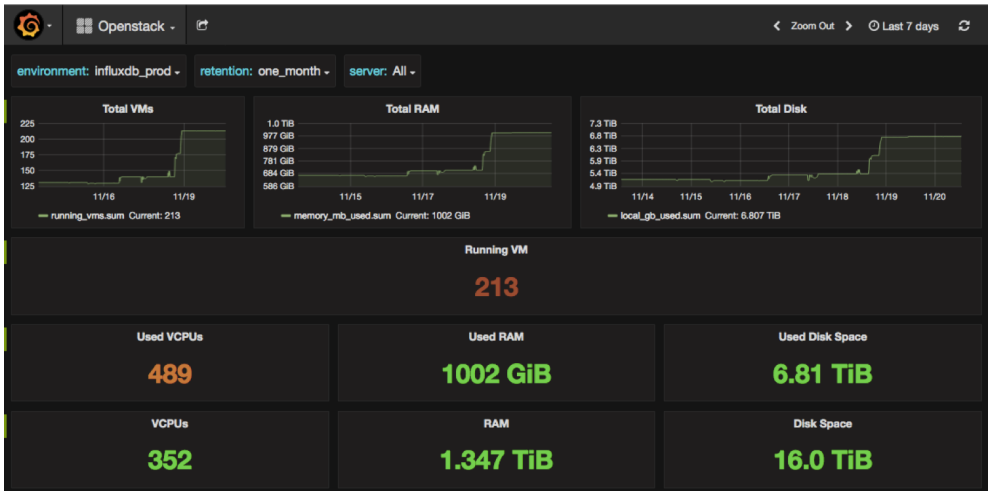


Figure 3. Monitoring for Cloud@CNAF

3. Users & Use cases

During last year the infrastructure saw also a growth on the number of users and use cases:

- the number of projects increased to 44 and the one of the users reached the number of 62, included the ones dedicated to the OpenStack services
- some of the project that used the cloud infrastructure are:
  - BioPhys - under the TTLab coordination, a collaboration started between biophysics scientists and computing experts finalized at optimizing the execution of the GATK - MuTect software pipeline used in genome sequencing analyses;

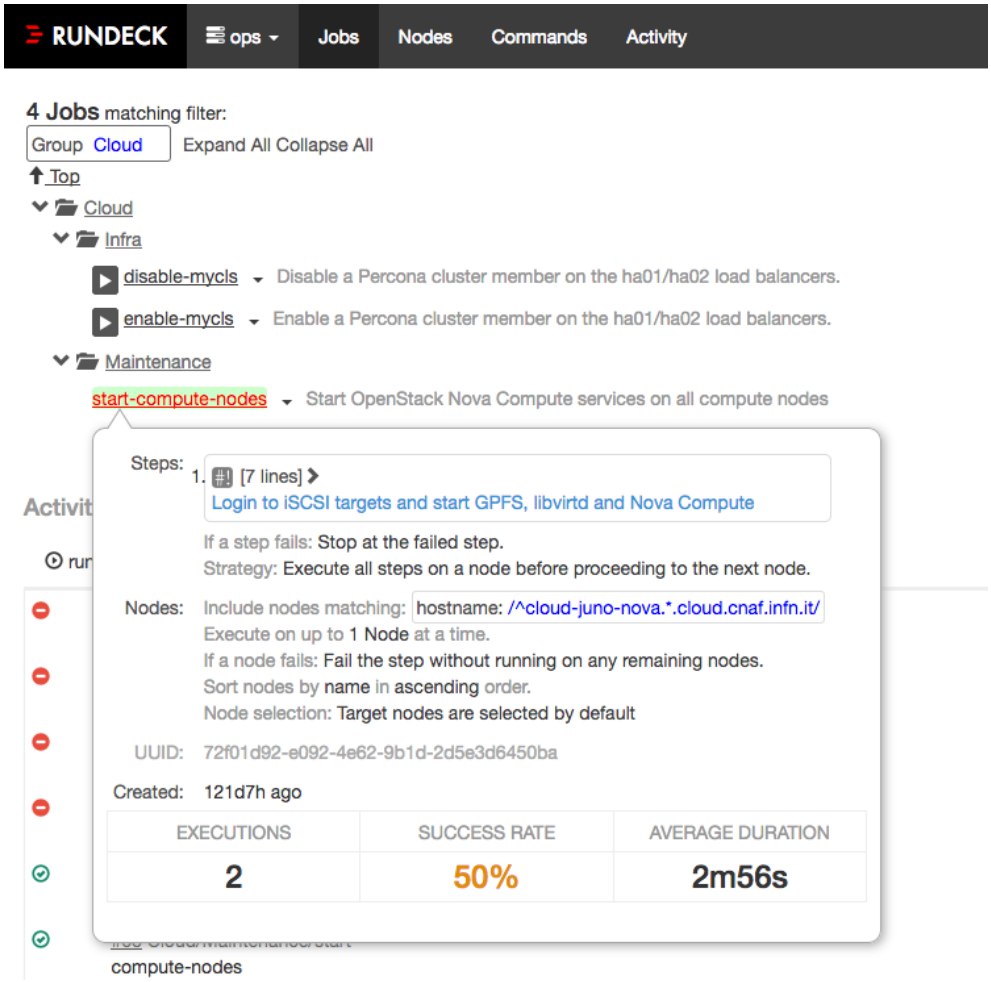


Figure 4. RunDeck for Cloud@CNAF

- USER Support - for the development of experiments dashboard and the hosting of the production instance of the dashboard, displayed on the monitor present on the CNAF hallway
- CMS - for Elastic Extension of Computing Centre Resources on External Clouds: - Extending a batch system working in a LAN (e.g. LSF) to external resources (CNAF Openstack) & Cloud Bursting of the Tier 3 Bologna (Tier3_as_a_Service)
- OpenCityPlatform - for Activities in the filed of BigData - build a platform based on Hadoop & Spark integrated with the OCP Software, Mesos and Marathon, and for Heat template development
- INDIGO-DataCloud - deployment of some of the solutions developed through the project like: Onedata - creating an INFN DataHub (activity described in another contribution); deployment of the PaaS layer as part of the project Preview testbed (activity described in another contribution), development and deployment of a Geographically distributed Batch System as a Service: leveraging the project approach exploiting HTCondor

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.

References

[1] OpenStack, *http://www.OpenStack.org/*
[2] Rundeck, *http://rundeck.org/*
[3] OpenCityPlatform (OCP), *http://opencityplatform.eu/*
[4] INDIGO - DataCloud, *https://www.indigo-datacloud.eu/*
[5] Foreman, *https://theforeman.org/*
[6] Sensu, *https://sensuapp.org/*
[7] Grafana, *https://grafana.com/*
[8] InfluxDB, *https://www.influxdata.com/*
[9] OpenStack Rally, *https://wiki.openstack.org/wiki/Rally*
[10] Apache Mesosphere (Mesos & Marathon), *https://www.digitalocean.com/community/tutorials/an-introduction-to-mesosphere*
[11] HTCondor, *https://research.cs.wisc.edu/htcondor/*

Software metrics thresholds: a mapping study

E. Ronchieri and M. Canaparo
INFN-CNAF, Bologna, IT
E-mail: elisabetta.ronchieri@cnaf.infn.it, marco.canaparo@cnaf.infn.it

Abstract. Many authors, over the years, have tackled the problem of defining suitable values or range of values for metrics' thresholds. Many different approaches have been proposed in literature so far: starting from the ones based on personal experience to those that exploited statistics; furthermore, in the last decades, machine learning methodologies seem to have become the most popular. In this report, we describe the last year activities on the identification of the influential software metrics threshold's papers by using the mapping study approach.

1. Background

Software metrics are reckoned as the most popular mean to measure code characteristics. IEEE defines them as *the quantitative measure of the degree to which a system, component or process possesses a given software attribute*. Although metrics have been successfully used for quantification purposes, they have marginally contributed to the decision-making process [1]. Thus, their use have involved the concept of threshold in order to determine if a certain measurement may be considered normal or anomalous. The determination of suitable threshold values is arduous as explained by Napaggan *et al.* [2]; in addition, threshold values can hardly be generalized from the programming language as stated by Zhang *et al.* [3].

In our research, we found a high number of papers that dealt with the topic of thresholds, nevertheless none of them seem capable of providing objective rules to apply them effectively. Therefore, we decided to identify the influential papers about software metrics thresholds following the evidence-based software engineering [4]. We adopted the mapping study approach, whose main purpose is to identify and categorize the considerable papers on a specific topic. In order to accomplish our objective, we focused our attention on the papers published from 1970 to 2015: to the best of our knowledge, there have been no surveys in the thresholds' field so far. We have used the SCOPUS tool to search for relevant peer-reviewed papers: SCOPUS is a general indexing system that includes publishers, such as IEEE, ACM, Elsevier, Wiley and Springer Lecture Notes publications. In addition to that, we employed the AlmaDL service of the University of Bologna to download the papers' full-text.

2. Research Methods and Results

The search process has been split in two phases ended in January 2016: the first one considered studies up to 2014; the second one only regarded year 2015. The reason why we decided to separate the papers published more recently is that in this case we omitted the filtration based on the number of citations. Due to the high number of papers found by the different searches (see Table 1) in SCOPUS, we decided to filter them according to various criteria: the number of citations, the title, common papers, abstract and text.

Table 1: Search queries

Data Set	Search String
papers 2015	(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 = 2015
papers 2015	(TITLE-ABS-KEY (metrics) AND TITLE-ABS-KEY (for) AND TITLE-ABS-KEY (object-oriented)) 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 = 2015
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
papers 1970-2014	(TITLE-ABS-KEY (metrics) AND TITLE-ABS-KEY (for) AND TITLE-ABS-KEY (object-oriented) AND TITLE-ABS-KEY (software)) 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

In the end, we obtained 33 papers published in 2015, 314 papers published in the range (1970-2014) and 15 papers in common (two of which are out of topic) [5]. On this sample, we started a classification of the papers, still on going, according to:

(i) the main topic: we identified 6 categories which describes the general purpose of the paper. Given the samples below we just explain 3 of them. A paper is defined as a "Development" if is about a specification of a new technique for calculating thresholds. Furthermore, an article can be classified as an "Assessment" i.e. it mainly evaluates an existing threshold or technique. A paper is reckoned as an application if it only applies an existing technique or calculated values of thresholds for a different.

(ii) whether the paper was empirical, theoretical or both. A paper is classified as Empirical if it assesses existing thresholds or techniques for calculating them. It is defined as theoretical if it discusses some issue about software engineering and may consider some theoretical aspects of software metrics thresholds.

(iii) the type of publication (e.g. proceedings, journal, books)

(iv) the software license of the projects analysed (e.g. open or commercial)

(v) the considered dataset of metrics (public or private)

(vi) the programming language of the analyzed projects

(vii) the type of metrics used (e.g. size, complexity, etc.)

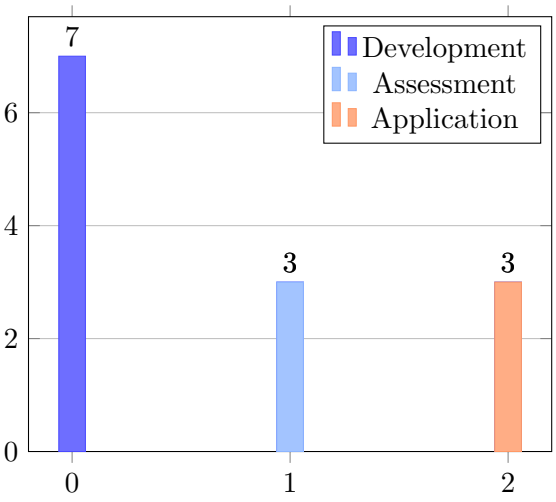
(viii) the type of the presented technique (e.g. statistical or artificial intelligence based)

Table 2: Metrics Thresholds

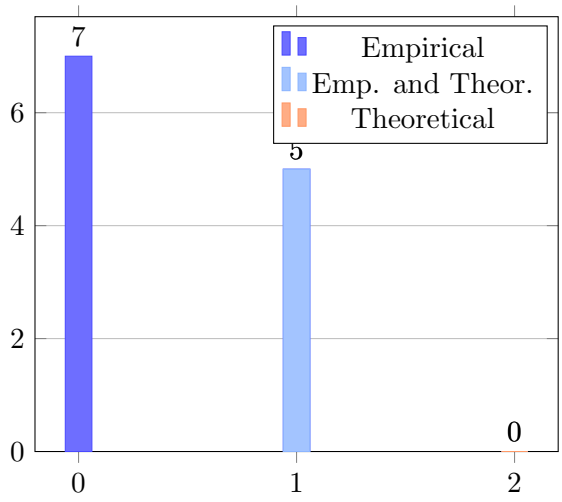
Reference	Type of Metrics	Thresholds
[6]	Size, complexity	McCabe <= 6 low risk,]6,8] moderate risk,]8,15] high risk, > 15 very high risk
[7]	Statements	Statements = 113(class)
[8]	Size and metrics derived from C&K	WMC > 31, TCC < 0.33, WOC < 0.95
[9]	A subset of C&K metric suites	CBO = 9, RFC = 40, WMC = 20
[10]	C&K, LK	CBO = 13, RFC = 44, WMC = 24, CTM = 33, NOC = 9 to identify low, medium and high risk

Table 2 and Figure 1 show some results of common publications between the two sets (see Table 1 from 1970 to 2015. The former is focused on the threshold values according to the type of metrics. The latter is related to 1.(a) the main topics, 1.(b) the type of paper and

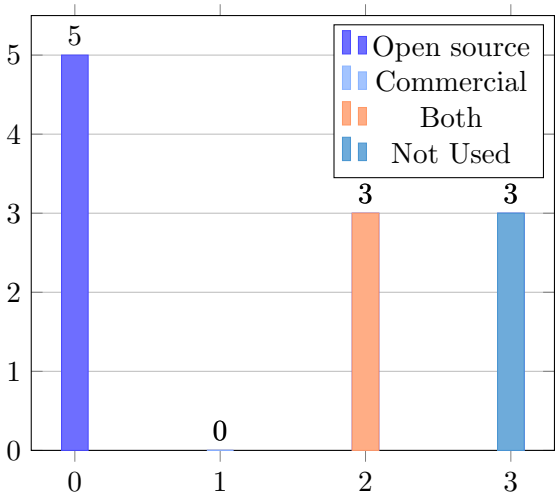
1.(c) if examined software is commercial or open source. Concerning Table 2, we specify the meaning of the metrics' acronyms used in threshold values columns: WMC (Weighted Method Count), TCC (Tight Class Cohesion), WOC (Weight of a Class), CBO (Class Between Object), RFC (Response for Class), CTM (Coupling Through Message passing), NOC (Number of Child classes), C&K (Chidamber & Kemerer), LK (Lorentz and Kidd).



(a) Categorization for main topic



(b) Categorization for empirical, theoretical or both



(c) Categorization for open source vs commercial

Figure 1: Results for papers

References

[1] Fenton N E and Neil M 2000 *ICSE - Future of SE Track* 357–370

[2] Nagappan N, Ball T and Zeller A 2006 *The 28th international conference on Software engineering (ICSE '06)* (New York, NY, USA,: ACM) pp 452–461

[3] Zhang F, Mockus A, Zou Y, Khomh F and Hassan A E 2013 *The 29th IEEE International Conference on Software Maintainability (ICSM'13)* (Eindhoven: IEEE) pp 350–359

[4] Kitchenham B A, Dyba T and Jorgensen M 2004 *The 26th International Conference on Software Engineering (ICSE'04)*

[5] Canaparo M and Ronchieri E 2016 *ICSOFT 2016 Proceedings* vol 1 pp 232–240 ISBN 978-989-758-194-6
[6] Alves T L, Ypma C and Visser J 2010 *The IEEE International Conference on Software Maintenance* pp 1–10
[7] Aman H, Mochiduki N, Yamada H and Noda M T 2005 *IEICE Transaction on Information and Systems* **E88-D**
[8] Mihancea P and Marinescu R 2005 *Ninth European Conference on Software Maintenance and Reengineering* (IEEE) pp 92 – 101
[9] Shatnawi R 2010 *IEEE Transactions on Software Engineering* **36** 216–225
[10] Shatnawi R, Li W, Swain J and Newman T 2009 *Journal of Software Maintenance and Evolution: Research and Practice* **22** 1–16

Continuous assessment of software characteristics: a step forward

E. Ronchieri, S. Antonelli, S. Longo, F. Giacomini
INFN-CNAF, Bologna, IT

E-mail: elisabetta.ronchieri@cnaf.infn.it, stefano.antonelli@cnaf.infn.it, stefano.longo@cnaf.infn.it, francesco.giacomini@cnaf.infn.it

Abstract. In this report, we describe the main activities performed in 2016 on the assessment of software characteristics. The main achievement has been the application of a solution in order to automate measurements of product metrics and static code analysis: we combine the existing INFN's GitLab installation with Docker assistance. The work is part of the INFN CCR Uncertainty Quantification project to assess characteristics of Geant4 software. Improvements and new objectives are also presented.

1. Introduction

The project about the assessment of software characteristics has been focused on the application of a solution that automates measurements of product metrics and static code analysis: the former can provide information about the characteristics of code, the latter is a critical technique to detect defects that are not visible to compilers.

Nowadays, there are several tools that measure metrics and perform analysis with excellent code editors. However, developers usually prefer traditional work environments discarding advanced code editors; furthermore, they like reducing any required effort to learn such tools. This is why developers support a solution where tools are set up once and used any time without strain.

In this report, we describe the applied solution that is based on the use of the existing INFN GitLab installation and Docker. The INFN GitLab installation (available at `https://baltig.infn.it`) is under the responsibility of the CNAF National ICT Infrastructures and Services department. Docker is the world's leading software container platform that is able to automate deployment of independent runtime environments inside a Linux hosts (see `https://www.docker.com/what-docker`).

The INFN GitLab installation by including GitLab-CI, where CI stands for Continuous Integration, eases the automation of some stages and their execution when the code changes. The various stages, such as build, test and deploy, compose what is known as the CI pipeline. Testing is one of these stages where some metrics tools and code analysers are executed.

2. Architecture summary

The used architecture is based on three entities (see Figure 1):

(i) users that push changes in the git repository;

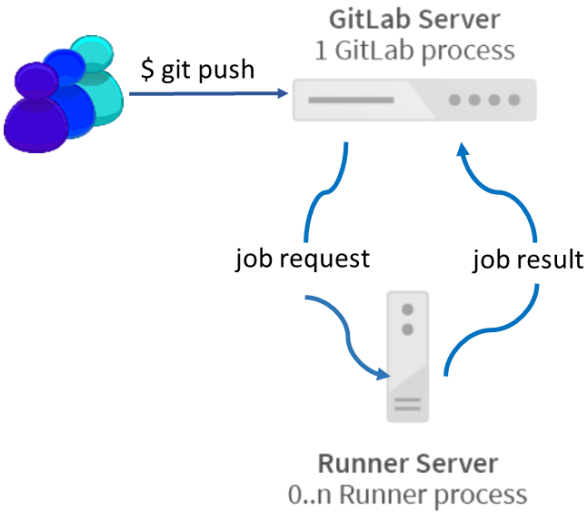


Figure 1: Architecture Summary

(ii) GitLab Server that stores all the repositories and supports container registry;

(iii) GitLab Runner that runs CI pipeline.

Users have to add a .gitlab-ci.yml file to the root directory of their repository and configure their GitLab project to use a GitLab Runner service. The latter runs the CI pipeline after each commit or push by using two type of executors, such as shell and docker, that can be used to run builds in different scenarios.

The shell executor executes builds locally to the machine where the Runner service is installed.

The docker executor uses the Docker service to run builds on user provided images. The Docker service runs on the Runner node, and each build runs in a separate and isolated container.

The .gitlab-ci.yml file contains some settings of the CI pipeline that details which operations the GitLab Runner has to perform. If their execution returns non-zero values, users obtain a green check-mark associated with the commit.

The Runner service is installed on a separate machine and associated to the GitLab Server by using a token-like authorization. The shell executor is used to build image and push it in the GitLab Container Registry. The docker executor is used to deploy a specific container according to the operations to be performed, such as metrics measures and static code analysis. On the Runner node it is requested to log in to the container registry before running docker commands.

At the end of the CI request the result is sent back to the main GitLab server, where any authorized user may check it by using a Web browser. In the project's page of the GitLab Web interface these users can access the Pipelines tab to check the result of the CI job, the Registry tab to check the storage of the pushed container image.

3. Use case

The described solution has been applied to automate the assessment of Geant4 software characteristics. The work is part of the INFN CCR Uncertainty Quantification project [1].

Geant4 is written in the C++ language and runs on different operating systems. In this work we have considered the latest version of CentOS as operating systems and the Geant4 10.2 release.

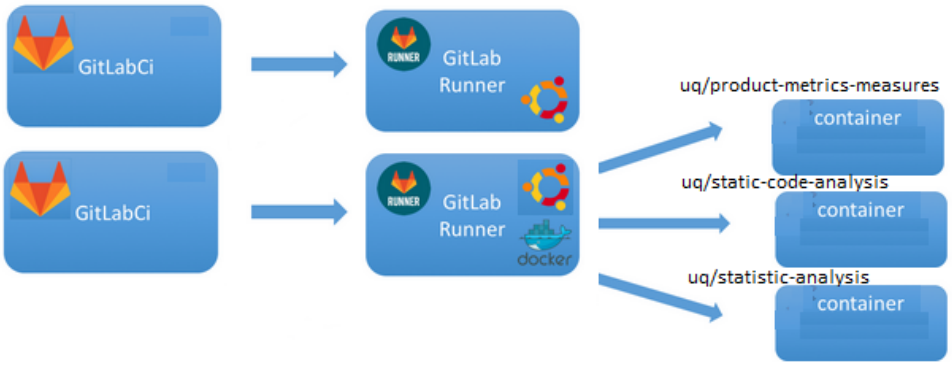


Figure 2: GitLab Ci

Figure 2 shows the use of a Runner node with a shell executor to create and upload a set of container images, whose names are set to the values of the CI_REGISTRY_IMAGE environment variable that follows the syntax: <CI_REGISTRY>:4567/<CI_PROJECT_NAMESPACE>/<CI_PROJECT_NAME> (see Table 1 that shows the values of the CI environment variables for each image).

Table 1: The CI environment variable values

Variable	Value
CI_REGISTRY	baltig.infn.it
CI_PROJECT_NAMESPACE	uq
CI_PROJECT_NAME	product-metrics-measures
	static-code-analysis
	statistic-analysis

Figure 2 also shows the use of a Runner node with a docker executor to run builds on different containers according to the different operations. At the moment the product-metrics-measures container includes the cloc and sloccount tools, while the static-code-analysis container makes available cppcheck (`http://oclint.org/`), flint++ `https://github.com/L2Program/FlintPlusPlus`, oclint (`http://oclint.org/`) and splint `http://www.splint.org/` tools. The statistic-analysis container includes the R tool to assess data collected by using those tools that are in the other described containers.

In the next step we are going to deploy results in the GitLab pages.

References

[1] Ronchieri E, Pia M G and Giacomini F 2015 *Journal of Physics: Conference Series* **664**

The INDIGO Identity and Access Management service

A. Ceccanti, E. Vianello, M. Caberletti
INFN-CNAF, Bologna, Italy

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

Abstract.
Contemporary distributed computing infrastructures (DCIs) are not easily and securely accessible by common users. Computing environments are typically hard to integrate due to interoperability problems resulting from the use of different authentication mechanisms, identity negotiation protocols and access control policies. Such limitations have a big impact on the user experience making it hard for user communities to port and run their scientific applications on resources aggregated from multiple providers in different organisational and national domains. INDIGO-DataCloud will provide the services and tools needed to enable a secure composition of resources from multiple providers in support of scientific applications. The core of the INDIGO Authentication and Authorization Infrastructure is the INDIGO Identity and Access Management Service (IAM), which has been designed and developed at CNAF by the middleware development group. In this contribution we give a description of the IAM service main functionalities and describe the problems it solves.

1. Introduction

The INDIGO AAI architecture primary objective is to fulfill the needs and requirements on authentication and authorization expressed by INDIGO user communities. The initial set of requirements of these user communities has been collected, organized and summarized [1, 2], and two generic use cases have emerged to drive the definition and development of the INDIGO platform.

The first generic user scenario is termed "Application portal as a service". In this scenario, computing applications are stored by the application developers in repositories as downloadable images (in the form of VMs or containers). Such images can be accessed by users via a portal, and require a back-end for execution; in the most common situation this is typically a batch queue.

The number of nodes available for computing should increase (scale out) and decrease (scale in), according to the workload. The system should also be able to do Cloud-bursting to external infrastructures when the workload demands it[1]. Furthermore, users should be able to access and analyse reference data, and also to provide their local data for the runs.

The second generic user scenario is described by scientific communities that have a coordinated set of data repositories and software services used to access, process and inspect the

[1] Cloud bursting is an application deployment model in which an application runs in a private cloud or data center and bursts into a public cloud when the demand for computing capacity spikes.

data. The processing is typically interactive, requiring access to a console deployed on the data premises [2].

2. INDIGO AAI requirements

To support the above user scenarios, the INDIGO AAI must satisfy the following requirements:

- Support for heterogeneous authentication mechanisms: The INDIGO AAI should not be bound to a single authentication technology, but should instead integrate and support federated authentication mechanisms like SAML [5] and OpenID Connect [6] and support X.509 [7] and username/password authentication;
- Identity harmonization and traceability: the INDIGO AAI should provide the ability to link multiple authentication credentials to a single INDIGO user profile, which provides a persistent and unique user identifier;
- Access to identity information: the INDIGO AAI must provide access to relying services to information regarding the user identity, presence and other attributes so that authorization and accounting can be implemented taking into account this information;
- Delegation: the INDIGO AAI must support constrained delegation, where a user can delegate part of his/her rights to an agent or service that acts on the user behalf. The service must be able to further delegate a subset of these rights to other services down the line, with the explicit or implicit approval of the user, to fulfill user requests; the INDIGO AAI must provide mechanisms and tools to safely define trusted delegation chains (i.e., which services can take part in a delegation chain, and which set of privileges can be delegated across the chain);
- Provisioning: the INDIGO AAI must provide the ability to provision, manage and deprovision identity information to relying services, to enable, for instance, local or service-specific account management;
- Virtual Organization/Collaboration management, registration and enrollment: the INDIGO AAI must provide the ability to define a VO/collaboration to group together users from distinct institutions sharing a common research goal, giving tools to manage the organization internal structure and registration flows;
- Integration and token translation: the INDIGO AAI must provide the ability to integrate with services that cannot be modified to directly support INDIGO AAI, for instance providing token translation functionality.

3. Authentication and Identity

The INDIGO AAI needs to accommodate heterogeneous authentication mechanisms and integrate with SAML identity federations like eduGAIN [8] and social identity providers (e.g., Google [10] or GitHub [11]).

In order to reduce the complexity of AAI integration at relying services and have the ability to decorate identities as provided by the upstream authentication mechanisms (SAML, OpenID Connect, X.509) with additional attributes, we developed a service whose responsibility is to authenticate users and agents (via the supported authentication mechanisms) and expose the authentication information to relying services trough standard OpenID Connect interfaces. This service is the Login Service component of the INDIGO Identity and Access Management (IAM) service [12].

This centralization of authentication responsibility in a single service, depicted in Figure 1, is an emerging architectural pattern (the *IdP-SP-Proxy* pattern [9]), which provides several advantages:

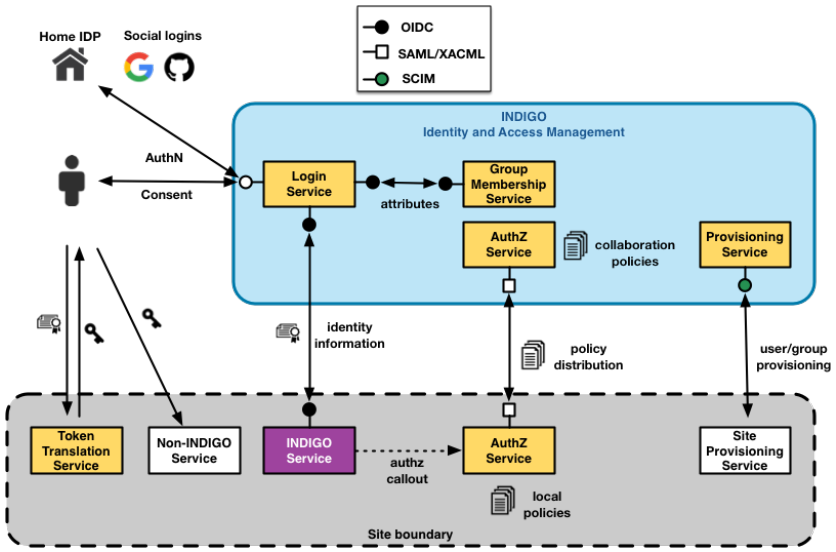


Figure 1. The INDIGO AAI overview. The INDIGO IAM Login Service is responsible for user and agent authentication, supporting several authentication mechanisms and exposing identity information through standard OpenID Connect interfaces. This approach simplifies registration in identity federations like eduGAIN and integration in relying services.

- A single point of control for authentication for all services: this approach simplifies auditing and control on enabled authentication mechanisms. On the other hand, from an operational point of view, the service has to be engineered in a way that allows it to scale horizontally to handle requests from services and not represent a bottleneck or single point of failure.
- Simplified registration in identity federations like eduGAIN: with this approach only one service, the Login Service, needs to be registered in the federation, not each individual service provider. This greatly simplifies support for identity federation at relying services.
- The ability to decorate the identity information obtained from upstream authentication mechanisms with additional attributes, like group membership attributes, roles, and, more importantly, a unique and non-reassignable persistent identifier that can be used to track down user activity in a way that is orthogonal to the authentication mechanism used in a given session.
- Natural support for guest users, i.e. users that do not have external credentials (i.e., are not part of any identity federation or do not want to use a personal social account), via the creation of local IAM Login Service username/password credentials.

3.1. OpenID Connect as the INDIGO identity layer

The INDIGO AAI identity layer is based on OpenID Connect. OpenID Connect is a Single Sign-On (SSO) protocol and identity layer recently standardized by the IETF, built on top of the OAuth 2.0 protocol framework [13], from which it inherits support for delegated authorization and offline access (that we discuss in more detail in section 5). Despite being a relatively new standard, OpenID connect is already widely adopted by many leading companies like Google, Microsoft, AOL, PayPal and LinkedIn.

OpenID Connect "allows clients of several types, including Web-based, mobile, and JavaScript clients, to request and receive information about authenticated sessions and end-users" in an interoperable and REST-like manner. Moreover, it provides a flexible trust model that

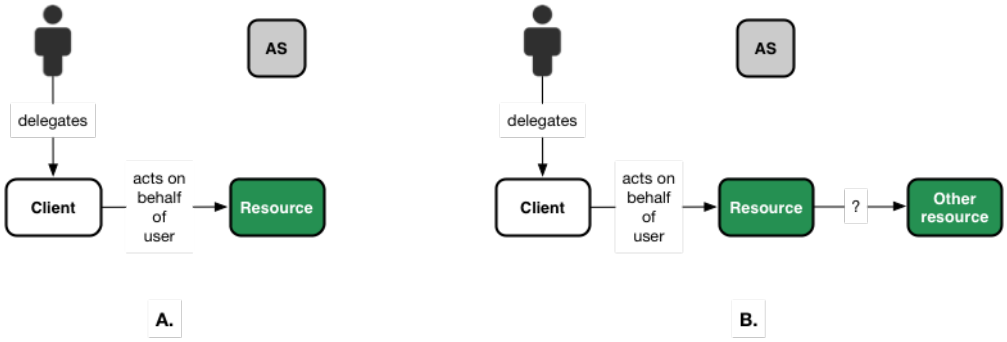


Figure 2. OAuth delegation vs OAuth chained delegation

accommodates the dynamic nature of our target infrastructure, as it standardizes protocols for dynamic registration [18] and identity provider discovery [17].

The main benefits of choosing OpenID Connect as the identity layer are:

- simplied integration in relying services, especially when compared to the SAML-based alternative;
- native support for delegated authorization on HTTP services (via OAuth, on which OpenID Connect is based);
- native support for long running computations, which is a common requirement for batch scientific computations;
- ability to accommodate heterogeneous authentication mechanisms (the OpenID Connect specification does not constrain how a user/agent should be authenticated);
- strong adoption of the technology in the industry, and availability of open source client and server libraries;
- support for mobile clients.

4. Authorization

Authorization, in the INDIGO AAI, follows the OAuth 2.0 delegated authorization model [13]: only agents presenting a valid and trusted OAuth access token are granted access to INDIGO services. Access tokens are obtained by client applications from the INDIGO IAM service and provide access to identity information (e.g., group membership and other attributes) and other authorization information (e.g., OAuth scopes) (see Figure 1).

Fine-grained, powerful and distributed attribute-based authorization is implemented by integrating the Argus authorization service [4] with the INDIGO AAI identity layer. This provides policy distribution and centralized XACML policy management.

5. Delegation and offline access

OAuth 2.0 was designed to solve the problem of a delegated access to resources across services, mediated by an authorization server, as shown in Figure 2A.

In scientific computing there are scenarios where a service, in order to satisfy a client request, needs to access resources hosted by other downstream services on behalf of the user, as shown in Figure 2B. In these scenarios, such a service acts both as an OAuth client application and an OAuth resource server. In our token model, access tokens are bearer tokens, so the first service could simply use the access token received from the client to interact, on behalf of the user, with the downstream service. There are, however, situations in which just using the received access token against the downstream service is not possible, like for instance if the token audience was

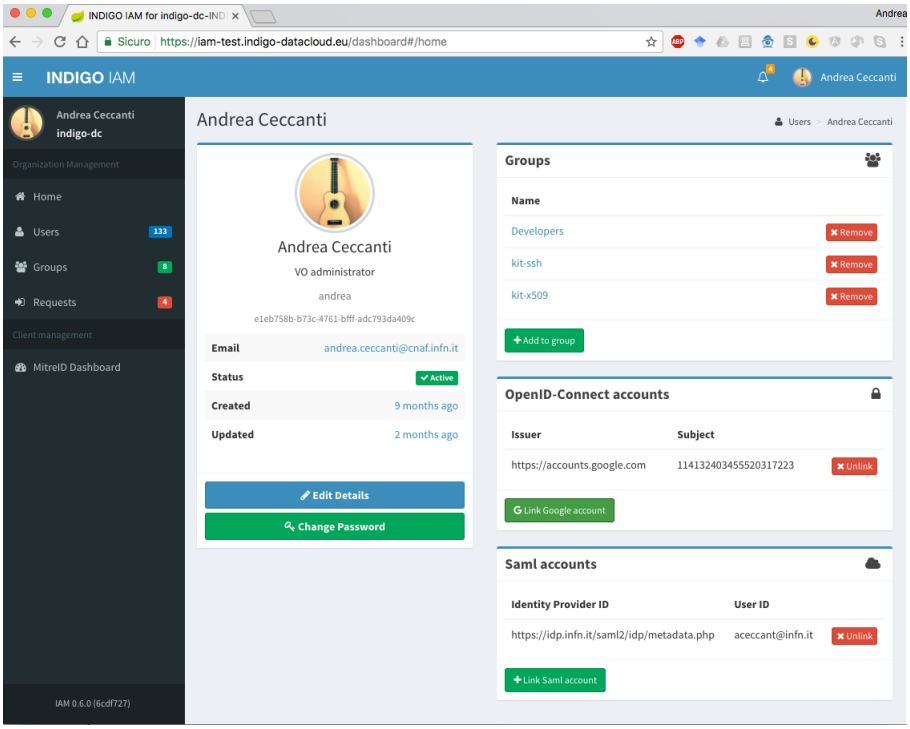


Figure 3. INDIGO IAM account management page

scoped to be valid only on the first resource server, or if the token does not grant scopes or privileges specific to the target service. Moreover, the resource server could need the ability to act on behalf of the user for an unbounded amount of time (e.g., to implement long-running computations), not limited by the validity of the received access token.

To support controlled chained delegation across services, in which a component can act both as a service and as a client for another downstream service, the INDIGO IAM implements the OAuth token exchange draft standard [15]. The token exchange is especially useful to implement controlled delegation of offline access rights across applications; i.e., the ability to execute tasks on behalf of a user while the user is not connected.

6. Provisioning

INDIGO IAM leverages version 2.0 of the standard System for Cross Domain Identity Management (SCIM) [16] to implement identity provisioning, deprovisioning and management.

The SCIM APIs provide a means to propagate identity and group information to relying services; for example, to implement dynamic account creation and other resource lifecycle management at various levels of the INDIGO infrastructure depending on events related to user identity status.

7. Identity harmonisation and account linking

Identity harmonisation is the process of providing consistent authorization and auditing across services based on account linkage information, so that, for instance, users' activity could be mapped to the same UNIX local account when they accesses a site through different services possibly relying on different authentication mechanisms (e.g., SSH keys, X.509 certificates, OpenID Connect tokens or SAML assertions).

The INDIGO IAM provides a powerful account linking mechanism (see Figure 3), where a

user can link several authentication credentials (OpenID Connect and SAML accounts, but also X.509 certificates and SSH keys) to a single user identity.

Access to this account linking information is then exposed to services via SCIM provisioning APIs.

8. Application integration

The INDIGO IAM has already been successfully integrated in almost all INDIGO services and in many external software components that INDIGO relies on (e.g., OpenStack [24]). The key for this successful integration is the adoption of the OpenID Connect standard. While OpenID Connect is widely supported, where that support is lacking, integration is still easy because it relies on technologies (e.g., JSON) that are already in use. This ensures the availability of libraries and SDKs in several programming languages. In the next section we describe in more detail how the Openstack integration works.

8.1. OpenStack integration

Starting from the native support for OpenID Connect implemented in OpenStack Keystone [25], a native OpenStack Mitaka deployment has been successfully integrated with IAM in order to support federated authentication and group-based authorization. The flow is as follows:

- a user that wants to access OpenStack IaaS service points his browser to the OpenStack dashboard URL. At this stage the user can choose to authenticate via INDIGO IAM credentials.
- The user is redirected to INDIGO IAM for authentication via a standard OpenID Connect authentication flow. Then, the user chooses how to authenticate (e.g., via home institution IdP or Google credentials) at the IAM. After successful authentication the user is redirected back to OpenStack;
- OpenStack Keystone is configured to grant access to specific tenants taking into account group membership information derived from the authentication token returned by the INDIGO IAM. For example, users in group CMS can access the CMS tenant.
- The same integration works also for the OpenStack command line client, through the OAuth password credential authentication flow [14] supported by the INDIGO IAM, that allows a user to directly authenticate with username and password.

9. Conclusions and future work

In this work we have introduced the main concepts behind the INDIGO Authentication and Authorization Infrastructure, the problems that it solves and its core service, the INDIGO Identity and Access Management (IAM) service.

The INDIGO AAI already represents a production-ready solution based on modern standards for securing access to infrastructure and services in support of scientific computing.

In 2017 we will focus on supporting the integration of the AAI (and in particular the INDIGO IAM) in all the target INDIGO use cases and on strengthening and improving the IAM code base and documentation.

References

[1] INDIGO-DataCloud D2.1: Initial Requirements From Research Communities `https://www.indigo-datacloud.eu/documents/initial-requirements-research-communities-d21`
[2] INDIGO-Datacloud: foundations and architectural description of a Platform as a Service oriented to scientific computing `https://arxiv.org/abs/1603.09536`
[3] The VOMS website `http://italiangrid.github.io/voms`
[4] The Argus authorization service website `http://argus-authz.github.io`

[5] The Security Assertion Markup Language (SAML) Wikipedia page https://en.wikipedia.org/wiki/SAML_2.0
[6] The OpenID Connect website http://openid.net/connect/
[7] The X.509 Wikipedia page https://en.wikipedia.org/wiki/X.509
[8] eduGAIN interfederation http://www.geant.org/Services/Trust_identity_and_security/eduGAIN
[9] First draft of the AARC Blueprint architecture https://aarc-project.eu/wp-content/uploads/2016/08/MJRA1.4-First-Draft-of-the-Blueprint-Architecture.pdf
[10] The Google Identity Platform https://developers.google.com/identity/
[11] The Github OAuth API reference https://developer.github.com/v3/oauth/
[12] The INDIGO IAM Github repository https://github.com/indigo-iam/iam
[13] The OAUTH 2.0 authorization framework https://tools.ietf.org/html/rfc6749
[14] The OAUTH 2.0 resource owner credential authentication flow https://tools.ietf.org/html/rfc6749#section-4.3
[15] The OAuth 2.0 Token Exchange draft standard https://tools.ietf.org/html/draft-ietf-oauth-token-exchange-07
[16] The System for Cross Domain Identity Management websitehttp://www.simplecloud.info/
[17] OpenID Connect discovery specification http://openid.net/specs/openid-connect-discovery-1_0.html
[18] OpenID Connect dynamic registration http://openid.net/specs/openid-connect-registration-1_0.html
[19] The eduGAIN website https://technical.edugain.org
[20] The WaTTS Github repository https://github.com/indigo-dc/tts
[21] Identity Harmonisation Service https://github.com/indigo-dc/identity-harmonization
[22] SCIM Event Notification https://tools.ietf.org/html/draft-hunt-scim-notify-00
[23] Amazon S3 https://aws.amazon.com/s3
[24] OpenStack website https://www.openstack.org/
[25] OpenStack Keystone https://docs.openstack.org/developer/keystone/

Partition Director

S. Taneja and S. Dal Pra
INFN-CNAF, Bologna, IT
E-mail: sonia.taneja@cnaf.infn.it

Abstract.
At CNAF the possibility to access the Tier-1 computing resources through an OpenStack based cloud service have been investigated as part of the INDIGO-DataCloud project, where a Partition Director component has been developed. A prototype have been set up for testing and development purposes, exploiting a dynamic partitioning mechanism, with the goal of avoiding a static splitting of the computing resources in the computing farm, while permitting a share friendly approach. The hosts in a dynamically partitioned farm may be moved to or from partitions, according to suitable policies for request and release of computing resources. Nodes being requested in a partition switch their role, becoming available to play a different one. In the cloud use case, hosts may switch from Worker Node in the Batch system farm to cloud Compute Node, made available to a Cloud Controller. This document describes the dynamic partitioning concept, its implementation and integration with a batch system based farm.

1. Partition Director

1.1. Introduction

A typical computing centre dedicated to HEP computing, such as a Grid site, is organized toward a quite specific usage pattern. Provisioning cloud resources from such an infrastructure is not straightforward, because of architectural differences:

- The existing computing resources are managed by a Batch System, dispatching *jobs* to the available *slots* and arbitrating access to them according to some sort of *fair share* policy, in order to enforce a convenient resource usage to competing job submitters. Worker Nodes are usually equipped with 3 to 4 GB RAM per core and one single physical network cable; online storage is designed to sustain an average I/O of 5 MB/sec per job.
- Conversely, a typical cloud infrastructure optimizes network traffic over three distinct channels, for *user activity*, *storage access* and *cluster management* and Computing Nodes are usually equipped with as many distinct network cables.

However, assuming to remain in the HEP usage scenario, the ability to provision cloud resources in a flexible and reversible manner is needed. Motivating use cases are:

- A Virtual Organization needs to start a "cloud computation campaign", dedicating a sub-quota of its pledges at the centre, then eventually reclaim back these resources to the usual batch mode.
- a VO wants to gradually migrate to cloud its computing resources.

- one user group needs resources for interactive usage. At CNAF this is currently handled by dedicating powerful hosts to work as "User Interface", bypassing the official pledge count.

The Partition Director is part of the INDIGO-DataCloud [1] suite and has been designed to be able to turn a given set of Worker Nodes in the batch cluster into Compute Nodes, enabling them to work for a Cloud Controller, and vice versa. This tool is intended for use by a site administrator, but it can also be driven by a user group, by triggering node conversion to add more resources toward a more *needing* partition. This document describes its functionalities and working principles.

1.2. Functionality

The Partition Director makes easy the management of a hybrid data center providing both batch system and cloud based services, with main focus at HEP applications.

Physical computing resources, in fact, can work as member of batch system cluster or as compute node on cloud resources.

Partition Director provides site administrators with an instrument to dynamically resize sub-quotas of a given user group among different infrastructures (HPC, Grid, local batch systems, cloud resources).

This ability can be directly used by site adinistrator, or can be driven by the user group itself, by tuning the amount of request at batch side and cloud side. At its basic level, the PD is responsible to:

- Switch the role of selected computing machines (Worker Nodes) from a batch cluster to a cloud one, enabling them as Compute Node (CN) and viceversa.
- Manage intermediate transition states, ensuring consistency.
- Adjust the shares at the batch side, to enforce an overall uniform quota to groups using both cloud and batch resources.

1.3. Working principles

The Partition Director works assuming that:

- A Batch System cluster is being used.
- A cloud instance, such as OpenStack is in place, having at least one Cloud Controller as privileged member of the batch cluster.
- On each node in the cluster, both batch and cloud services are active. However, they are enabled in a mutually exclusive manner.
- A Draining phase is needed at each node switching role.
- Computing resources (WN or CN) have read access to a shared file system.

1.4. Node transitions

1.4.1. From Batch to Cloud, [WN → CN] Assuming that new cloud resources are requested for a tenant, a number of WN are to be moved from the batch cluster and enable them as Compute Node (CN) to the Cloud Controller. When a WN is selected for *switching*, it usually have running jobs, hence a *draintime* is needed. During this *transition phase* the batch system stops dispatching jobs there and when the host is finally free, the WN is activated in the Cloud partition (using the nova-compute enable api call) as a hypervisor (Compute Node). The CN can then be assigned to the given tenant.

1.4.2. From Cloud to Batch, [CN → WN] Assuming that the cloud resources to a given tenant are not needed anymore, then Compute Nodes (CN) can be enabled back again as Worker Nodes in the batch cluster. Of course the selected CNs for *switch* could have Virtual Machines (VM) running. In this scenario also a *draintime* is foreseen.

When a CN is assigned to the batch partition, the cloud controller disables (nova-compute disable) node (Draining), so that no new VM's are instantiated there. The draining phase of a CN ends when no more VMs are hosted there. Unfortunately no obvious time limit exists, nor a direct way to know that a VM has done its tasks. Because of this, a reasonable timeout must be defined, after which the VMs are destroyed. To help preventing computational losses, the Machine Job Features Task Force [2] has defined a protocol that permits the VM to have access to a variety of information, in particular to *shutdowntime* which is a file on shared file system which stores the value of a TimeToLive (TTL) which the cloud user can consider as a timeout. During the Draining phase, running VM's have a grace period of time to finish their tasks. After the draintime finishes all the active VM's on the CN are destroyed and the node is enabled to the batch system and new batch jobs can be dispatched to it.

1.5. Implementation

1.5.1. Batch side components Specific to LSF batch system, it has two components: External Load Index Managers (ELIM) and External SUBmission (ESUB) modifier scripts.

- elim script: The elim script is launched by LSF on each node in the cluster and reports at regular times the value of a custom `dynp` flag, whose value indicates the partition to which the node belongs.
- esub script: The esub script is run at job submission time on the submitting host (UI or CE) and simply alter its submission parameters by adding the requirement of running on a node having `dynp!=2`. This mechanism is a variant of the one implemented to provision multi–core resources with LSF and more deeply detailed in [3].

1.5.2. Cloud side components Specific to Openstack cloud manager (Kilo or newer), it has two components:

- validation script : This tool primarily enforces consistency by evaluating the eligibility of each selected node for the indicated transition and put the nodes in the intermediate transition state. It is provided with a file containing the list of hostnames whose role is to be switched.
- director script: The director script implements the partioning logic and runs at a regular interval on a master node i.e Cloud controller or LSF Master and triggers role switching for nodes from Batch to Cloud or vice-versa. It checks for status changes of each node under transition and takes needed action accordingly. For example, when a node switching from Batch to Cloud has no more running jobs, it is enabled on cloud resources as hypervisor, and from then on, new VM can be instantiated on that node. The states of the nodes are stored as a json file in the shared file system.

1.5.3. Overall working of Partition Director The Partition Director is implemented as a finite state machine. Fig. 1 represents the algorithm implemented by the Partition Director. It also gives a pratical insight on how and when transition from one state to other take place.

- At $T = 0$, all nodes are $w_i \in B = \{w_1, \ldots, w_N\}$
- When k Compute Nodes are requested, they are moved from B to $B2CR$ (for validation) and then to $B2C = \{c_1, \ldots, c_k\}$ by the director.

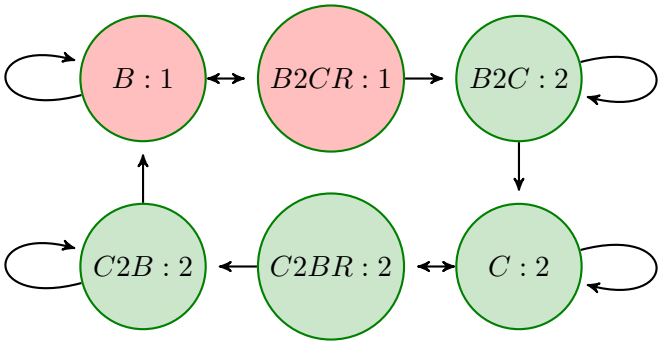


Fig. 1. The Status Transition Map

- When the drain finishes, it is moved from $B2C$ to C and becomes available as a Compute Node.
- When a Compute Node $c_i \in C$ must work again as a WN, it is moved from C to $C2BR$ (for validation)and then to $C2B$ and begins a drain time. The duration can be specified through the `shutdowntime` parameter from the machinejob features.
- When a Compute Node $c_i \in C2B$ expires its `shutdowntime`, Existing VMs are destroyed and the node moves to B as WN.
- The elim script on each node w_i updates its dynp status:

$$dynp(w_i) = \begin{cases} 1 & if\ c_i \in B \cup B2CR \\ 2 & if\ c_i \in C \cup C2BR \cup C2B \cup B2C \end{cases}$$

References

[1] Salomoni D, Campos I, Gaido L, Donvito G, Antonacci M, Fuhrman P, Marco J, Lopez-Garcia A, Orviz P, Blanquer I *et al.* 2016 *arXiv preprint arXiv:1603.09536*
[2] Alef M, Cass T, Keijser J, McNab A, Roiser S, Schwickerath U and Sfiligoi I 2016 URL `https://twiki.cern.ch/twiki/bin/view/LCG/MachineJobFeatures`
[3] Dal Pra S 2015 *Journal of Physics: Conference Series* vol 664 (IOP Publishing) p 052008

Middleware support, maintenance and development

A. Ceccanti, E. Vianello, M. Caberletti, F. Giacomini
INFN-CNAF, Bologna, IT
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 2016 by the CNAF middleware development team.

1. Introduction

The CNAF middleware development team has focused, in 2016, 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

The work on our CI infrastructure in 2016 focused on:

- Developing an integrated, dockerized packaging system that could be used for producing packages for all components [8];
- Porting our CI infrastructure to Kubernetes. This work is described in a separate contribution in this annual report;
- Integrating our infrastructure with CNAF Monitoring system [30]: on every node of the infrastructure a Sensu agent is installed, to collecting performance metrics and check the health of the node and the other services.

3. VOMS

During 2016, 48 new issues were opened in the VOMS issue tracker [6] to track maintenance, development and release activities. In the same period, 52 issues were resolved.

It has been a busy year for VOMS. Among the middleware products developed and maintained at CNAF, VOMS has been the one that caused more work, leading to 4 new releases for VOMS Admin [10, 11, 12, 13], 2 releases for the VOMS server [14, 15], and 1 release for the VOMS Java APIs, VOMS clients and VOMS MySQL plugin components [18, 17, 16].

Maintenance and development work focused on:

- Various improvements and bug fixes for VOMS admin to:
  - Improve integration with the CERN HR database;
  - Address problems in the handling of Sign AUP notifications;
  - Provide a functional audit log of all administrative operations performed on VOMS Admin;
  - Provide full text search functionality on the VOMS admin audit log;
  - Implement support for VO memberships that do not expire;
  - Persist notification in the VOMS database in order to support replicated, multi-master VOMS admin deployments;
- Fixing an important security vulnerability in the VOMS daemon which allowed impersonation due to improper RFC certificate chain validation [29];
- Porting the VOMS codebase to OpenSSL 1.1;
- Improved error reporting on VOMS daemon, native APIs and MySQL plugin component;

4. StoRM

During 2016, 16 new issues were opened in the StoRM issue tracker [19] to track maintenance, development and release activities. In the same period, 15 issues were resolved [20].

The main highlights for StoRM were:

- The releases of StoRM v1.11.11 [21] providing:
  - fixes for several issues found in production and during development;
  - a mechanism to gather metrics about synchronous operations load and performance;
  - changes to the internal garbage collector system which now adapts himself to better keep the request database size under control in high-load scenarios;
- Evolution of build and packaging with Docker [22].

It was a stable year, with no accidents or vulnerabilities detected, in which it was possible to focus on the optimization of our continuous integration infrastructure.

5. Argus

During 2016, 14 new issues were opened in the Argus project to track maintenance, development and release activities. In the same period 7 issues were resolved.

The main activities were:

- Release of Argus 1.7.0 on UMD-4 [27]: this is the first middleware product maintained by INFN which has been ported and released on RHEL7 based operating systems;
- Build a new packaging system, based on Docker containers, to build and provide RPMs for the RHEL6 and RHEL7 platforms [26];
- Improve the pool accounts mapping system on the PEP server, optimizing the access to the Gridmap directory, specially when an NFS server is used as storage backend for pool accounts;

- Migration and review of Argus official documentation from the legacy CERN wiki to a new website, hosted by ReadTheDocs [28];
- Build of a Docker-based deployment test environment, for development, testing and continuous integration purposes.

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 simplify the services management;
- Providing an alternative interface for QoS management on files hosted on a StoRM storage area, through a plugin for the INDIGO CDMI server [23];
- 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] VOMS on INFN JIRA https://issues.infn.it/jira/browse/VOMS
[7] Argus on INFN JIRA https://issues.infn.it/jira/browse/ARGUS
[8] The pkg.base system https://github.com/italiangrid/pkg.base
[9] The VOMS clients testsuite https://github.com/italiangrid/voms-testsuite
[10] VOMS Admin v. 3.4.1 http://italiangrid.github.io/voms/release-notes/voms-admin-server/3.4.1/
[11] VOMS Admin v. 3.4.2 http://italiangrid.github.io/voms/release-notes/voms-admin-server/3.4.2/
[12] VOMS Admin v. 3.5.0 http://italiangrid.github.io/voms/release-notes/voms-admin-server/3.5.0/
[13] VOMS Admin v. 3.5.1 http://italiangrid.github.io/voms/release-notes/voms-admin-server/3.5.1/
[14] VOMS Server and native APIs v. 2.0.13http://italiangrid.github.io/voms/release-notes/voms-server/2.0.13/
[15] VOMS Server and native APIs v. 2.0.14http://italiangrid.github.io/voms/release-notes/voms-server/2.0.14/
[16] VOMS MySQL plugin v. 3.1.7http://italiangrid.github.io/voms/release-notes/voms-mysql-plugin/3.1.7/
[17] VOMS clients v. 3.0.7 http://italiangrid.github.io/voms/release-notes/voms-clients/3.0.7/
[18] VOMS API Java v. 3.0.6 http://italiangrid.github.io/voms/release-notes/voms-api-java/3.0.6/
[19] StoRM issues created in 2016 https://issues.infn.it/jira/issues/?filter=15310
[20] StoRM issues resolved in 2016 https://issues.infn.it/jira/issues/?filter=15311
[21] StoRM v. 1.11.11 http://italiangrid.github.io/storm/2016/05/10/storm-v1.11.11-released.html
[22] StoRM packaging code https://github.com/italiangrid/pkg.storm
[23] INDIGO Datacloud Definition of QoS classes, policies and protocols for storage https://docs.google.com/document/d/15MYURV-57iwiQVUlMo9POEUG2VVe0wFgfv35i25sw-E/edit
[24] Docker https://www.docker.com/
[25] Docker compose https://docs.docker.com/compose/
[26] Argus dockerized packaging code https://github.com/argus-authz/pkg.argus
[27] EGI UMD-4 Middleware release http://repository.egi.eu/category/umd_releases/distribution/umd-4
[28] Argus Documentation website http://argus-documentation.readthedocs.io/
[29] VOMS certificate validation vulnerability https://wiki.egi.eu/wiki/SVG:Advisory-SVG-2016-11495
[30] S Bovina, D Michelotto, G Misurelli, *CNAF Monitoring system*, Annual Report 2015, pp. 111-114

Building an elastic continuous integration and delivery infrastructure with OpenStack and Kubernetes

M. Caberletti, A. Ceccanti, E. Vianello
INFN-CNAF, Bologna, IT
E-mail: marco.caberletti@cnaf.infn.it, andrea.ceccanti@cnaf.infn.it, enrico.vianello@cnaf.infn.it

Abstract. The software development and delivery process needs an environment where building, testing, and releasing software can happen rapidly, frequently, and reliably. The deployment of a new software release should be automated and require little or no operational effort, while allowing at the same time easy rollback when things go wrong. In this contribution we describe how our experience in building a continuous integration, testing and delivery infrastructure (CITD) on top of Kubernetes, the leading open-source platform for the automated deployment, scaling and operation of application containers.

Introduction

Moving middleware services to containers has been one of the top priorities in the last years for the CNAF middleware development group. The key objective is to exploit the many advantages of containerized applications, i.e. development and production environment consistency, simplified configuration, deployment and scaling.

Kubernetes [1] is an open-source system for automating deployment, scaling, and management of containerized applications. It groups containers that make up an application into logical units for easy management and discovery.

Kubernetes builds upon 15 years of experience of running production workloads at Google, combined with best-of-breed ideas and practices from the community. It is a completely open-source project, with a very active community: on GitHub [3] it counts more than 1000 contributors, and a fast release cycle (a new release every three months).

Kubernetes provides many interesting features, such as automatic binpacking [4], self-healing, automatic scaling, rollout and rollback of containerized application as well as service discovery and load balancing.

The key feature is the focus on applications: users ship applications inside containers and describe their requirements. Kubernetes then does the work: schedules the workload on the best node, checks application health, ensures the desired number of replicas, restarts or migrates the containers in case of failures.

For these reasons, we decided to build a continuous integration, testing and delivery infrastructure (CITD) based on Kubernetes to understand how this could improve our software quality and release processes.

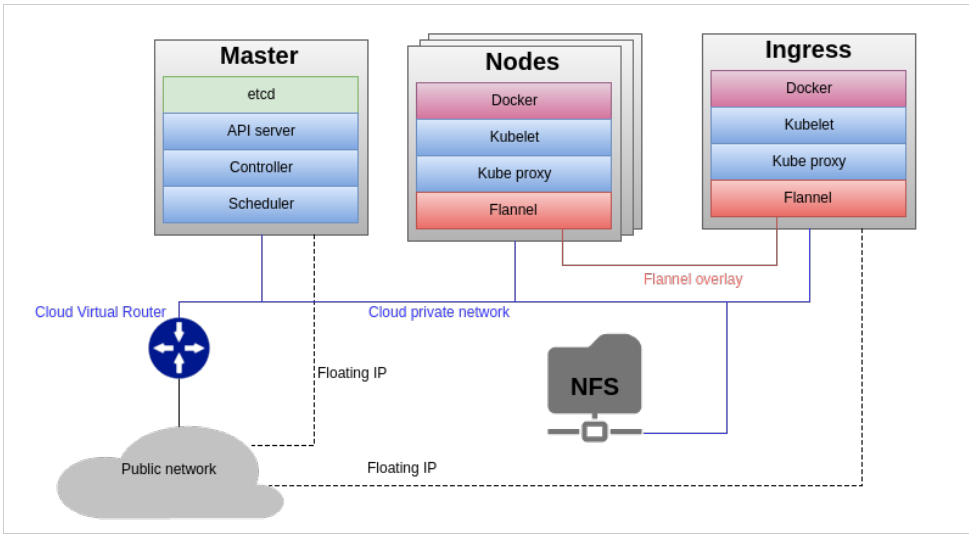


Figure 1. The Kubernetes infrastructure

The CNAF middleware development CITD infrastructure overview

Our Kubernetes pilot deployment is built on top of the infrastructure-as-a-service (IAAS) services provided by the Cloud@CNAF Openstack[5] infrastructure.

As figure 1 shows, the main components of the infrastructure are:

- the **Master node**: this node runs Etcd [7], a key-value data store used by Kubernetes to persist the cluster status, and the Kubernetes services that orchestrate the workloads execution on the cluster. Our pilot infrastructure has a single master node, which is enough to handle our current workloads.
- the **Worker node**: this node runs the Docker daemon and the Kubernetes agents that create/destroy containers and mount/unmount volumes. The Docker daemons are connected to each other through a Flannel [8] overlay network. Currently our infrastructure counts three worker nodes.
- the **Ingress node**: this is a "special" worker node, that runs a single Kubernetes service, called "ingress-controller", which acts as access controller and exposes selected applications to the public network. Our infrastructure has currently a single ingress node.
- the **NFS server**: this node is used to implement persistent volume storage via NFS and is used by docker containers to persist non-volatile data.

Typically, a Kubernetes installation deployed within a cloud provider does not need an ingress node to expose applications to the public Internet (cloud providers provide such functionality as a service).

However, in our pilot infrastructure, the ingress node is needed because the Cloud@CNAF infrastructure does not yet provide "LoadBalancer as a Service" functionality (this is going to change with the upgrade of Cloud@CNAF to OpenStack Mitaka).

Persistent data volumes are provided with a NFS (Network File System) server, shared between worker nodes inside the private network. We choose NFS volumes because, with this solution, volumes can be pre-populated with data, and that data can be "handed of" between pods. Kubernetes takes care of mounting/unmounting the shared NFS directory on the right worker node. Moreover, an NFS volume can be shared by multiple nodes at the same time. This functionality is needed to enable "rolling updates" of applications that do not incur any downtime and cannot be easily obtained by using other persistent volume approaches that do

not support volumes shared across nodes (such as, in our case, persistent volumes implemented via OpenStack Cinder).

Pilot continuous delivery use case: the Indigo IAM service
The first application that we have deployed on the Kubernetes infrastructure is the Indigo IAM Login Service and the Indigo IAM test client applications. Both applications are packaged and released as Docker containers.

The following requirements were identified for the IAM login service:

- it needs a MySQL compliant database to persist data such as collaboration members and token information;
- it requires read-only configuration files, such as SAML IdP metadata information;
- a single instance should be enough to handle most workloads, but replication should be possible if the need arises;
- its services are exposed to the public Internet only on TLS;
- service updates should not require any service downtime and should be invisible for client applications.

The IAM test client has similar requirements, with the exception that it does not need persistent storage. These requirements fit perfectly into a Kubernetes worload scenario. The Kubernetes deployment for the IAM login service basically defines:

- a *Pod* made of a single Docker container, with a read-only volume, mounted into the pod from a NFS shared directory, previously created and populated with the configurations files needed by the application;
- a *ReplicaSet* that ensures that at least one instance of a IAM login service pod is running and can answer client requests; this approach provides a rolling update strategy with zero-downtime;
- a *Service* that defines an high level endpoint to address the running pod instances.

The SQL database used by the application is managed outside Kubernetes: the connection information are injected into the IAM login service pod via environment variables, leveraging the Kubernetes secret management system [2].

To expose the IAM login service to the Internet, we defined an *Ingress* resource that configure the ingress controller as the TLS termination service and HTTP reverse proxy for the IAM login service pods.

A similar configuration is made for the IAM test client.

Zero-downtime rolling updates
When a new version of the IAM login service must be deployed in production, the only thing to do is change the Docker image name in the deployment definition and apply the new configuration. This operation is currently triggered by hand, but we plan to further automate it by including it in our continuous integration and delivery process, to really achieve automated continuous delivery.

When Kubernetes detects that a different configuration is submitted, it applies the rolling update strategy as follows:

(i) creates a new *ReplicaSet* with the new configuration;
(ii) this new *ReplicaSet* starts a pod running the new version of the software to be deployed in production;

(iii) when the pod with the new version is up and running, according to readiness and liveness probes defined in the configuration, Kubernetes configures the ingress controller to send requests to the new instance and scales to zero the old *ReplicaSet*, shutting down the old version of the service;

In this way, no downtime is observed during an update, because the old replica isn't removed until the new one is up and running. If the update fails, the old version keeps running.

Future work

The experience done with the management of the Indigo IAM login service deployment and updates has shown us that Kubernetes greatly simplifies services management and reduces the operational effort required to ensure service availability.

The good results obtained so far convinced us to plan the migration of other key services to Kubernetes, starting from our Jenkins's based continuous integration server, our Nexus repository manager and all the CI build nodes. This migration, started at the end of 2016, will allow us to reduce the CI infrastructure operational costs and will provide a completely elastic infrastructure to develop, test and run our services.

References

[1] Kubernetes official documentation `https://kubernetes.io`
[2] Kubernetes secret management `https://kubernetes.io/docs/concepts/configuration/secret/`
[3] Kubernetes GitHub project `https://github.com/kubernetes/kubernetes`
[4] The binpacking problem `https://en.wikipedia.org/wiki/Bin_packing_problem`
[5] D Michelotto, F Capannini, E Fattibene, D Salomoni and P Veronesi, *Cloud@CNAF*, Annual Report 2014, p. 136
[6] Docker website `https://www.docker.com/`
[7] Etcd GitHub project `https://github.com/coreos/etcd`
[8] Flannel GitHub project `https://github.com/coreos/flannel`

Development and tests of the Large Scale Event Builder for the LHCb upgrade

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

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

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

Abstract. The LHCb experiment will undergo a major upgrade during the second long shutdown (2019 - 2020), aiming to let LHCb collect an order of magnitude more data with respect to Run 1 and Run 2. The maximum readout rate of 1 MHz is the main limitation of the present LHCb trigger. The upgraded detector will readout at the LHC bunch crossing frequency of 40 MHz, using an entirely software based trigger. A new high-throughput PCIe Generation 3 based readout board, named PCIe40, has been designed on this purpose. The readout board will allow an efficient and cost-effective implementation of the DAQ system by means of high-speed PC networks. The network-based DAQ system reads data fragments, performs the event building, and transports events to the High-Level Trigger at an estimated aggregate rate of about 32 Tbit/s. In this contribution we present an Event Builder implementation based on the InfiniBand network technology. This software relies on the InfiniBand verbs, which offers a user space interface to employ the Remote Direct Memory Access capabilities provided by the InfiniBand network devices. We will present the performance of the software on a cluster connected with 100 Gb/s InfiniBand network.

1. Introduction

LHCb [1] is one of the four main experiments at the Large Hadron Collider (LHC). A major upgrade of the detector is foreseen during the second long shutdown of the LHC (2019-2020). The upgrade will concern both the detector and the Data Acquisition (DAQ) system. They will be reviewed in order to optimally exploit the expected collision rate of about 40 MHz. In fact one of the main limitations of the way the LHCb detector is currently operated is that the collision rate must be reduced to match the maximum readout rate of 1.1 MHz. The rate reduction is achieved by a hardware trigger operating within a fixed latency of a few microseconds. Due to its implementation the hardware trigger causes the largest inefficiencies in the entire trigger chain. One of the main objectives of the LHCb upgrade is to remove this bottleneck, thanks to the adoption of an approach characterized by a trigger-less readout and a fully software-driven event selection [2].

2. The upgraded DAQ design

The Event Builder (EB) is the enabling component of the upgraded DAQ system. It collects and reassembles the event fragments delivered by the sub-detector readout boards, called PCIe40 [3], at collision rate. Each PCIe40 is equipped with up to 24 optical links coming from the detector

and it is directly connected to a node of the EB through a Peripheral Component Interconnect Express (PCIe) slot. Consecutive event fragments transmitted from the front-end electronics are received and buffered by the PCIe40 and then copied into the EB node memory by means of Direct Memory Access (DMA) operations.

Given the foreseen nominal event size of 100 KB and the maximum event rate of 40 MHz, the aggregated bandwidth of the EB network can be estimated to be of the order of 32 Tb/s. Assuming an input rate, through the PCIe40 board, of 100 Gb/s per server, the size of the EB cluster can be estimated of the order of 500 nodes. The final foreseen readout system for the upgrade is shown in Figure 1, in which each of the estimated 500 nodes of the EB cluster will receive data from the front-end electronics and will exchange data with its peers at about 100 Gb/s full-duplex. Moreover, in case of no data filtering performed on an EB node, the built events will be sent out to an Event Filter Farm (EFF) again at about 100 Gb/s.

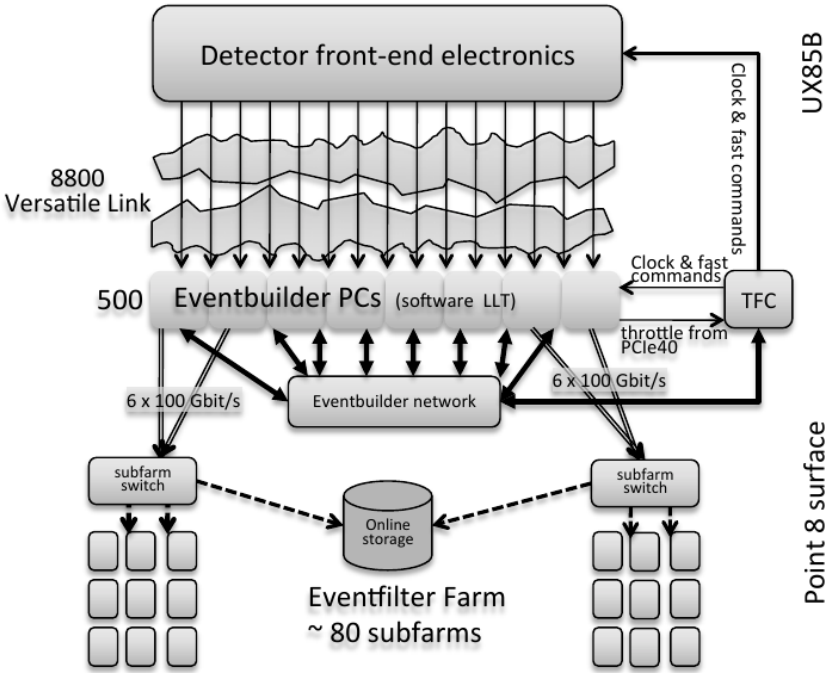


Figure 1. The architecture of the upgraded LHCb readout system.

The EB network can be effectively implemented by using commercial LAN technologies, such as 100G Ethernet [4], OmniPath [5] or InfiniBand [6]. In the following we first present the design choices adopted for the Large Scale Event Builder (LSEB) software [7], an EB software implementation designed for an InfiniBand interconnect infrastructure. Then we show some scalability tests aimed at validating those choices.

One of the key capabilities of InfiniBand is the support for Remote Direct Memory Access (RDMA), that is the ability to transfer data directly between applications over the network with no operating system involvement and while consuming negligible CPU resources on both sides (zero-copy transfers). This makes InfiniBand a high-speed, low-latency and low CPU-overhead interconnect technology.

3. The Event Builder implementation

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.

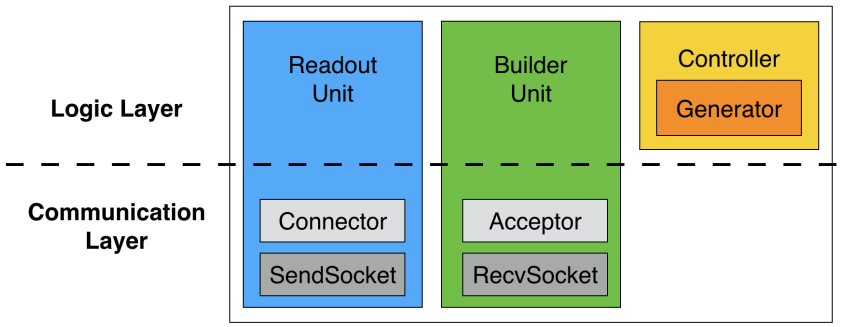


Figure 2. Schematic view of the main blocks of LSEB.

In order to keep the communication management separated from the event-building logic, LSEB can be logically 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 [8] library, which offers a user-space interface to access the 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.

3.1. Options for RDMA programming

RDMA programming offers many more options and is more complex than traditional sockets programming, as it requires the programmer to directly manipulate data structures defined by the network interface in order to directly control all aspects of RDMA message transmission. Therefore, the programmer must take decisions which may drastically affect performance. The following sections explain which implementation choices have been taken in order to develop the Communication Layer of LSEB.

3.1.1. Completion detection strategy. The basic principle of RDMA is the ability to access memory on a remote machine without involving the CPU, leaving it free to perform other tasks. The verbs implement this principle and they are designed to be asynchronous. Each transfer call is performed posting a work request on a dedicated First-In First-Out (FIFO) queue, named Work Queue (WQ); the call returns immediately instead of waiting for the completion of the requested operation. There are two ways for an application to know about the completion of a work request:

- *Busy polling.* The first completion detection strategy is to poll a dedicated FIFO queue, named Completion Queue (CQ), for a work completion.
- *Event notification.* The second strategy is to set up a dedicated channel, named Completion Channel, that allows an application to wait until a notification is signaled on it.

Repeatedly polling the CQ allows immediate reaction to completions at the cost of full CPU utilization, but requires no operating system intervention. On the other hand, the event notification approach is somewhat less efficient than the polling approach since it introduces the overhead of an interrupt handler and context switches between user and kernel modes [9].

Furthermore, the second strategy forces to adopt an event oriented design for the whole application in order to avoid a busy waiting call somewhere in the logic of the program.

The main aim of the Readout Units in the EB design is to send data to the Builder Units as soon as it is available. When all the send operations are started, the Readout Units cannot do anything but wait for a work completion and, with the high data rates expected, it is likely that the poll operation is almost always successful. Thus, LSEB implements the busy polling strategy, privileging the performance and keeping a linear and simple design.

3.1.2. Memory management. The DMA engine responsible for transferring data from main memory to the network device handles only physical addresses. Thus, the virtual addresses of the communication buffer have to be translated into physical ones; this process is called memory registration. Registering a Memory Region (MR) is a time-consuming process which requires several operations to be performed by the operating system and the driver [10]. In order to avoid registration and deregistration of memory for every transfer, it is preferable to use designated memory buffers, which are registered only once. Furthermore, registering physical contiguous memory can allow the low-level drivers to perform better since fewer memory address translations are required [11]. Thus, it is a good practice to register a large MR and to access a subset of it each time.

These considerations fit well with the design of the EB, in which a large and contiguous memory area of the Readout Units is written by the readout boards and subsets of it are sent remotely to the Builder Units. Therefore, LSEB allows the registration of a single MR, keeping all the memory contiguous, and avoids temporary registrations/deregistrations.

3.1.3. Work request type. Starting from all the data transfer options foreseen by the InfiniBand architecture [12], we identified two viable paradigms to implement the communication logic of the EB:

- *send/receive*. In this paradigm a receiver first posts a *receive* work request that describes a MR into which the HCA should place a single message. The sender then posts a *send* work request which refers to a MR containing the message to be sent. The HCAs transfer data directly from the sender's memory to the receiver's memory without any intermediate copy. Since both sides of the transfer are required to post work requests, this is called a "two-sided" transfer.
- *rdma_write/receive*. The second paradigm is a "one-sided" transfer in which a sender posts an *rdma_write* request that pushes a message directly into a MR that the receiving side previously communicated to the sender. The receiving side is completely passive during the transfer. When the sender needs to notify the receiver, it posts a *rdma_write_with_immediate* request to push a message directly into the receiving side's MR, as for *rdma_write*, but in this case the work request also includes 4 bytes of immediate (out-of-band) data that is delivered to the receiver on completion of the transfer. The receiving side needs to post a *receive* work request to catch these 4 bytes, and the work completion for the *receive* indicates the status and the amount of data written with the *rdma_write_with_immediate* operation.

In the EB logic each Builder Unit needs to know the address and the size of each fragment (or collection of fragments) sent by the Readout Units in order to build complete events. With the *send/receive* paradigm each Builder Unit implicitly obtains such information, being notified of the memory address and the transferred size of each transfer. With the *rdma_write/receive* paradigm the immediate data can be used to communicate the address, but this requires to post a *receive* work request by the Builder Unit for each data transfer. Moreover, using this second paradigm requires the Readout Units to be aware of the remote memory in order to write

only on those MRs no longer used by the Builder Units. Efficiency studies have shown that the performance of these two approaches are equivalent [13]. However, the use of the first paradigm has the advantage of avoiding the implementation of memory management and synchronization mechanisms. For these reasons LSEB adopts the *send/receive* paradigm to perform RDMA data transfers.

4. Performance tests

LSEB has been tested on a variety of clusters [14]. Here we describe the tests carried out on a 84-node High Performance Computing cluster in order to study the performance and the scalability of the software to validate the design choices described above.

4.1. Testbed details

We had the possibility to run LSEB on a cluster composed by 84 nodes and fully connected with InfiniBand EDR interconnect. Each node of the cluster is a "C6320 PowerEdge" server [15] and it contains two Intel Xeon Haswell E5-2697 v4 processors, each with 18 cores (36 hardware threads) and a clock of 2.3 GHz. The EDR (Enhanced Data Rate) standard is the highest data rate currently available with InfiniBand; it foresees 25 Gb/s of raw signaling data rate per link and in a classic four-link configuration it allows to reach a raw throughput of 100 Gb/s. Considering the overhead induced by the 64b/66b encoding scheme, the maximum theoretical bandwidth allowed by the InfiniBand EDR interconnect is of 96.97 Gb/s.

4.2. Measurements

Generically speaking, before testing an application on one or more nodes, it is required to execute standard benchmarks on those machines in order to establish how that application performs and how it can be improved. In case of applications that make use of RDMA, there is a set of micro benchmarks provided by the OFED package [16] that allow 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 data transfer operations for each different buffer size from 2^1 to 2^{22} bytes. Figure 3 shows the average bandwidth obtained running the ib_write_bw tool on two nodes of the cluster: the benchmark reaches about 95 Gb/s with buffer sizes greater than 32 KB (2^{15} bytes).

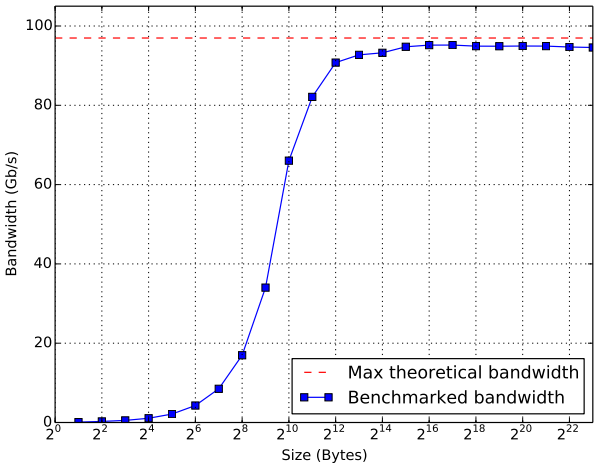


Figure 3. Bandwidth benchmarked with *ib_write_bw* on two nodes of the cluster.

After the benchmark with *ib_write_bw*, we started running LSEB on different scales. The scalability plot is shown in Figure 4, where the average bandwidth is plotted as a function of the number of nodes. In all the tests LSEB runs with the same fragment aggregation cardinality of 600 fragments; this implies that each data transfer has an average size of about 128 KB. This size is big enough to potentially saturate the bandwidth, as shown by the benchmark previously described. Due to availability issues we were able to exploit only up to 64 of the 84 nodes of the cluster. Starting from a 2-node setup, several tests were performed doubling each time the number of nodes, up to 64 nodes. Moreover, in order to better understand the bandwidth gap between 32 and 64 nodes, additional tests were performed with 36, 42, 48 and 56 nodes.

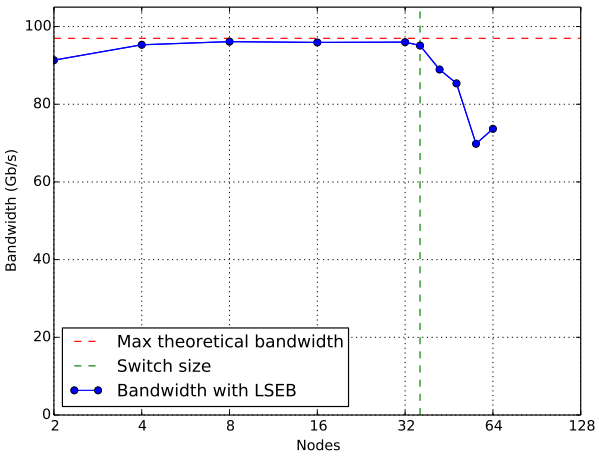


Figure 4. Bandwidth measured running LSEB on a different number of nodes.

The network topology of the cluster is a two-level non-blocking fat-tree network with 5 edge switches and 3 core switches. Both the edge and the core switches are Mellanox EDR InfiniBand SB77X0 systems [17] with 36 ports each. In a fat-tree network [18] the servers represent the leaves. In a two-level fat-tree network there are two layers of switches: servers are connected to the edge switches and the edge switches are interconnected through the core switches. When the up-links and down-links in an edge switch are in a 1:1 proportion the network is non-blocking. The presumed switch hierarchy present in the cluster is sketched in Figure 5. The plot in Figure 4 clearly shows an unexpected bandwidth drop while moving across the switches. This might be explained in case of a discontinuous node allocation on the cluster, which might create link contention. However, we do not expect such contention in a non-blocking fat-tree network. More studies are needed in order to understand the ultimate causes of this bottleneck and if the implementation of a software aware of the specific network topology can reduce the performance drop with an increasing number of nodes.

5. Conclusions

LHCb will under a major upgrade during the second long shutdown. Apart from detector upgrades, also the readout system will be redesigned in order to allow for a trigger-less data acquisition. We developed an evaluator in order to study the performance of the EB with high-throughput network technologies and we tested this software on a cluster connected with an InfiniBand EDR network, able to provide 100 Gb/s full-duplex bandwidth. The performed scalability tests show that the solution we implemented is promising and that the InfiniBand EDR interconnect could cope with the event-building requirements of the upgraded LHCb experiment. However, a complete and deep knowledge of the network topology is needed in

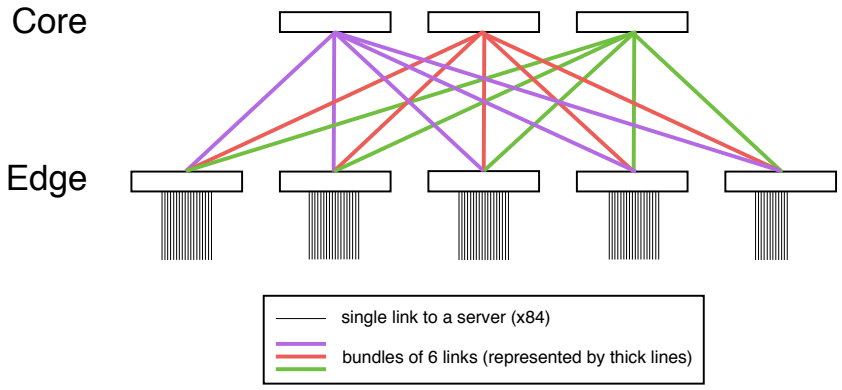


Figure 5. Presumed switch hierarchy of the cluster.

order to optimally exploit the bandwidth capacity with a number of nodes that require a switch hierarchy.

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] M. Bellato, G. Collazuol, I. D'Antone, P. Durante, D. Galli, B. Jost, I. Lax, G. Liu, U. Marconi, N. Neufeld, R. Schwemmer and V. Vagnoni, *A PCIe Gen3 based readout for the LHCb upgrade*, Journal of Physics: Conference Series (2104).
[4] 100 Gigabit Ethernet Technology Overview, `http://www.ethernetalliance.org/wp-content/uploads/2011/10/document_files_40G_100G_Tech_overview.pdf`.
[5] Intel Omni-Path Architecture, `http://www.intel.com/content/www/us/en/high-performance-computing-fabrics/omni-path-architecture-fabric-overview.html`.
[6] R. Buyya, T. Cortes, H. Jin, *An Introduction to the InfiniBand Architecture*, Wiley-IEEE Press (2002).
[7] M. Manzali, *lseb 2.0*, DOI 10.5281/zenodo.46935.
[8] RDMA Protocol Verbs Specification (Version 1.0), `http://www.rdmaconsortium.org/home/draft-hilland-iwarp-verbs-v1.0-RDMAC.pdf`.
[9] A. Cohen, *A Performance Analysis of 4X InfiniBand Data Transfer Operations*, Parallel and Distributed Processing Symposium (2003).
[10] F. Mietke, R. Rex, R. Baumgartl, T. Mehlan, T. Hoefler and W. Rehm, *Analysis of the Memory Registration Process in the Mellanox InfiniBand Software Stack*, Proceedings of Euro-Par 2006 Parallel Processing (2006).
[11] Tips and tricks to optimize your RDMA code, `http://www.rdmamojo.com/2013/06/08/tips-and-tricks-to-optimize-your-rdma-code/`.
[12] InfiniBand Architecture Specification (Release 1.3), March 2015, `https://cw.infinibandta.org/document/dl/7859`.
[13] P. MacArthur and R. Russell, *A Performance Study to Guide RDMA Programming Decisions*, Proceedings of the 2012 IEEE 14th International Conference on High Performance Computing and Communication (2012).
[14] A. Falabella, M. Manzali, F. Giacomini, U. Marconi, B. Voneki, N. Neufeld and S. Valat, *Large-scale DAQ tests for the LHCb upgrade*, 2016 IEEE-NPSS Real Time Conference (2016).
[15] Dell PowerEdge C6320 Datasheet, `http://i.dell.com/sites/doccontent/shared-content/data-sheets/en/Documents/Dell-PowerEdge-C6320-Spec-Sheet.pdf`.
[16] OpenFabrics Enterprise Distribution (OFED), `https://www.openfabrics.org/index.php/openfabrics-software.html`.
[17] Mellanox 1U EDR 100Gb/s InfiniBand Switch Systems Hardware User Manual Models: SB7700/SB7790, `http://www.mellanox.com/related-docs/user_manuals/1U_HW_UM_SB77X0.pdf`.
[18] C. Leiserson, *Fat-Trees: Universal Networks for Hardware-Efficient Supercomputing*, IEEE Transactions on Computers (1985).

Development and tests of TriDAS for the KM3NeT-Italy neutrino telescope

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

[3] INFN, Bologna, IT
[2] INFN-CNAF, Bologna, IT
[1] Università degli Studi di Ferrara, Ferrara, IT
[4] Università di Bologna, Bologna, IT

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

Abstract. KM3NeT is a deep-sea research infrastructure being constructed in the Mediterranean Sea. It will host a large Cherenkov neutrino telescope that will collect photons emitted along the path of the charged particles produced in neutrino interactions in the vicinity of the detector. The philosophy of the DAQ system of the detector foresees that all data are sent to shore after a proper sampling of the photomultiplier signals. No off-shore hardware trigger is implemented and a software selection of the data is performed with an on-line Trigger and Data Acquisition System (TriDAS), to reduce the large throughput due to the environmental light background. In this contribution the TriDAS developed for the KM3NeT-Italy detector is presented.

1. Introduction

The INFN's project KM3NeT-Italy, supported with Italian PON (National Operative Programs) fundings, is the inner core of the KM3NeT Cherenkov neutrino telescope for astrophysical searches using neutrinos as a probe (Fig. 1). The detector consists of 8 vertical structures, called towers, instrumented with a total number of 672 Optical Modules and its deployment is ongoing 3500 meters deep in the Ionian Sea in front of the south-east coast of Portopalo di Capo Passero, Sicily (Italy) [1][2]. A tower is a semi-rigid vertical structure composed by a sequence of 14 horizontal structures in marine grade aluminum named floors. Each tower is anchored to the seabed and kept vertical by an appropriate buoyancy on the top. Each floor, which is 8 m long, hosts six optical modules: two at either end and other two in the middle. Each optical module contains a 10-inch Photo-Multiplier Tube (PMT), a high-voltage supply circuit, a Front End Module (FEM) and an optical system for timing calibration.

The detection principle exploits 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 to a Trigger and Data Acquisition System (TriDAS) [3] software running at the shore station. The collected data stream from all the towers is largely affected by the optical background in the sea [4], 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.

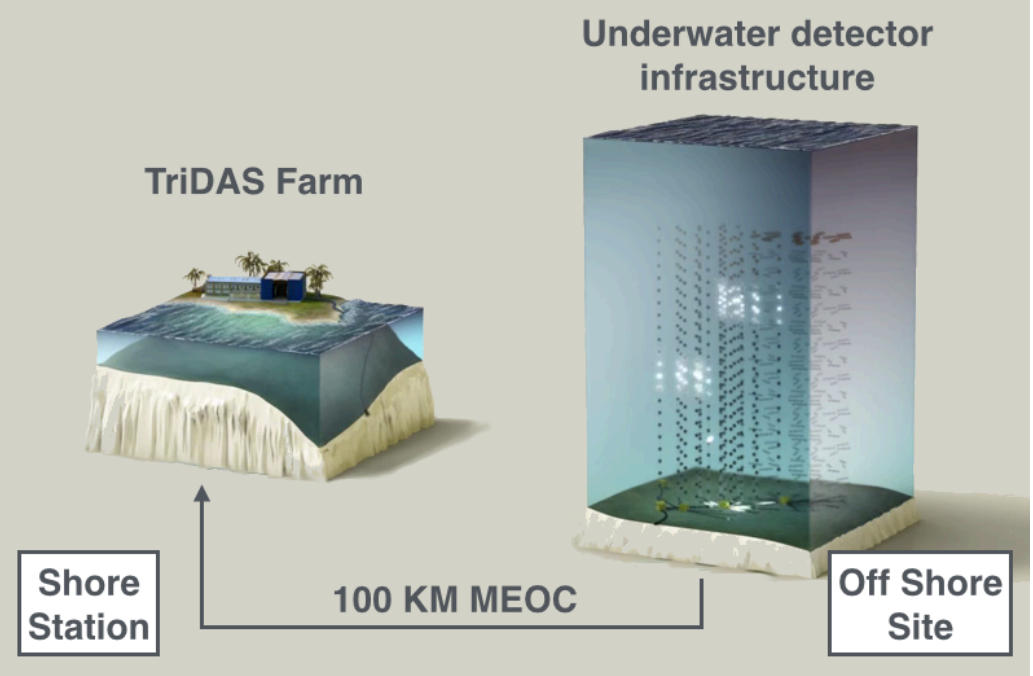


Figure 1. Reconstruction of the full detector linked to the shore station by a 100km Mechanical Electro-Optical Cable.

In the following sections we describe the final implementation of the acquisition system, the user and management interfaces and the testbed infrastructure. Finally, we present results of the scalability tests that demonstrate the system capabilities.

2. The TriDAS software

The TriDAS software has been developed to acquire and filter the data stream coming from the KM3NeT-Italy detector. Figure 2 shows the core acquisition components (TriDAS Core), the control components (WebServer and GUI) and their interactions with external services, presented in the next subsections. The C++ programming language is used for the development of the TriDAS Core programs, whereas the WebServer component is written in PHP and the GUI is based on the AngularJS framework. For the development of TriDAS, git [5] has been chosen as version control system, due to its strong support for non-linear development that makes it ideal for projects that require an extensive collaboration. All the code is open-source and hosted on BitBucket [6], a freely available web-based service able to host git repositories. In the following subsections each component of TriDAS is briefly described.

2.1. Floor Control Module Server (FCMServer)

The FCMServers represent the interface of TriDAS with the data from the off-shore detector. They perform the read-out of data through dedicated electronic boards called NaNet3 [7], which is a custom FPGA-based readout board hosted on a FCMServer through a Peripheral Component Interconnect Express (PCIe) slot. Each NaNet3 is point-to-point connected to up to 4 floors through optical fiber links. The communication is bidirectional and allows to propagate the on-shore GPS time to the floor electronics and the read-out of the data that are stored in the host memory. Each FCMServer merges the data coming from each connected floor into a single TCP/IP data-stream in the form of *hits*, that are variable length structures. Each hit

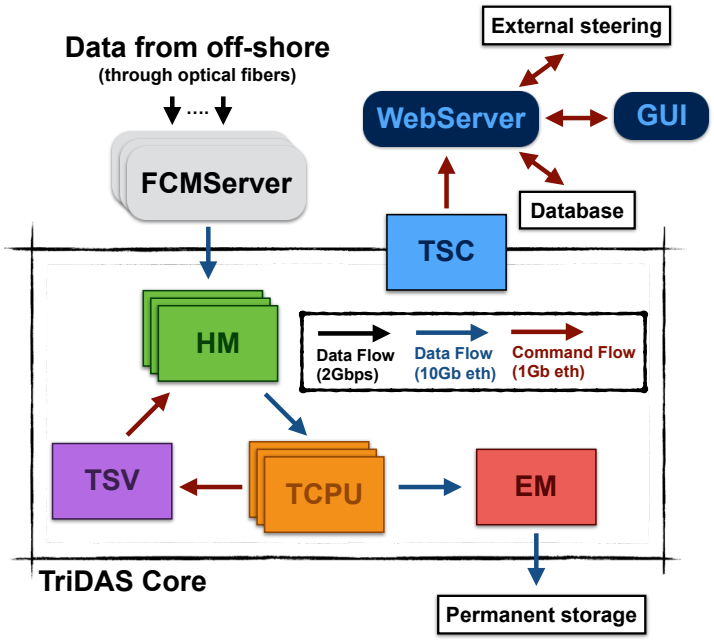


Figure 2. Scheme of TriDAS components and their interactions with external services.

is made of one or more *DataFrames*, another variable length structure, with fixed maximum size. Each DataFrame contains the OM identifier, a GPS absolute timestamp, the total electric charge collected by the PMT and an array of charge signal samples. If the number of samples exceeds the limit for a single DataFrame, a new one is created by the electronics and marked as a subsequent fragment of the first frame. Upon successful TCP connection, each FCMServer sends the data-stream to the connected HitManager.

2.2. HitManager (HM)

The HMs are the first aggregation stage for the incoming data-stream. Each HM is connected to a subset of FCMServers and it receives data from a portion of the detector called *Sector*. Hits are separated into arrays, one for each OM, and divided into temporally disjoint subsets called *SectorTimeSlices*. A *TimeSlice* is a temporal window of fixed duration, typically 200 ms. On completion of one of its SectorTimeSlices, each HM sends it to one TriggerCPU. The TriDAS SuperVisor assigns one TimeSlice to a precise TriggerCPU so that each HM sends its SectorTimeSlices, belonging to the same TimeSlice, to one TriggerCPU.

2.3. TriggerCPU (TCPU)

The TCPUs are responsible for the online analysis. Each TCPU receives from the HMs all the SectorTimeSlices that belong to the same TimeSlice, creating the so called *TelescopeTimeSlice* (TTS). This means that each TCPU has the snapshot of the whole detector during a specific TimeSlice. After its creation, each TTS is analyzed with a 2-steps trigger system:

- The Level 1 (L1) algorithms search for hit coincidences and charge excess along the whole TTS, identifying interesting portions of data, called *events*, made of the hits occurring in a time window 6 μs-wide centred in the trigger seed. The event window is increased if another trigger condition is satisfied within it.
- The Level 2 (L2) algorithms implement more complex conditions that operate on L1 events. If a L2 is satisfied, the event is marked to be saved on permanent storage.

The Trigger CPU implements a parallel processing of the TTS by means of a fixed number of worker threads, defined by the run-setup file.

2.4. Event Manager (EM)
The EM is the software component of TriDAS dedicated to the storage of the triggered data. It collects the events from all the TCPUs and writes them on the local storage in the form of a Post Trigger (PT) file. The PT file contains also the full run-setup and the positions and calibration information of the OMs. In this way, each PT file is self consistent with the bulk of information needed for the first stage of offline reconstruction.

2.5. TriDAS SuperVisor (TSV)
The TSV supervises the data exchange between HM and TCPU, taking note of the processed TelescopeTimeSlices. When a TCPU is ready to handle new data, it sends a token to the TSV, that assigns to that TCPU a new TimeSlice among those not yet processed.

2.6. 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 simple hierarchical state machine with four states.

2.7. WebServer
The WebServer is the unique entry point for operating the TriDAS Core. This component provides a set of RESTful APIs which allows to communicate with the TSC. Therefore, it provides user authentication based on hierarchical configuration implemented via different privileged groups. Despite the TSC is a local-single-client program, the WebServer can be contacted from several different concurrent users at a time. The WebServer allows to control the DAQ only one user at a time via an escalation procedure that permits users to acquire the control of the TSC. Finally, the WebServer can communicate instantly feedback and alarms through the use of WebSockets, implementing a real-time feedback to the users.

2.8. GUI
The GUI is a web application that graphically represents information provided by the WebServer and acts as control interface for the user.

3. Preliminary tests

We set up a testbed at the INFN section of Bologna in order validate the TriDAS software in all of its parts. The testbed is called Bologna Common Infrastructure (BCI) and it is a scaled version of the official acquisition farm present at the shore station infrastructure of Portopalo di Capo Passero. The hardware technologies present at the BCI and its network topology are similar to those adopted for the official acquisition farm. Figure 3 presents a schematic view of the BCI and it introduces all the servers used in the tests. Hereby we report the legend explaining the connection typologies:

(A) (B) Private LANs for data transfer from FCMServers and simulators to HMs.

(C) Private LAN for data transfer from HMs to TCPUs.

(D) Private LAN for data transfer from TCPUs to EM.

(E) INFN-Internal Public LAN (This LAN grants the control services and access to the servers).

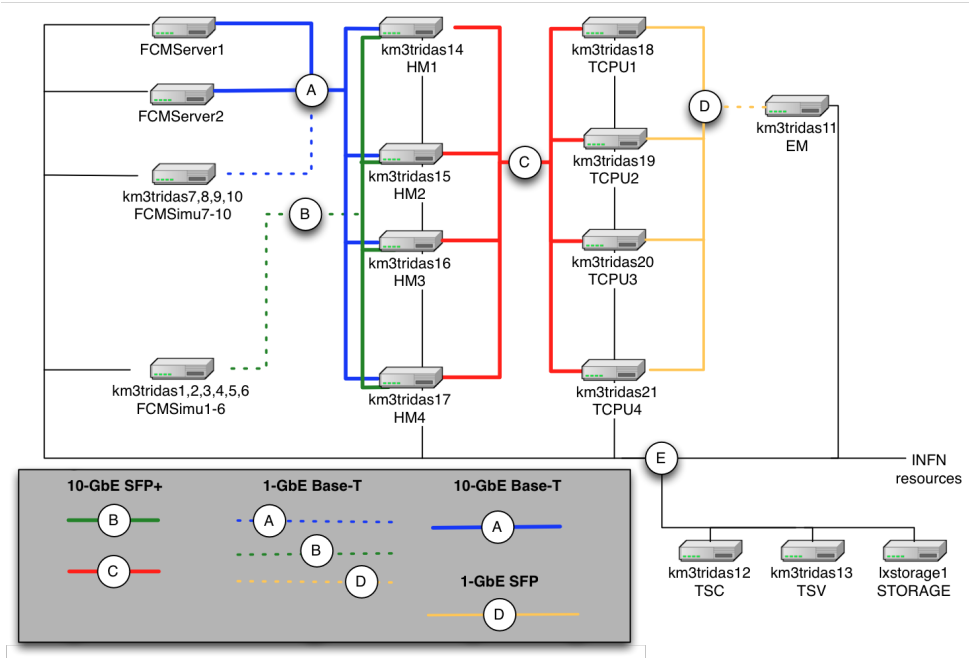


Figure 3. BCI network topology.

3.1. Scalability tests

In order to prove the scalability of TriDAS we performed different simulation tests at the 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. In order to reproduce the data throughput coming from the detector we developed a software called FCMSimu that is able to simulate the data stream coming from a single floor with a parametric rate of random hits. Preliminary studies performed at the Capo Passero deep-sea site have shown a background noise with an average hits rate of $\sim$ 55 kHz and less than 10% of peaks over $\sim$ 100 kHz [4]. We decided to perform the scalability tests simulating an heavier background in order to validate the software in a critical scenario. For this reason we ran the FCMsimu processes with a constant hit rate of $\sim$ 100 kHz for each optical module.

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 20 TTS in parallel, for a total of 80 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 processed in parallel by the system; in the described setup every TCPU node requires at most 0.2 s $\times$ 80 = 16 s for processing a TTS. Scalability tests were performed scaling from 1 to 4 towers and even in the worst case the mean time required to analyze a TTS is below the maximum time allowed by this setup, being 15.34 s against 16 s. Figure 4 shows the computation times measured divided by the number of TTS processed in parallel for each test performed.

4. Conclusions

An on-line Trigger and Data Acquisition System (TriDAS) has been design and implemented to handle the data coming from the KM3NeT-Italy underwater neutrino detector. Different

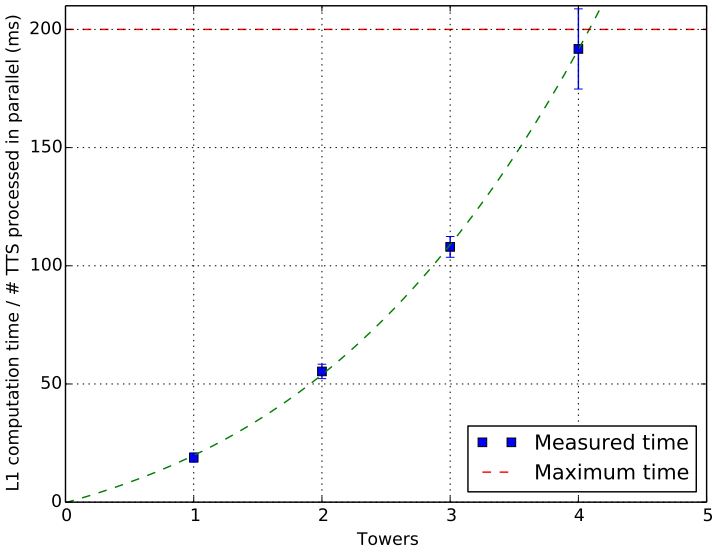


Figure 4. TTS computation time as function of the number of Towers

validation and scalability tests have been performed running the software on a dedicated testbed that is a scaled version of the official acquisition farm. Each software component has been successfully validated and the scalability tests have shown that TriDAS is stable and matches the requirements. More investigations will be carried out, thanks to the expansion of the testbed hosted at the BCI that will allow to test the system against a 8 towers detector.

References

[1] 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).

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

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

[4] 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).

[5] Git web site, http://www.git-scm.com/.

[6] TriDAS web site, https://bitbucket.org/chiarusi/tridas.

[7] R. Ammendola, A. Biagioni, O. Frezza, F. Lo Cicero, M. Martinelli, P.S. Paolucci, L. Pontisso, F. Simula, P. Vicini, F. Ameli, C.A. Nicolau, E. Pastorelli, F. Simeone, L. Tosoratto, and A. Lonardo, *NaNet3: The on-shore readout and slow-control board for the KM3NeT-Italia underwater neutrino telescope*, DOI: 10.1051/epjconf/201611605009, EPJ Web of Conferences 116, 05008 (2016).

A VOMS module for the NGINX web server

S. Paulon[1], F. Giacomini[2] and A. Ceccanti[2]

[1] University of Ferrara, Ferrara, IT
[2] INFN-CNAF, Bologna, IT

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

Abstract. This paper introduces the design and the implementation of a module that allows the NGINX web server to properly understand an X.509 proxy certificate presented by a client during authentication and extended with an attribute certificate issued by a VOMS server.

The main use case enabled by this module is a deployment of NGINX as a reverse proxy that handles all aspects of TLS communication on behalf of a business service running as a back-end.

1. Introduction

When developing a service that offers access to valuable or sensitive resources, security aspects such as authentication and authorization must be adequately addressed.

In the Grid world the established security model relies on X.509 certificates [1], both for server and user authentication. The model has proved to offer a strong base that, despite some usability issues, has served well for almost 20 years.

To enable the typical use cases of Grid systems and provide single sign-on functionality, client authentication is in fact performed presenting proxy certificates [2] which provide information about the user (or client application) identity and other signed information, such as the name of the scientific collaboration to which the user belongs. This information is presented in the form of VOMS [4] attribute certificates [3], and can be used for authorization purposes by Grid services (e.g., to allow submission of computational activities or grant access to data).

A web service wanting to perform authentication and authorization based on VOMS proxies must carry out a number of security-related actions before even starting its own business logic:

- offer an HTTPS endpoint [5]
- demand and validate certificate-based client authentication
- validate and extract the VOMS attributes

The goal of this work is to enable a deployment model of a service whereby all the above steps are isolated in a generic front-end server, which forwards a pre-processed request to a back-end server where the business logic is implemented. The reply from the back-end server is then relayed by the front-end server to the client through the already-opened secure channel. The model is shown in Figure 1. Note that if the back-end and front-end servers are in a trusted zone, the communication between the two can even happen over plain HTTP.

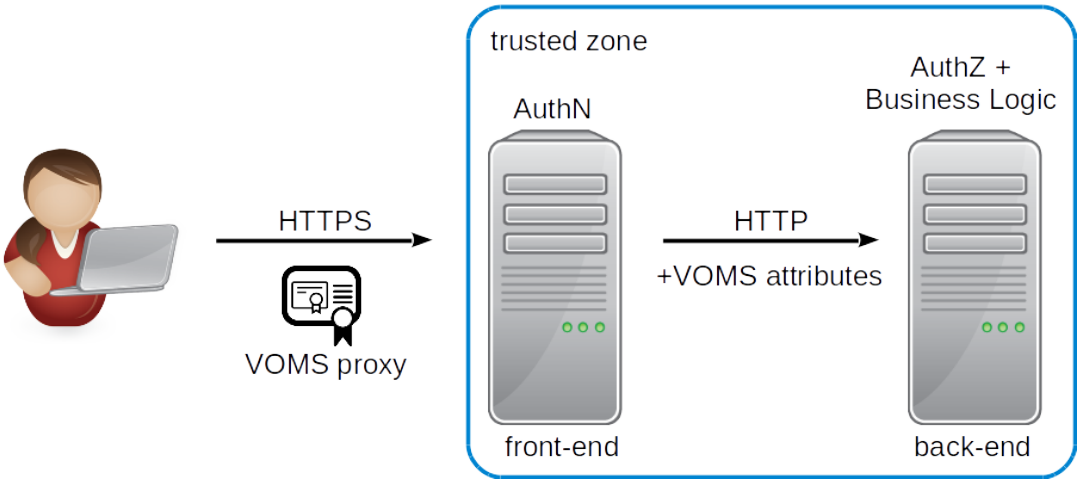


Figure 1. Deployment model enabled by the VOMS module for NGINX.

2. The NGINX web server

NGINX [6] is an HTTP and reverse proxy server, a mail proxy server and a generic TCP/UDP proxy server. Its design and implementation focus on scalability, efficiency, small resource footprint and portability on diverse operating systems. It is one of the most used web servers, especially for heavily-loaded sites.

NGINX has a modular architecture and comes with a number of ready-to-use modules. However an installation that requires a module not foreseen in the distribution needs to be recompiled because at runtime NGINX will load only modules that it knows about during compilation.

An important concept in NGINX is the *embedded variable*. An embedded variable is a variable that NGINX itself or one of its modules creates, associates with a value and makes available to others through a shared hash table. A variable is then accessible for example from the NGINX configuration file or from another module.

3. A VOMS module for NGINX

NGINX comes with the *ngx_http_ssl_module* module, which already takes care of SSL-protected HTTP communications, including the management of X.509 proxy certificates. The module provides a number of useful embedded variables, notably the *ssl_client_raw_cert* variable, containing the client certificate, in PEM format, for an established SSL connection. The SSL module however does not provide immediate access to X.509 extensions, one of which corresponds to the attribute certificate issued by VOMS.

The purpose of the VOMS module, called *ngx_http_voms_module*, is then to apply the VOMS API to the client certificate made available by the SSL module through the above-mentioned *ssl_client_raw_cert* variable, in order to access the VOMS attribute certificate and extract from it the VO membership information in the form of FQANs.

The VOMS module makes the FQAN information available to others in the form of two embedded variables:

voms_fqan The value is the list of all FQANs found in all the attribute certificates included in the client certificate.

voms_primary_fqan The value is the first FQAN of the previous list.

4. Using the NGINX VOMS module

Once the embedded variables *voms_fqan* and *voms_primary_fqan* are available, they can be used to configure an NGINX deployment such as the one envisioned in the introduction. Listing 1 shows some relevant excerpts taken from a possible configuration file for the front-end server.

Listing 1. Configuration file for the front-end server.

```
env OPENSSL_ALLOW_PROXY_CERTS=1;
load_module "/usr/lib/nginx/modules/ngx_http_voms_module.so";

http {

    upstream backend {
        server back-end:8080;
    }

    server {
        listen 443 ssl;
        ssl_verify_client on;

        location / {
            proxy_set_header Voms-Fqan         $voms_fqan;
            proxy_set_header Voms-Primary-Fqan $voms_primary_fqan;
            proxy_pass       http://backend;
        }
    }
}
```

The back-end server would find the FQAN information readily available in the headers of the incoming HTTP request.

Acknowledgments

This work was the subject of the stage project of one of the authors (Paulon) in collaboration with CNAF and in preparation of his thesis in Computer Science at the University of Ferrara [7].

References

[1] Cooper D, Santesson S, Farrell S, Boeyen S, Housley R and Polk W, Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile, RFC 5280, May 2008. http://www.rfc-editor.org/rfc/rfc5280.txt.

[2] Tuecke S, Welch V, Engert D, Pearlman L and M. Thompson, Internet X.509 Public Key Infrastructure (PKI) Proxy Certificate Profile, RFC 3820, June 2004, http://www.rfc-editor.org/info/rfc3820.

[3] Farrell S, Housley R and Turner S, An Internet Attribute Certificate Profile for Authorization, RFC 5755, January 2010.

[4] Alfieri R, Cecchini R, Ciaschini V, dell'Agnello L, Frohner Á, Gianoli A, Lõrentey K and Spataro F, VOMS, an Authorization System for Virtual Organizations, pp. 33–40. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004.

[5] Rescorla E, HTTP Over TLS, RFC 2818, May 2000. http://www.rfc-editor.org/rfc/rfc2818.txt.

[6] Nginx, https://www.nginx.com/.

[7] Paulon S, Progettazione e implementazione di un modulo per il server web NGINX dedicato alla comprensione di X.509 Attribute Certificates, https://gitlab.com/bigPaul/ngx_http_voms_module/.

A web application to analyse XRF scanning data

L. Cappelli[1], F. Giacomini[2], F. Taccetti[3], L. Castelli[3], L. dell'Agnello[2]

[1] University of Ferrara, Ferrara, IT
[2] INFN-CNAF, Bologna, IT
[3] INFN, Firenze, IT

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

Abstract. This article summarizes the design and the development of a prototype web application aimed at the visualization and analysis of images obtained from the application of XRF scanning techniques to works of art. The scanning is carried out using a spectrometer developed at the INFN-LABEC in Florence.

The goal of the project is to verify that replacing the current native application with one based on web technologies is indeed possible. This would make the application easily accessible to a wider user base.

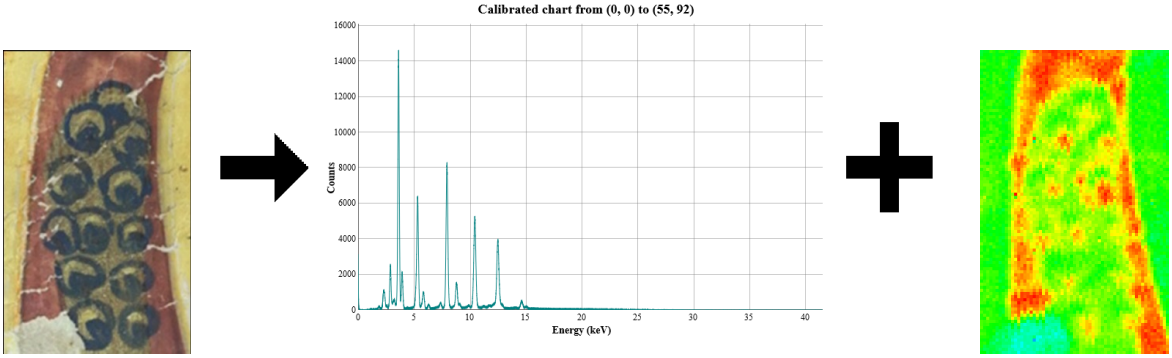
1. Introduction

The *X-Ray Fluorescence* (XRF) analysis [1] is applied in the field of preservation and restoration of artistic heritage with the purpose to discover the chemical elements that were used during the preparation of a work.

The XRF technique is based on physical principles that describe the interaction between energy and matter, in particular the photoelectric effect: when an X-ray photon interacts with an atom, its energy is absorbed by an electron, which, as a consequence, is ejected from the atom, forming a gap and leaving the atom in an unstable state. To restore the atom stability, one of the electrons belonging to an external orbital fills up the gap; in doing so it emits energy in the form of a photon with a characteristic energy, which can be used to identify the atoms and the orbitals involved. The raw result of an XRF analysis is thus a spectrum of the detected energies, which can be represented as a histogram, as shown in Figure 1b for the work of art in Figure 1a. The spectrum gives the elemental composition of the analysed area, the dimensions of which depend on the spot of the X-ray source used; the spot diameter ranges from tens of microns for microfocus tubes equipped with polycapillary optics to about a *mm* for conventional X-ray tubes.

XRF spectrometry can be performed also in scanning mode, that is by moving the X-ray spot on the sample surface and acquiring a spectrum for each "point", called *pixel*, of a selected area. In this way, areas up to a few m^2 can be investigated. A scanning analysis, though more complex to perform, is more representative of the characteristics of the work of art. The result of a scan is a matrix that associates a spectrum to each pixel. The elemental maps can then be rendered assigning a color to each pixel depending on the associated spectrum, as shown in Figure 1c.

The instrument that has produced the XRF analyses used in this project is a spectrometer developed at the INFN Laboratory of nuclear techniques for Cultural Heritage (*LAboratorio*



(a) (b) (c)

Figure 1. The XRF scanning executed on a work of art produces a histogram (*spectrum*) and an image (*map*), which represent the distribution of detected energies in the scanned area.

di tecniche nucleari per i BEni Culturali, LABEC [2]) in Florence, which is part of the INFN *Cultural Heritage Network* (CHNet) [3]. The spectrometer, shown in Figure 2, consists of four main components:

- an X-ray tube;
- a detector;
- an ADC with a resolution of 14 bits. Each of the 16384 possible values is called a *channel*;
- a computer where the instrument management software runs.

The management software has three different purposes:

(i) setting the scanning parameters;
(ii) acquiring data;
(iii) visualizing the results of an analysis.

In order to address the first two purposes above the software is necessarily oriented to specialists. However the latter functionality, namely the visualization of the results, should be easily accessible also to non specialists, notably people working with cultural heritage. This has suggested the creation of a dedicated application pursuing just this goal and possibly extending the existing analysis functionality.

2. Towards a new analysis software

The main high-level requirements identified for the new analysis software are:

(i) accessibility to both CHNet researchers and cultural heritage operators;
(ii) an in-depth study of the data format;
(iii) the possibility to display the map;
(iv) the possibility to display the spectrum;
(v) the possibility to interact with the map and the spectrum, with each modification reflected in both entities.

The first requirement and the expected limited amount of computation required to the application have suggested to experiment with a web-based application, to avoid the development and especially the deployment of a native, multi-platform program. A user needs to simply point

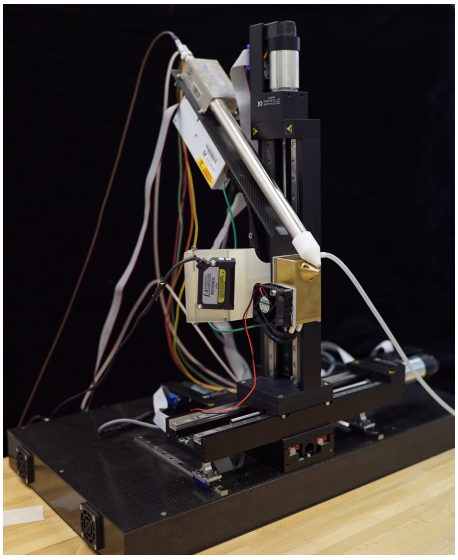


Figure 2. The INFN-LABEC Scanner.

her browser to a web page, currently hosted, together with some sample data files, on a server at CNAF.

As far as the language is concerned, HTML and CSS are the natural choices to implement the hierarchical structure of the web page. The application itself is instead written in Typescript [4], a statically-typed variant of Javascript, which is compiled into Javascript.

3. Image file and implementation

The image file created after a scanning is a text file consisting of a sequence of numeric rows. For each position of the work of art that is subject to the scanning, the position and the sequence of detected energies are stored. The scanning area, for this reason, is divided into small squares of the order of one mm^2, called pixels.

The format of the file has shown some shortcomings:

- some essential information (*metadata*) needed for a correct data interpretation are absent, for example the two constants to convert from a channel to the corresponding energy or the orientation of the work of art;
- the file size is excessively large;
- the map edges include inconsistent data and they have to be removed to correctly handle the map;
- if the scan is horizontal, the pixels on even rows are stored shifted to the right and their position needs to be corrected before being displayed. Similarly for pixels on even columns, if the scan is vertical.

Some of the problems can be fixed in software (e.g. the horizontal or vertical shifts), others are inherent in the file format (e.g. the lack of essential metadata) and could be properly addressed in the data acquisition software.

After the image file has been read, it is possible to represent both the map and the spectrum. The tool used to draw the map is a HTML canvas, which defines a drawable region of the web page. This region is filled with small colored squares, each corresponding to one pixel of the image. The color is computed as a function of the number of counts detected for that pixel. The

spectrum is instead drawn using the open-source Dygraphs library [5], which at the moment is the best fit for our needs.

Other features implemented so far and worth mentioning are:

- to zoom both the map and the spectrum;
- to export both the map and the spectrum in jpg format;
- to display the distribution of a selected chemical element on the map;
- to represent the spectrum in linear or logarithmic scale.

4. Conclusions and future outlook

The outcome of the project summarized in this article confirms that the use of a web application for an in-depth study of the data obtained by an XRF scanning analysis is a viable option.

We have also identified some areas that would require additional developments:

- Image file format: the use of an established standard format to store a large amount of hierarchical data (such as HDF5 [6], FITS [7] or EDF [8]) would allow to include the necessary metadata in the image file and possibly reduce the load and processing time;
- Remote access to the data files: this depends on satisfactory authentication and authorization mechanisms to discriminate among different users or groups;
- Overhaul of the graphical layout: the graphical aspect of the web application has not been the focus of the project so far and the adoption of one of the various available frameworks will certainly make it more attractive.

Acknowledgments

This work was the subject of the stage project of one of the authors (Cappelli) in collaboration with CNAF and in preparation of her thesis in Computer Science at the University of Ferrara [9]

References

[1] A Mazzinghi, 2016, Sviluppo di strumentazione XRF a scansione per applicazioni ai Beni Culturali.
[2] LABEC – LAboratorio di tecniche nucleari per i BEni Culturali, `http://labec.fi.infn.it/`.
[3] CHNet – Cultural Heritage Network, `https://chnet.infn.it/`.
[4] TypeScript, `https://www.typescriptlang.org/`.
[5] Dygraphs, `http://dygraphs.com`.
[6] The HDF Group, 2017, What is HDF5?, `https://support.hdfgroup.org/HDF5/whatishdf5.html`.
[7] NASA, 2016, The FITS Support Office, `https://fits.gsfc.nasa.gov/`.
[8] D Alvarez-Estevez, 2003, European Data Format, `http://www.edfplus.info/`.
[9] L Cappelli, Progettazione e sviluppo di un'applicazione web per la fruizione di dati di analisi XRF a scansione, `https://gitlab.com/Laura115312/Tesi`.

CNAF Monitoring system

S. Bovina, D. Michelotto, E. Fattibene, A. Falabella, A. Chierici
INFN-CNAF, Bologna, IT
E-mail: stefano.bovina@cnaf.infn.it, diego.michelotto@cnaf.infn.it, enrico.fattibene@cnaf.infn.it, antonio.falabella@cnaf.infn.it, andrea.chierici@cnaf.infn.it

1. Abstract
Last year, the CNAF monitoring system was reviewed to obtain a centralized and scalable monitoring infrastructure, being able to manage monitoring of a constantly growing infrastructure at reduced cost. In this report we are going to describe the consolidation activity and the new features introduced in 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. Monitoring consolidation
The main activity in 2016 has been the consolidation of CNAF monitoring infrastructure.

This task involved a lot of effort which was spent in several tasks such as the upgrade of InfluxDB from 0.9.0 version to 1.0.0, required to improve read and write performance, and to resolve blocker bugs related to downsampling and incremental backup procedures.

The database upgrade task involved the following steps:

- Migration of operating system from CentOS 6 to CentOS 7;
- Setup of InfluxDB instances dedicated to each functional unit, in order to be able to manage workload more efficiently and to separate contexts among functional units.

Following database migration, the backup procedure was reviewed and to allow the extrapolation, viewing and analysis of data on a high timespan, appropriate retention policies and continuous queries have been create to downsampled data, reducing its granularity.

Thanks to this modification, data is written into the default *one_week* retention policy. Every 15 minutes/30 minutes/1 hour we downsample raw data with continuous query into *one_month*/*six_months*/*infinite* retention policy respectively, in order to aggregate and reduce data granularity. In this way, we have a fine granularity for the data retrieved during last week, and a wider one for the following months, keeping all the data in the *infinite* retention policy.

In addition to this activity, a lot of effort was spent to connect farming nodes to the monitoring infrastructure (now 1700 nodes are using it) and to create custom dashboards related to CNAF mission.

3. CNAF Tier1 monitoring dashboards

Before using this new infrastructure, different tools were used to perform monitoring operations, such as Lemon [5] (developed at CERN and no longer maintained), and a system based on Graphite database and ad-hoc web pages. In order to completely migrate to the new monitoring system, we had to adapt Lemon sensors to be compliant with the Sensu/InfluxDB system. These sensors are custom Perl or Python scripts used to collect specific metrics, such as the number of recall and migrations from/to the tape system, the data throughput from the GPFS servers, and some metrics related to the StoRM [6] services running at CNAF. Other Lemon sensors used to collect common system metrics were replaced by Sensu community plugins.

Once saved monitoring data in InfluxDB using Sensu, we started to realize plots using Grafana, to easily show monitoring information at CNAF operators and users. The aim of this activity was to realize the same web views and plots given by the Lemon web interface. At the end of this activity, all Lemon web views and plots were rebuilt in Grafana. Moreover different dashboards have been realized on the basis of data stored in Graphite database.

The Grafana instance running for Tier1 monitoring contains 43 dashboards, grouped in 4 categories:

- *Overview*: the most significant plots aggregated per functional units (farming, storage, network, facility) and LHC experiments (ALICE, ATLAS, CMS, LHCb);
- *Storage*: information about the storage systems and services, resources (disk and tape) usage by VOs, data throughput;
- *Farming*: containing job monitoring dashboards, accounting information and computing services status;
- *Facility*: showing power consumption and temperatures.

As an example, Figure 1 shows an overview dashboard for CMS experiment.

Figure 1. CMS overview dashboard

In the top-left plot "CMS jobs", jobs executed at CNAF over time are represented, grouped per job status. The other plots give information on data throughput from and to CNAF disk and tape systems: CMS data at CNAF are on a unique GPFS filesystem (gpfs_tsm_cms reported in the plots), shared between an only-disk section and another for tape and disk buffer.

The top-right plot "CMS GPFS server traffic" shows in and out data throughput from/to GPFS NSD servers. These are data network-accessed by CNAF Worker Nodes. In the bottom-left plot "CMS GridFTP servers traffic" writing and reading data throughput through GridFTP servers is represented: these can be transfers from/to outside CNAF (via WAN) or internal (such as to CNAF Worker Nodes).

The bottom-right plot "CMS HSM server traffic" shows writing and reading data traffic on the filesystem from HSM (Hierarchical Space Management) server connected to the tape library; In particular "gpfs_tsm_cms read" is done by data read from disk buffer (previously written there from outside) that are written on tapes, and "gpfs_tsm_cms write" shows data written on disk buffer after have been read from tapes.

A comparison between the Lemon and Grafana dashboards can be seen in Figures 2 and 3.

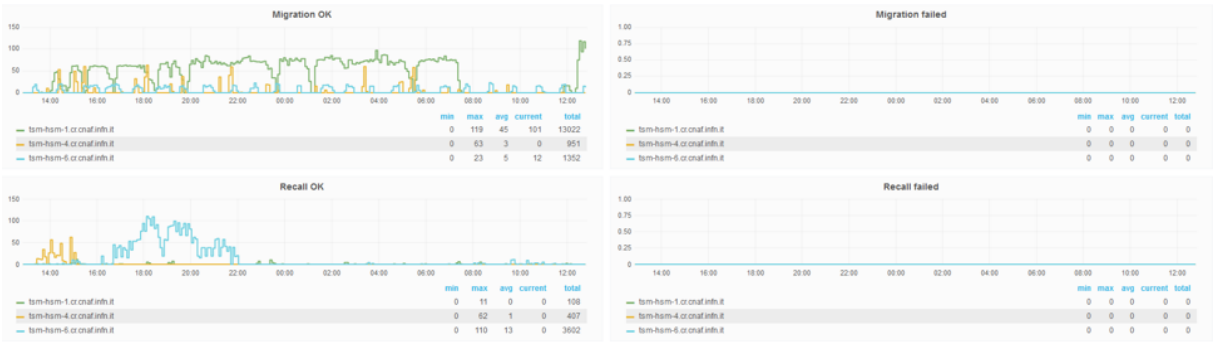


Figure 2. Tape recalls and migrations on Grafana

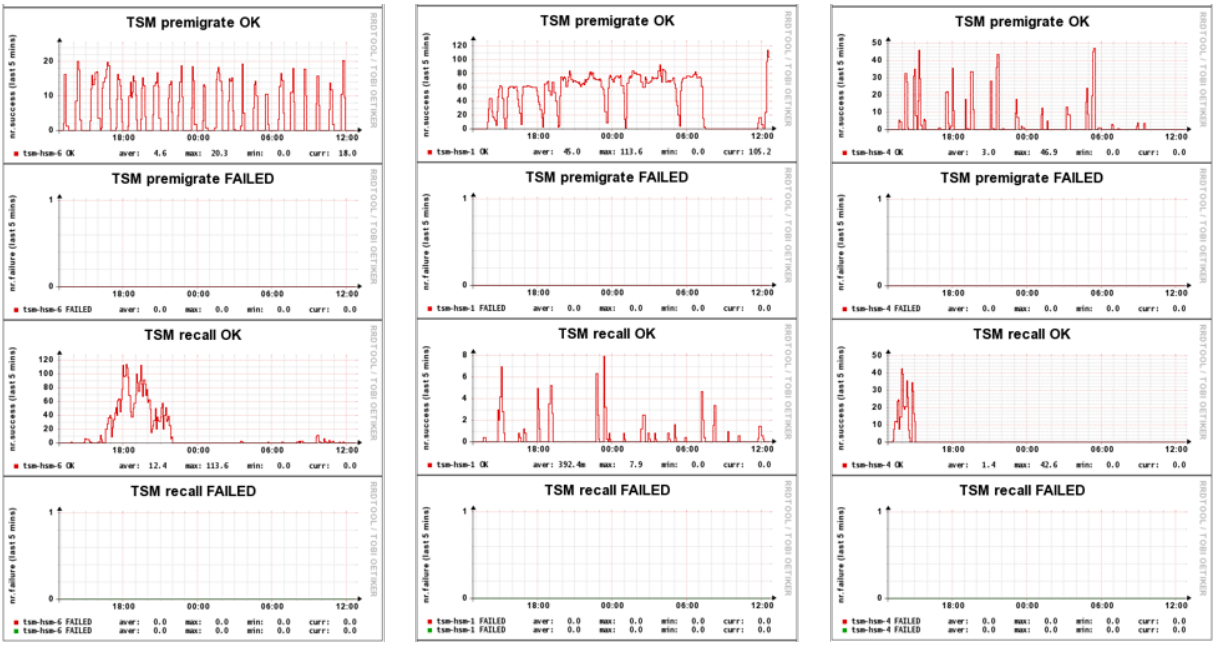


Figure 3. Tape recalls and migrations on Lemon

Figure 2 show plots taken from 3 different web pages (one for each VO) as there is no way to plot all these data together in Lemon. Figure 3 shows the same information as depicted in

a unique dashboard, where a dropdown menu gives the possibility to chose a desired subset of experiments (in this case ATLAS, CMS and LHCb) and HSM servers. In order to minimize the number of monitoring infrastructures at CNAF, Tier1 group have planned to migrate historical monitoring data from Lemon/Graphite to InfluxDB.

4. 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] Lemon paper: 2011 LAS LEMON - LHC Era Monitoring for Large-Scale Infrastructures J. Phys.: Conf. Ser. 331 052025
[6] StoRM web page: https://italiangrid.github.io/storm/

Developing software in a conservative environment

D. Bonino[1], V. Capobianco[1], L. Corcione[1], F. Fornari[2], F. Giacomini[3], S. Ligori[1], L. Patrizii[2] and G. Sirri[2]

[1] INAF-OATo, Pino Torinese (TO), IT
[2] INFN, Bologna, IT
[3] INFN-CNAF, Bologna, IT

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

Abstract. The software controlling the NISP instrument on board the Euclid space mission, whose launch is foreseen in 2020, is an application based on the RTEMS Operating System running on a LEON2 processor. The official development environment is very conservative and consists of an old Debian distribution augmented with space-qualified, and similarly old, versions of a gcc-based cross-compiler suite and of RTEMS.

Doing daily software development on a host platform that is about ten years old prevents the use of more modern tools, such as version control, editors, integrated development environments and analyzers, that would greatly improve developers' productivity and software correctness.

The modernization of the development ecosystem has started with the provision of a continuous integration infrastructure, based on containerization technology. Container images that have been made available include both the official toolset and some more advanced tools.

1. The NISP instrument on board the Euclid mission

Euclid [1, 2] is an ESA space mission whose launch is planned for 2020. It aims at addressing questions related to fundamental physics and cosmology on the nature and properties of dark energy, dark matter and gravity.

Euclid, shown in Figure 1, will be equipped with a 1.2 m telescope feeding two instruments: a high-quality visible imager (VIS) and a near-infrared spectrometer and photometer (NISP).

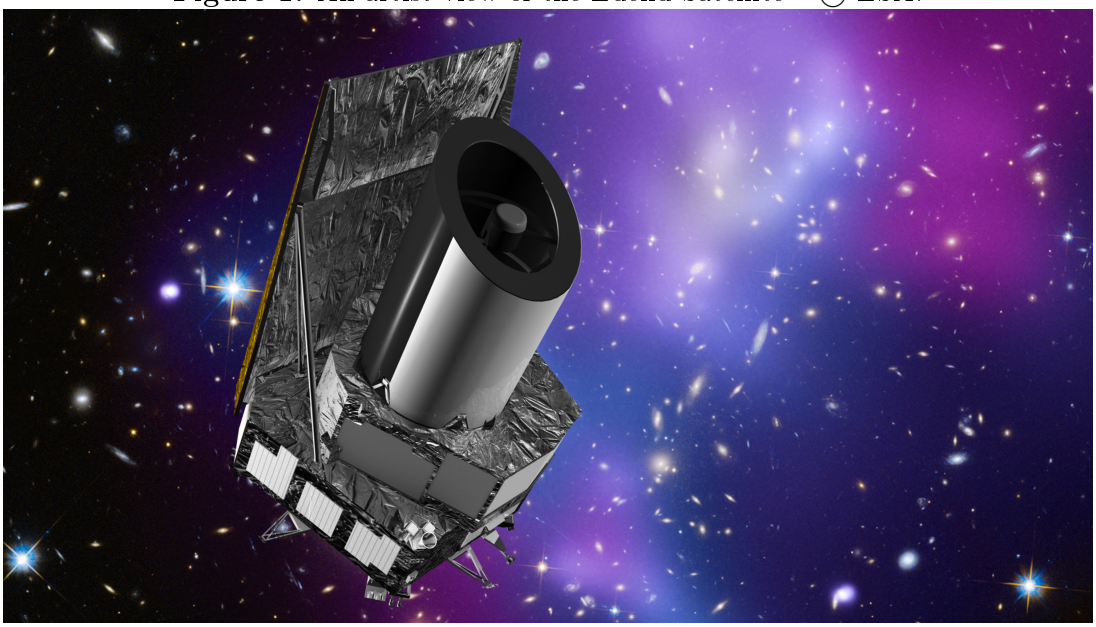
The work described here is focused on the NISP instrument [3]. NISP is composed of multiple units:

- the Opto-Mechanical Assembly, which includes the optics, the filter wheel assembly and the grism wheel assembly;
- the Detection System, which includes the detectors and the ADCs;
- the Data Processing Unit and Data Control Unit, which process and store the data coming from the Detection System;
- the Instrument Control Unit (ICU), which controls all the other subunits and communicates with the spacecraft via telecommands and telemetries.

2. The NISP-ICU software ecosystem

The application software running on the ICU is written in the C language on top of the RTEMS operating system. The processor is a LEON2, an implementation of the 32-bit SPARC-V8 architecture.

Figure 1. An artist view of the Euclid Satellite – © ESA.



Such a platform is not available as a desktop machine usable for the daily software development. Software development happens instead on a typical Linux host and binaries for the ICU native platform are produced thanks to a cross-compiler.

2.1. Operating systems and compilers

The operating system running on the ICU is a space-qualified variant of the Real-Time Executive for Multiprocessor Systems (RTEMS) [4], an open-source Real Time Operating System widely used in avionic, medical, networking and other embedded applications. It has been ported to many architectures, including SPARC, ARM, PowerPC, Intel. The RTEMS variant used for the ICU has been space-qualified by Edisoft [5] and is based on the vanilla version 4.8.0, released in 2008. It is distributed by Edisoft either as a *zip* file of the entire modified source tree or as a *patch* against the vanilla source tree.

The application binary code has to be obtained through *cross-compilation*, a process by which the compiler running on a host produces a binary for an architecture other than the one where it is running on.

The cross-compiler used for the development of the ICU application software is a certified variant of the *gcc* compiler suite [6], based on the vanilla version 4.2.1, released in 2007. Like for RTEMS, it is distributed by Edisoft, either as a *zip* file of the entire modified source tree or as a *patch* against the vanilla source tree.

The instructions for the use of the space-qualified RTEMS and of the certified gcc are documented for a Debian 5 ("lenny") distribution [7], first released in 2009 and last updated in 2012. This is then the platform where an official binary for the ICU software has to be produced.

2.2. Issues with an old computing environment

Relying on an old computing environment, especially if properly qualified, has some advantages in terms of stability and economies of scale. It presents however at least two issues that make it less than ideal for a modern software development process:

- the platform may contain security vulnerabilities unknown at the time of qualification and never fixed afterwards. The risk increases if the machine is exposed to the network;
- a developer cannot benefit from advancements in tools such as version control, editors, compilers and analyzers.

The next sections presents a few fundamental techniques that have been introduced in the development process for the ICU application software in order to address the above-mentioned issues, keeping at the same time the reproducibility of the correct production of the binary artifacts.

3. Reproducible development environment

A reproducible build ensures that a given source code is always transformed into the same binary code, provided the build environment is the same. This guarantee represents a solid foundation for all the activities aimed at providing high-quality software products.

A reproducible development environment requires at least that:

- all the changes to the source code are uniquely identifiable;
- the build environment is well defined and itself reproducible.

3.1. Source code management

The source code for the ICU application software is kept under the *git* software configuration management tool. The authoritative repository is kept on a GitLab service hosted by INFN, which also integrates a continuous integration platform.

The development workflow is based on feature branches and pull requests, which give the possibility to do some peer-review on the proposed changes. Each push operation to the central repository and each merge request are automatically built on a continuous integration infrastructure and the corresponding logs kept.

The continuous integration at the moment is limited to the build task, but static analysis tools are being integrated (see below).

3.2. Build environment management

The authoritative build environment is represented by a *Docker* [8] image. Docker is a software containerization platform: a container wraps a piece of software in a complete filesystem that contains everything needed to run it. Every time a container is started, the environment that it defines is pristine and retains no memory of previous executions.

The image used to build the application software is built in stages:

(i) a 32-bit Debian "lenny" image, built thanks to the *debootstrap* command and the Debian archives, is used to build the *sparc-rtems4.8-gcc* cross-compiler;

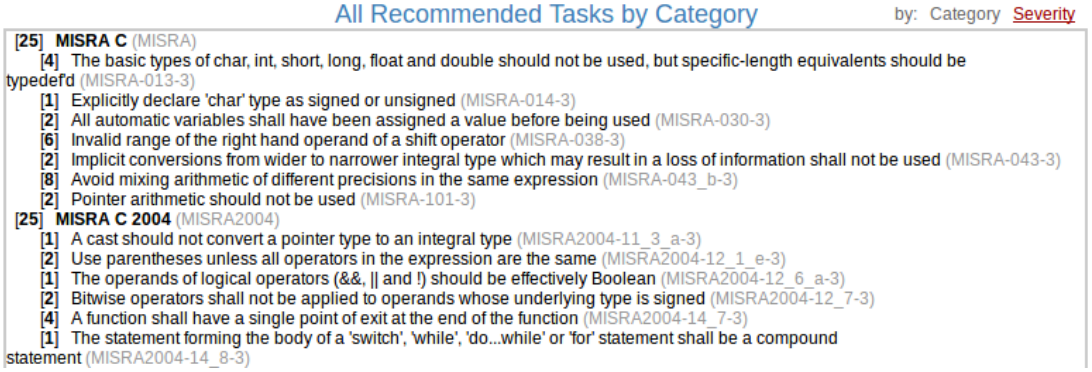
(ii) the same 32-bit Debian "lenny" image is augmented with the cross-compiler. The new image is used to build the RTEMS binary distribution;

(iii) the final image is obtained extending the image of the previous point with the RTEMS distribution, both in source and binary forms.

The full procedure is reproducible thanks to shell scripts and *Dockerfiles*, simple text files describing how to create an image. The scripts and the Dockerfiles for all the containers mentioned above are themselves kept in a git repository, hence guaranteeing the reproducibility of the build environment.

Listing 1 shows an example Dockerfile, representing the last step of the above procedure.

Figure 2. Some MISRA violations detected by C/C++Test.



Listing 1. Example of Dockerfile.

```
FROM icu-devel-stage1

ADD edilib.tgz /opt/
COPY rtems-impr /home/rtems/rtems-impr
ADD rtems.tgz /home/rtems/
```

4. Static analysis

The first validation of the application software code is performed with static analysis tools.

The most used at the moment is *C/C++Test* by Parasoft [9]. The configurations applied to the code are "MISRA C", to check the compliance with the MISRA C rules, and "BugDetective Aggressive", which implements flow-based checks; no customization has been applied yet to the two configurations. Figure 2 shows an example of the output produced applying the "MISRA C" configuration.

Some experimentation with tools of the *clang* family [10] is ongoing. More specifically *clang-format* is used to format the code according to an agreed-upon specification and *clang-tidy* is used to check for violations to a set of rules. The functionality offered by the clang tools overlaps, at least in part, with that offered by C/C++Test; however its open-source license makes its use certainly more flexible. clang-format and clang-tidy are also being included in the continuous integration system in order to provide early feedback to the developers.

Other tools are under considerations, namely *Coverity* and *PVS-Studio*, whose use is free for open-source projects.

5. Conclusion

Developing code for a space mission requires necessarily the adoption of a conservative environment to produce the application binary artifacts. This however does not mean that the software development itself needs to be done using decade-old tools.

The use of modern technologies such as git, Docker and static analysis tools makes it easy to implement a fully reproducible process that allows the automatic and continuous production and testing of binary artifacts for the target architecture.

References

[1] Renè J. Laureijs, et al., "Euclid Definition Study Report", arXiv:1110.3193 [astro-ph.CO], `https://arxiv.org/abs/1110.3193` (2011).

[2] Renè J. Laureijs, et al., "Euclid mission status," SPIE **9143** (2014).

[3] Thierry Maciaszek, et al., "Euclid near infrared spectro photometer instrument concept and first test results at the end of phase B," SPIE **9143** (2014).
[4] RTEMS Real Time Operating System, `https://www.rtems.org/`.
[5] Helder Silva, et al., "RTEMS CENTRE – RTEMS Improvement," in [Proceedings of DASIA 2010 Data Systems In Aerospace], Ouwehand, L., ed., **ESA-SP 682**, id 38 (2010).
[6] GCC, the Gnu Compiler Collection, `https://gcc.gnu.org/`.
[7] Debian "lenny" Release Information, `https://www.debian.org/releases/lenny/`.
[8] Docker, `https://www.docker.com/`.
[9] Parasoft C/C++Test, `https://www.parasoft.com/product/cpptest/`.
[10] clang: a C language family frontend for LLVM, `https://clang.llvm.org/`.

Technology transfer and other projects

External Projects and Technology Transfer

C. Vistoli, B. Martelli, A. Ferraro
INFN-CNAF, Bologna, IT
E-mail: cristina.vistoli@cnaf.infn.it

Abstract.
In 2016 we saw the take-off of the new External Projects and Technology Transfer CNAF Organizational Unit. Its mission is the regional coordination of external fund projects and technology transfer activities carried out by INFN-Bologna, INFN-Ferrara and INFN-CNAF departments, as required by the Regione Emilia Romagna funding policies. In order to effectively participate in the Emilia Romagna industrial research and development activities, this unit has created the INFN Technology Transfer Laboratory, which has been accredited to the Emilia Romagna High Technology Network (HTN) at the end of 2015 and immediately submitted his first project ideas as proposals to the Emilia-Romagna-POR-FESR 2014-2020 call. The proposals ranged from nanotechnologies to nuclear medicine, cultural heritage, internet of things and embedded systems and brought together the expertise of all INFN personnel in Emilia Romagna.

1. Introduction

During 2016 the External Projects and Technology Transfer (PETT) Organizational Unit has contributed to various projects in the field of computing, communication of science, technology transfer and education. Some of the most relevant ones are: FiloBlu (POR-FESR Regione Marche), Opus Facere (MIUR), European Research Night, Emilia Romagna Plan for high competencies in Research, Technology Transfer and Entrepreneurship, Open-Next (POR-FESR 2014-2020), Harmony, EEE (Extreme Energy Events). Great effort has been put on the start up phase of the new Technology Transfer Laboratory which put together heterogeneous competencies (physics, computing, mechanics and electronics) from Emilia Romagna INFN Sections and Centers (Bologna, Ferrara and CNAF) in order to promote the transfer of INFN know-how toward regional enterprises. At the end of the year we started an activity finalized to the ISO-27001 certification of a subset of the INFN Tier1 resources. This is required in order to store and manage private and sensitive personal data and could open new opportunities of exploitation of the Tier1 resources in the near future.

2. Commissioned Research on Immersion Cooling

In the field of the Emilia Romagna High Technology Network, a local star-up has asked the Technology Transfer Laboratory (TTLab) to validate an innovative, patent pending, immersion cooling solution for big data center based on a biological fluid. TTLab is drawing up a commissioned research contract for testing and optimizing the immersion cooling prototype. The activity will start with computational validation tests on the Tier1 data center resources. In case of success, the next step will perform fluid dynamics and thermodynamics tests and

Table 1. Estimated cost of different low-power clusters performing 50 pipelines in a day .

Head 1	Cost/Motherboard() 2	N. Motherboards 3	Estimated total cost () 4
Xeon D-1540	1000	11	11000
Atom C2750	430	24	10320
Pentium N3700	85	31	2604

studies aimed at helping the start-up in the process of optimizing and bringing on the market its product.

3. Genome Sequencing Pipeline Optimization

Thanks to the heterogeneous competencies represented by the TTLab personnel (G. Castellani, D. C. Duma, I. Do Valle, A. Ferraro, B. Martelli, D. Remondini, E. Ronchieri, C.Vistoli), in 2016 a collaboration started between biophysics scientists and computing experts finalized at optimizing the execution of the GATK MuTect software pipeline used in genome sequencing analyses. We wanted to characterize the execution of this pipeline on different platforms: low power architectures, HPC cluster, Virtual Machines provisioned by the Cloud@CNAF infrastructure and multicore servers. The activity aims also at studying the parallelism mechanism natively provided by the software and investigating possible improvements by using e.g. the GATK-Queue framework. Queue, in conjunction with the LSF batch system, enables users to semantically define multiple pipeline steps, run independent jobs in parallel, handle transient errors and run multiple copies of the same program on different portions of the genome to speed up the analysis. A monitoring system based on Telegraph and Grafana has been installed on all the computing platforms and monitoring data about CPU load, RAM usage and disk access patterns have been collected during the pipeline execution. Our preliminary results seem to state the high efficiency of low power architectures. More investigation is needed to confirm these results and to deepen our understanding of parallelism in GATK and Queue. Some results from the low-power (Intel only architectures) testbed are shown (see Fig. 1).

The low-power cluster show interesting points of discussion and debate. We point out that Xeon D-1540 rough performance overtakes those of all the other architectures, such as Atom C2750 and Pentium N3700 (see Tab. 2): however, looking results from different perspective (as economic issue), the results lead to different conclusions. For example an economic costs/performance evaluation can be inspired by the following statement: “How much does it cost a low-power cluster in order to run at least 50 pipelines in a day?”

4. OPEN-NEXT project

TTLab is involved in the OPEN-NEXT project that won a call for strategic industrial research projects directed to priority areas of the Strategy of Intelligent Specialization fonded by the POR-FESR 2014-2020 program of Regione Emilia-Romagna (PG/2015/737416).

The aim of the TTLab research in the OPEN-NEXT project is to investigate the best parallelization strategies of real applications that fulfil the requirements of low power consumption. The analysis wants to offer a few applications performance comparative study both on single and many core platforms. The activity will port some not-optimized server-grade applications to embedded low-power n-core platforms.

Energy consumption is one of the most relevant issue for the TTLab activity in the project. A few parallel programming paradigms for applications running in multi-core low-power

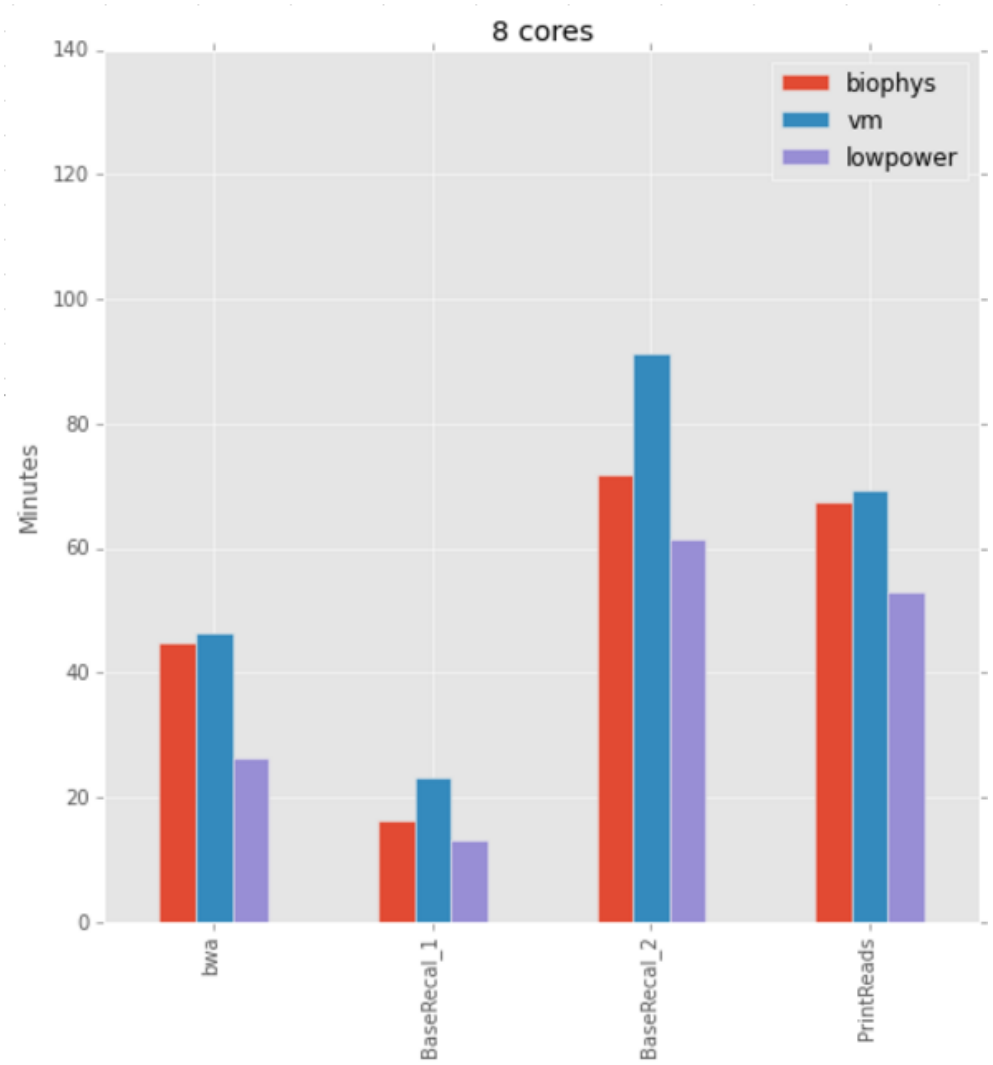


Figure 1. First results showing low-power machines leading in performing tests of four multicore pipeline stages

architectures were studied and investigated. We evaluated all the strenghts of energy-efficient SoC (System on Chip) boards ranging from multi-core embedded CPUs to many-core embedded GPUs designed to meet the demands of the mobile and embedded market. The case of off-the-shelf SoCs various limitations (32-bit architectures, small CPU caches, small RAM sizes, high latency interconnections, ECC memory not available, etc.) are not a problems because OPEN-NEXT applications are mainly industrial and embedded applications usually run in simple 1-core CPUs (we did not evaluate HPC or HTC server-grade applications).

The activity of TTLab in OPEN-NEXT was inspired by the collaboration with COSA (COmputing on SoC Architectures) project, a three-year INFN project funded by the Scientific Commission V of INFN and led by CNAF, which aims to investigate the performances and the total cost of ownership offered by computing systems based on commodity low power Systems on Chip (SoC) and high energy-efficient systems based on GP-GPUs. We want to point that COSA results are relevant for the data center HPC/HTP scientific community, while OPEN-NEXT results are relevant for the industrial and automation sector.

Brand		Xeon D-1540 (TDP: 45W)						Atom C2750 (TDP: 20W)					Pentium N3700 (TDP: 4.5W)			
Cores/Freq		8 / 2.00						8 / 2.40					4 / 1.60			
Cache		12MB						4MB					2MB			
Lithography		14nm						22nm					14nm			
Release date		Q1'15						Q3'13					Q1'15			
	Multi Cores	16c(HT)	8c	4c	3c	2c	1c	8c	4c	3c	2c	1c	4c	3c	2c	1c
bwa mem	*	1246	1573	3000	3985	5906	11695	3659	6738	9961	15094	30201	7927	10130	15371	30201
SortSam		1791						3297					3529			
MarkDuplicates		1506						3348					3569			
BuildBamIndex		190						376					409			
IndelRealigner		1368						2773					3134			
BaseRecalibrator	*	863	786	895	977	1153	1662	2160	2744	2905	3470	4843	2714	2827	3280	4654
BaseRecalibrator 2nd	*	3824	3688	3779	3915	4068	4588	7080	8176	8557	9098	10479	8071	8346	8741	10187
AnalyzeCovariates		2						5					6			
PrintReads	*	3101	3186	3173	3576	4475	7868	5160	7190	8367	10917	15613	7937	8669	11163	15303
HaplotypeCaller		4915						13569					15462			
SelectVariants SNP		18						42					43			
VariantFiltration		15						41					44			
SelectVariants INDEL		17						39					42			
VariantFiltration		7						18					17			
CombineVariants		12						34					38			
TIME (s)		18875	19074	20688	22294	25443	35654	41601	48390	53332	62121	84678	52942	56265	64848	86638
TIME (m)		314,58	317,89	344,79	371,56	424,04	594,23	693,35	806,50	888,87	1035,35	1411,30	882,37	937,75	1080,80	1443,97
TIME (h:m:s)		05:14:35	05:17:54	05:44:48	06:11:34	07:04:03	09:54:14	11:33:21	13:26:30	14:48:52	17:15:21	23:31:18	14:42:22	15:37:45	18:00:48	00:03:58

Figure 2. Low-power architectures performance. The result of three low-power Intel families of CPU. Yellow rows show the four pipelane stages (bwa mem, BaseRecalibrator, BaseRecalibrator 2ndm PrintReads) that benefit from multicore runs

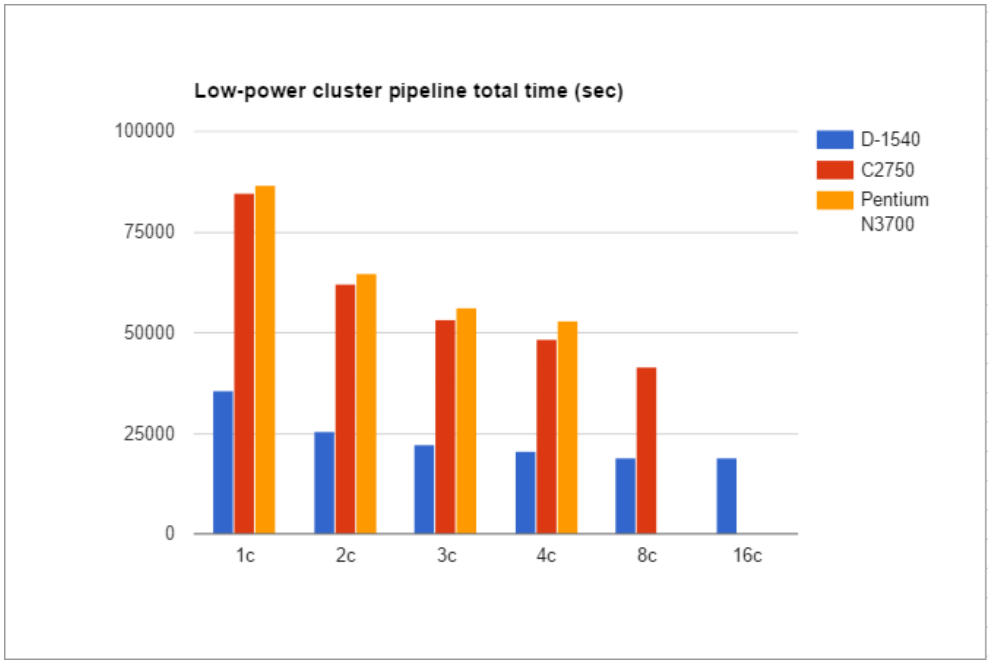


Figure 3. Low-power architectures perfomance

TTLab collaborates with CIRI-ICT (University of Bologna) and Softech (University of Modena and Reggio-Emilia) that lead the respective workpages: OR2 (CIRI-ICT): Programming model for n-core real-time architectures OR3 (SOFTECH-ICT): Optimization of real-time operating systems

The OR2 collaboration with CIRI-ICT is focused on identify and use wide-spread standard

programming models as OpenMP, OpenCL, OpenVX, and evaluate run-time libraries tat minimize the memory usage and power consumptions. We identified several applications to port and test to embedded low-power hardware as a computer tomography application developed by INFN. INFN-CNAF already ported (in COSA project) the computer tomography application from an Intel x86 platform to an ARM SoC with embedded GPU. In OR2 TTLab and CIRI-ICT efforts are directed towards unusual platforms as Kalray and PULP (Parallel Ultra Low-Power Processing) platforms. The activity (in collaboration with COSA project) consisted in porting several computing workloads to low-power architectures, investigating the computing and energy performance and comparing them with traditional data center servers. We have evaluated different programming paradigms and manual hardware/OS tuning (as manual clock frequency tuning with respect to the hardware governor, looking for an optimal trade-off between energy-to-solution and time-to-solution, etc.).

The OR3 collaboration with SOFTECH-ICT is focused on testing SOFTECH-ICT proposed RT Linux scheduling mechanisms in TTLab laboratory low-power hardware and evaluate their benefit for selected real-time applications.

Regular meetings were organized in order to exchange experience among partners. Presentations of OPEN-NEXT and the relative TTLab activities were performed at seminars for small-medium local companies and researches from INFN and universities.

5. Setup of an Information Security Management System for ISO 27001 certification

TTLab personnel, together with Tier1 staff and biophysics people from University of Bologna, started the design and establishment of an ISO27001-compliant ISMS (Information Security Management System) for data sensitive information to be hosted in the Tier1 data center in 2017. The adoption of an ISMS was triggered by the INFN participation to international projects requiring high-security standards and can be viewed as an opportunity for INFN-CNAF for future projects and collaborations requiring strict security rules. So the design and implementation of the ISMS is influenced by needs, scope and objectives of the current projects, but is expected that the designed implementation will be scaled in accordance with the needs of the organization and future projects.

The design and establishment activity started in the last quarter of 2016 and will continue throughout the 2017 year. The ISMS design will converge in a ISO27001 certification expected in the middle of 2017 year. After obtaining INFN-CNAF management support and acquiring awareness that ISO 27001 implementation of an ISMS is a complex issue involving various activities, lots of people, and lasting several months, we chose to follow a project management approach in order to clearly define what is to be done, who is going to do it and in what time frame. After the performing of the ISO 27001 certification, the ISMS will be operated, monitored, reviewed, maintained and improved in a continuos cycle. The 2016 activity was focused to complete the following very first preliminary phases required for establishing a ISO27001-compliant ISMS: - define the scope and boundaries (we decided to implement ISO 27001 only in one part of our organization, thus significantly lowering overall risk) - define the overall ISMS policy that should be approved by the INFN-CNAF management (it is the highest-level document in the ISMS, it is not very detailed, but it defines some basic issues for information security in the scope of ISMS. The purpose is for management to define what it wants to achieve, and how to control it) - define the risk assessment approach of the INFN-CNAF (identify a risk assessment methodology suited to the ISMS, develop criteria for accepting risk. The point is to define the rules for identifying the assets, vulnerabilities, threats, impacts and likelihood, and to define the acceptable level of risk.)

References

[1] GATK, *https://software.broadinstitute.org/gatk*
[2] Queue, *http://gatkforums.broadinstitute.org/gatk/discussion/1306/overview-of-queue*
[3] LSF, *https://en.wikipedia.org/wiki/Platform_LSF'*

COmputing on SoC Architectures: the INFN COSA project

D. Cesini[1], E. Corni[1], A. Falabella[1], A. Ferraro[1], G. Guidi[1], L. Morganti[1], E. Calore[2], S. F. Schifano[2], M. Michelotto[3], R. Alfieri[4], R. De Pietri[4], T. Boccali[5], A. Biagioni[6], F. Lo Cicero[6], A. Lonardo[6], M. Martinelli[6], P. S. Paolucci[6], E. Pastorelli[6], P. Vicini[6]

[1] INFN-CNAF, Bologna, IT
[2] University of Ferrara and INFN, Ferrara, IT
[3] INFN, Padova, IT
[4] INFN, Parma, IT and University of Parma, IT
[5] INFN, Pisa, IT
[6] INFN, Roma, IT

E-mail: daniele.cesini@cnaf.infn.it

Abstract. Energy consumption represents one of the most relevant issues by now in operating HPC systems for scientific applications. The investigation and assessment of unconventional processors with high ratio of performance per watt is interesting for a better trade-off between time-to-solution and energy-to-solution. Computing on SoC Architectures (COSA) is a three-year project (2015-2017) funded by the Scientific Commission V of INFN and led by CNAF, which aims to investigate the performances and the total cost of ownership offered by computing systems based on commodity low power Systems on Chip (SoC) and high energy-efficient systems based on GP-GPUs. Here we present the results of the project and we describe the methodology we have used to measure energy performance and the tools we have implemented to monitor the power drained by applications while running.

1. Introduction

Energy consumption has increasingly become one of the most relevant issue for scaling up the performance of modern HPC systems and applications, and this trend is expected to continue in the foreseeable future. This implies that the costs related to the run of applications are more and more dominated by the electricity bill, and therefore the adoption of energy-efficient processors is necessary, ranging from many-core processors architectures like Graphic Processor Unit (GPU) to System on Chip (SoC) designed to meet the demands of the mobile and embedded market.

SoC hardware platforms are integrated circuits typically composed of low power multicore processors combined with a GPU and all the circuitery needed for several I/O devices. These processors feature an high performance-per-watt ratio, aimed at energy efficiency, but require carefully programming and optimization to be, at same time, computing efficient. Moreover, for the case of off-the-shelf SoCs various limitations may arise: 32-bit architectures, small CPU caches, small RAM sizes, high latency interconnections, ECC memory not available.

Investigating and assessing the performance of these systems for scientific workloads is the aim of the COmputing on SoC Architectures (COSA) project [1], a 3-year initiative funded by

the INFN, coordinated by CNAF and started in January 2015, as a natural prolongation of the INFN-COKA (COmputing on Knights Architecture, [2]) project.

In this work we explore the performance of energy efficient systems based on multi-core GPUs, low-power CPUs and SoC systems. We have ported several scientific workloads and we have investigated the computing and energy performance comparing them with traditional systems mainly based on x86 multi-core systems. We have also evaluated the benefits of manual clock frequency tuning with respect to the hardware governor, looking for an optimal trade-off between energy-to-solution and time-to-solution.

2. The COSA clusters

The COSA project built and maintains three computer clusters, located at CNAF department (Bologna), ROMA1 department and Padova department. Moreover, a fourth cluster has been installed in Ferrara with the main contribution of the University of Ferrara.

2.1. The low power cluster based on SoCs

CNAF hosts an unconventional cluster of ARMv7, ARMv8 and x86 low-power SoCs nodes, interconnected through 1Gb/s and 10Gb/s Ethernet switches and used as a testbed for synthetic benchmarks and real-life scientific applications in both single-node and multi-node fashion.

The ARM cluster is composed by eight NVIDIA JETSON TK1, four NVIDIA JETSON TX1 (64bit), two ODROID-XU3, a CUBIEBOARD, a SABREBOARD, an ARNDALE OCTA board, all interconnected with standard 1Gb/s ethernet. We notice that the X1 cluster is connected with 1Gb/s ethernet eventhough a USB-ethernet bridge provides the physical connection with the SoC, increasing the latency with respect to standard 1Gb/s ethernet connections.

The 64bit x86 cluster is composed by four mini-ITX boards powered by the Intel C-2750 AVOTON SoC, by four mini-ITX motherboard based on the Intel Xeon D-1540 CPU and by four mini-ITX boards based on the Intel Pentium N3700 processor. The "Avoton", the "XeonD" and the "TX1"clusters are connected with both 1Gb/s and 10Gb/s Ethernet connections while the N3700 only with 1Gb/s Ethernet network. The 1 Gb/s connections are provided by onboard connectors, while the 10 Gb/s links are obtained with PCI Host Bus Adapter (HBA).

Table 1 summarizes and compares the relevant features of the boards in the low-power COSA cluster at CNAF. The Thermal Design Power (TDP) of the SoCs in this cluster, when declared, ranges from 5W of Intel Pentium N3700 to 45W of the 8-cores Intel Xeon-D Processor.

Ubuntu is installed on all the platforms. A master server is used as a monitoring station and an external network file system hosting all softwares and datasets is mounted on every cluster node. CPU frequency scaling is used by the Linux operating system to change the CPU frequency for saving power depending on the system load, and the recommended "ondemand" governor is enabled by default. We set governor to "performance" level, so to avoid dynamic CPU frequency scaling and maximize CPU performance. The GPU frequency was set to its maximum value (see Table 1).

2.2. The high-end hardware clusters

At INFN-CNAF, the traditional reference architecture is a x86 node from an HPC cluster, equipped with two Intel Xeon E5-2620v2 CPUs, 6 physical cores each, HyperThread enabled (i.e. 24HT cores in the single node), and with a NVIDIA K20 GPU accelerator card with 2880 CUDA cores. The HPC server is rated with a TDP of 160W for the two CPUs (about 80W each) and 235W for the GPU.

At Ferrara another traditional HPC reference architecture is available [3]. This cluster is made of 5 computing nodes, hosting each 2× Intel Xeon E5-2630v3 CPUs and 8× NVIDA K80 dual-GPU boards, interconnected with Mellanox MT27500 Family [ConnectX-3] Infiniband HCA (two per node). The TDP of each CPU is 85 W, while for an NVIDIA K80 board is

Table 1: The COSA cluster at INFN-CNAF. Top: ARM; bottom: Intel.

Platform	CPU	GPU	RAM
Freescale Sabreboard (iMX.6Q SoC)	4xA9 1.2 GHz	Vivante GC2100 600 MHz	2 GB
Hardkernel ODROID-XU-E (Exynos5 5410)	4xA15+4xA7 1.6 GHz	PowerVR SGX544 384 MHz	2 GB
Hardkernel ODROID-XU3 (Exynos5 5422)	4xA15+4xA7 2 GHz	ARM Mali-T628 533/695 MHz	2 GB
HiSilicon Kirin 6220	8xA53 1.2 GHz	ARM Mali-450 MP4 700 MHz	1 GB
Hardkernel ODROID-XU3 (Exynos5 5422)	4xA15+4xA7 2 GHz	ARM Mali-T628 533/695 MHz	2 GB
ARNDALE OCTA (Exynos 5420)	4xA15+4xA7 1.7 GHz	ARM Mali-T628 533/695 MHz	1 GB
Nvidia Jetson K1	4 cores +1 2.3 GHz TDP 10W	Nvidia K1 192 Kepler cores 852 MHz	2 GB
Nvidia Jetson X1	4 cores 1.73 GHz TDP 10W	Nvidia X1 256 Maxwell cores 998 MHz	4 GB
Intel Avoton C2750	8 cores 2.4 GHz TDP 20W	-	16 GB
Intel XeonD 1540	8 cores 2.6 GHz TDP 45W	-	16 GB
Intel Pentium N3700	4 cores 2.4 GHz TDP 6W	HD Graphics	16 GB

300 W, amounting for a total maximum power of 3.2 kW for each computing node or 16 kW for the whole cluster.

3. Power monitor tools

In the context of the COSA project, we developed and implemented a few fine grained power monitoring systems.

The first of these systems is a software wrapper [4] exploiting the PAPI Library [5] to read appropriate hardware registers containing power related information, often available on modern processors.

The second system requires a custom hardware [6] and is suited for those situations were a purely software approach is not viable due to unavailability of appropriate hardware register or

difficulties in their readout. The setup uses an analog current to voltage converter (using a LTS 25-NP current transducer) and an Arduino UNO board to digitize and store readings.

Another available power measurement equipment consists of a DC power supply, a high precision Tektronix DMM4050 digital multimeter for DC current measurements connected to a National Instruments data logging software, and a high precision AC power meter. In this case, AC power of the high-end server node is measured by a Voltech PM300 Power Analyzer upstream of the main server power supply (measuring on the AC cable). Instead, for the SoCs, the DC current absorbed downstream of the power supply.

4. Benchmarks and applications

In order to characterize the performances of the various architectures available in the COSA clusters, we performed the following synthetic benchmarks and scientific applications:

- home-designed tests based on PI-computation in a fixed number of iterations, computation of prime numbers within a given integers interval, and matrix multiplication [7, 8];
- High Performance Conjugate Gradients (HPCG [9]) Benchmarks;
- applications from the Theoretical Physics domain, e.g. Lattice Boltzmann Methods [6, 10, 11, 12, 13];
- High Energy Physics workloads, e.g. LHCb simulation and offline reconstruction softwares, as reported at the CCR Workshop in La Biodola in May 2016 (see also [8]);
- Spiking neural networks simulations [14, 15];
- X-Ray Computed Tomography applied to the field of Cultural Heritage [16];
- Einstein ToolKit [17, 18] for the high resolution simulation of inspiral and merger phase of binary neutron star systems, as presented at the CCR Workshop in La Biodola in May 2016 (see also [8]);
- Bioinformatics applications like the porting of GROMACS (GROningen MAchine for Chemical Simulations, [19]) to Jetson K1 [7], and of a space-aware system biology stochastic simulator to the newer Jetson X1 boards [20].

On the whole, our various works show that it is feasible, and sometimes even straightforward, to compile and run benchmarks and scientific workloads on low-power SoCs.

Computing performances are obviously worse than those obtained on HPC servers, particularly when attempting multi-node runs due to high latency connections, but gains in terms of power consumptions can be relevant.

As for high-end platforms, whenever the scientific workloads manage to exploit the large number of graphics cores available in the SoCs it is possible to obtain good speed gains and increase in computing power per Watt.

5. Image recognition using deep learning

In the context of the COSA project, at CNAF we are currently exploring the possibility of using SoCs for image recognition and classification making use of deep learning techniques. This work, which is still very much in progress, looks promising.

First, we generated large image datasets using ImageNet [21], and we improved the quality of such datasets in two ways, namely *i)* adding also images which can be found in sub-categories (not only in categories), and *ii)* removing junk web-images.

Then, we performed the training, which is the most computationally demanding task, on server GPUs, e.g. a Nvidia K20 (but mounted on a XeonD board) and a node from HPC cluster equipped with 2 Nvidia K40. The Nvidia Deep Learning GPU Training System (DIGITS, [22]) and the deep learning framework CAFFE [23] developed by the Berkeley Vision and Learning Center were used for the training task.

Finally, we moved the trained network on the low power Jetson X1 board, and used Nvidia TensorRT framework [24] for fast, or better real-time, inference. The compatibility was not instantaneous, and it required a bit of coding.

In this way, we could test image classification for several datasets and categories, and this looked satisfying and basically independent from the background of the images.

In the immediate future, we plan to explore image segmentation inference on Jetson X1.

6. Conclusions

The INFN COSA (Computing on SoC Architecture) project led by CNAF aims at investigating viable energy efficient computing systems for scientific applications.

System On Chip platforms based on both Intel and ARM architectures have been considered and benchmarked with synthetic tests and real life applications belonging to various scientific fields, ranging from High Energy Physics to Systems Biology, Theoretical Physics, Computed Tomography and Neural Network simulations. Moreover, in the framework of the COSA project, INFN-Roma is investigating the implementation of low latency toroidal network architectures to connect low power CPUs [15].

We found that it is actually possible, and in some cases even easy, to compile and run complex scientific workloads on low-power, off-the-shelf, devices not designed for high-performance computing.

For certain applications, the computing performances are satisfying and even comparable to those obtained on traditional high-end servers, with much lower power consumptions, in particular when the GPU available on the SoC is exploited.

Of course, many limitations still affect low power SoCs-based platforms. To name a few: small maximum RAM size, high latency of the network connections, few and small PCI slots, missing support for ECC memory. Hence, off-the-shelf systems can be hardly used for extreme scale applications and highly demanding HPC computations. However, the experience of the COSA project shows that systems based on low-power SoCs can be an interesting possibility to reduce power consumption if a proper integration is carried on.

Finally, we recall that another research line that the project is investigating to reduce energy consumption of HPC application runs does not focus on hardware but on software. Modern operating systems, in fact, allow to control the performance of various hardware components (i.e. tuning the RAM bus, CPU and GPU clocks) with libraries calls and APIs. In particular, the research group from Ferrara presented several works following precisely this line of investigation [6].

References

[1] URL `http://www.cosa-project.it/`
[2] Alfieri R, Brambilla M, De Pietri R, Di Renzo F, Feo A, Giacomini F, Manzali M, Maron G, Salomoni D, Schifano S and Tripiccione R 2014 *INFN-CNAF Annual report* ISSN 2283-5490
[3] URL `http://www.fe.infn.it/coka`
[4] URL `https://baltig.infn.it/COKA/PAPI-power-reader`
[5] Jagode H, YarKhan A, Danalis A and Dongarra J 2016 *Power Management and Event Verification in PAPI* (Springer International Publishing) pp 41–51
[6] Calore E, Schifano S F and Tripiccione R 2015 *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **9523** 737–748
[7] Morganti L, Cesini D and Ferraro A 2016 PDP Proceedings 2016 pp 541–544
[8] D C *et al.* 2017, submitted *Scientific programming*
[9] URL `http://www.hpcg-benchmark.org/index.html`
[10] Biferale L, Mantovani F, Pivanti M, Pozzati F, Sbragaglia M, Scagliarini A, Schifano S F, Toschi F and Tripiccione R 2013 *Computers & Fluids* **80** 55 – 62 ISSN 0045-7930
[11] Biferale L, Mantovani F, Pivanti M, Pozzati F, Sbragaglia M, Scagliarini A, Schifano S F, Toschi F and Tripiccione R 2013 *Computers & Fluids* **80** 55 – 62 ISSN 0045-7930

[12] Biferale L, Mantovani F, Sbragaglia M, Scagliarini A, Toschi F and Tripiccione R 2011 *Physical Review E* **84** 016305
[13] Biferale L, Mantovani F, Sbragaglia M, Scagliarini A, Toschi F and Tripiccione R 2011 *EPL* **94** 54004
[14] Paolucci P S *et al.* 2013 *arXiv:1310.1459[cs.DC]* (*Preprint* 1310.1459)
[15] Ammendola R, Biagioni A, Frezza O, Cicero F L, Lonardo A, Paolucci P S, Rossetti D, Simula F, Tosoratto L and Vicini P 2012 *Journal of Physics: Conference Series* **396** 042059 URL http://stacks.iop.org/1742-6596/396/i=4/a=042059
[16] Corni E, Morganti L, Morigi M, Brancaccio R, Bettuzzi M, Levi G, Peccenini E, Cesini D and Ferraro A 2016 *24th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing 2016* PDP Proceedings pp 369–372
[17] "the einstein toolkit" URL http://www.einsteintoolkit.org/
[18] De Pietri R, Feo A, Maione F and Leoffler F 2016 *Phys. Rev.* **D93** 064047 (*Preprint* 1509.08804)
[19] Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark A and Berendsen H 2005 *Journal of Computational Chemistry* 1701–1718
[20] Morganti L, Corni E, Ferraro A, Cesini D, D'Agostino D and Merelli I 2017, to be published PDP Proceedings 2017
[21] URL http://image-net.org/
[22] URL https://developer.nvidia.com/digits
[23] URL http://caffe.berkeleyvision.org/
[24] URL https://developer.nvidia.com/tensorrt

Open City Platform project: advances in Cloud Environment Automation and beyond

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

[1]INFN CNAF, Bologna, IT

[2]IFIN - "Horia Hulubei", Bucharest - Magurele, RO

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 and intended to research, develop and test new technological solutions open, interoperable and usable on-demand in the field of Cloud Computing. OCP is aimed at making available new innovative and sustainable organizational models for Local Public Administration (PAL) and Regional to citizens, companies and other public administrations. The IaaS layer of OCP is based on OpenStack [2], 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 research activity aimed at implementing new relevant features on our semi-automatic IaaS installation method.

1. Introduction

As already discussed in previous publications [3, 4], the Open City Platform (OCP) project 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).

Thanks to the new developments implemented in our semi-automatic installation method, already presented in [4], we have been able to implement a more complex Network architecture together with the full support to OpenStack Mitaka and CEPH Jewel functionalities both described in the next section. All the new functionalities have been implemented by maintaining the compatibility with the Ubuntu 14.04 Operating System as it is an OCP project requirement.

The paper is organized as follow. In Section 2 the advancements designed and implemented for the semi-automatic installation method are proposed and in Section 3 the future work is presented to conclude the paper.

2. New advancement in the semi-automatic installation method

To cope with the nature of the Open City Platform project focused on making available a Cloud Computing platform open source, flexible, scalable and compliant to international standards,

we started a set of activities aimed at supporting new functionalities in the semi-automatic installation method developed by us and fully described in [4].

The semi-automatic installation method and related tools, known also as OCP-Automatic tools, is leveraging on two of the most popular open source automation tools, namely Foreman [5] and Puppet [6], making use as much as possible of the official OpenStack Puppet modules [7], as well as of other community supported Puppet modules (see full description in [4]).

Such new functionalities are related to three different aspects of the semi-automatic installation model that are here described (see Figure 1):

- Full support to Openstack Mitaka version
- Full support to CEPH Jewel for the storage service
- Implement a more complex Network architecture

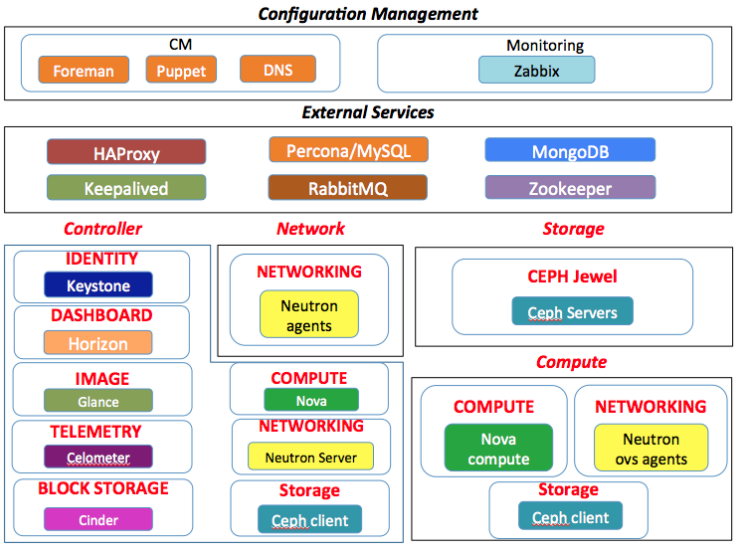


Figure 1. Overview of the OCP Infrastructure architecture services where new functionalities (in red) have been added.

2.1. Full support to OpenStack Mitaka

The current release cycle of OpenStack project and related software (6 months) make extremely difficult to stay updated to the last available version of the software. For such reason we decided to support the last stable OpenStack release, considered to be Mitaka at time of writing, included the support to the OpenStack Identity API v3 [8]. The work performed in this direction enable the entire IaaS deployed with our semi-automatic installation tool to be compliant with all the new features and services made available with the Mitaka release [9]:

- New, improved and easy to use Dashboard with a workflow-based approach for handling common tasks
- A simplified configuration for Nova compute service
- Improved Layer 3 networking and Distributed Virtual Router (DVR) support
- All-in-one client and Identity API v3 functionalities

2.2. Full support to CEPH Jewel

Looking at the new advancements of Ceph [10] as distributed storage solution, we decided to implement in the semi-automatic tools the full support to the CEPH stable version: Jewel. Ceph is a free-software open-source storage platform aimed at replicating data making it fault-tolerant using commodity hardware and requiring no specific hardware support [11].

The improved support of Ceph for the OpenStack Mitaka distribution made it eligible to be used as a storage platform. Also in this case an important work have been done to implement the installation and configuration of Ceph via the OCP semi-automatic tools. In particular, considerable time has been spent in the implementation of the new Ceph Object Gateway [12] and on the support of the related RADOSGW daemon, a FastCGI module that provides interfaces compatible with OpenStack Swift [13] and Amazon S3 [14].

2.3. Complex Network architecture implementation

Particular attention has been paid to the improvement of the Network architecture and on the deployment of the related Network services (see Figure 2). The OpenStack Mitaka Network service, in fact, is designed to run in High Availability but it can be deployed as a single service joining the Controller node if imposed by architectural needs. OCP Network currently support the configuration of five different networks as from the OpenStack documentation [15]: (i) Public network, used to provide internet access to nodes and services, (ii) Management network, used for internal communication between OpenStack Components and RHMK services, (iii) Data network, used for VM data communication within the cloud deployment, (iv) Storage Network, used to provide storage access from OpenStack components and tenants, (v) External network, used to provide VMs with Internet access in some deployment scenarios. The architecture support the implementation of multiple External networks that can be configured and deployed by setting appropriate variables.

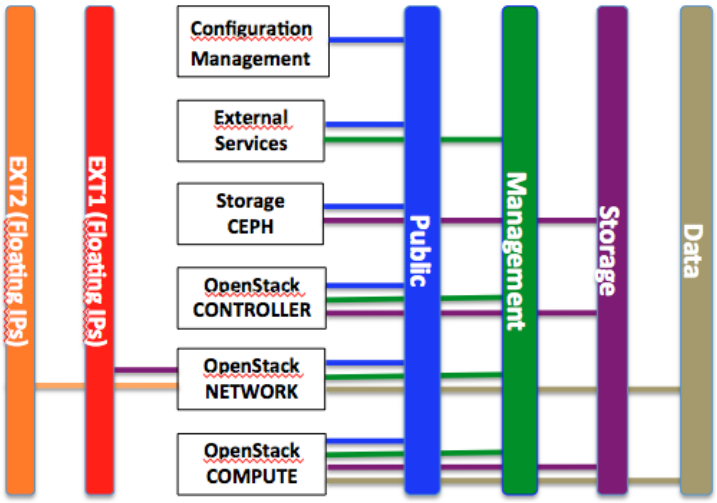


Figure 2. New OCP Network architecture and connections among nodes.

3. Future work

In the present paper advancements semi-automatic installation method for IaaS has been presented. The method fully support the most stable version of OpenStack Mitaka and Ceph Jewel, bringing to the OCP-IaaS the new features delivered by those releases.

The method still has proved to be flexible enough (i) to meet the architectural requirements dicted by the OCP project as well as the Data Center where the OpenStack Infrastructure will be deployed (ii) to easily integrate new software releases in the installation and configuration procedures.

The positive results obtained and the experience gained during the testing phase, led us to investigate an new semi-automatic procedures able to performing the upgrade of the OCP-IaaS layer currently in use by a Data Center, to a new OpenStack version.

References

[1] OpenCityPlatform Project, *http://www.opencityplatform.eu/*
[2] OpenStack, *http://www.OpenStack.org/*
[3] 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 (2015) ISSN 2283-5490
[4] C. Aiftimiei, A. Costantini, R. Bucchi, A. Italiano, D. Michelotto, M. Panella, M. Pergolesi, M.Saletta, S. Traldi, C. Vistoli, G. Zizzi and D. Salomoni: Cloud Environment Automation: from infrastructure deployment to application monitoring, INFN-CNAF Annual Report 2015, pp 129-133 (2016) ISSN 2283-5490
[5] Foreman framework, *http://theforeman.org/*
[6] Puppet, *https://puppetlabs.com/*
[7] OpenStack Puppet, *https://wiki.openstack.org/wiki/Puppet*
[8] OpenStack API v3, *http://developer.openstack.org/api-ref/identity/v3/*
[9] https://www.openstack.org/software/mitaka/
[10] Ceph storage platform, *http://ceph.com/releases/v10-2-0-jewel-released/*
[11] Ceph official documentation, *http://docs.ceph.com/docs/master/*
[12] Ceph object storage, *http://docs.ceph.com/docs/master/radosgw/*
[13] OpenStack Swift, *https://docs.openstack.org/developer/swift/*
[14] Amazon Simple Storage Service, *https://aws.amazon.com/it/s3/*
[15] OpenStack Networking documentation, *http://docs.openstack.org/security-guide/networking/architecture.html*

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	E. Corni
G. Misurelli[2]	D. Cesini	D. De Girolamo	A. Mazza	A. Falabella
S. Virgilio	E. Fattibene	F. Rosso		L. Lama
	A. Prosperini			L. Morganti
	P. Ricci			F. Noferini
				M. Tenti
				S. A. Tupputi[3]

[1] GARR employee relocated at CNAF
[2] Until 16th Feb
[3] Until 14th May

Software Development and Distributed Systems

Head: D. Salomoni

Software Development

F. Giacomini, A. Ceccanti, M. Manzali[5], E. Vianello, M. Caberletti, M. Favaro[4], E. Ronchieri

Distributed Systems

C. Aiftimiei, V. Ciaschini, D. Michelotto, S. Taneja, R. Bucchi, A. Costantini, M. Panella, G. Zizzi[6]

External Projects and Technology Transfer

Head: M. C. Vistoli

A. Ferraro

ICT

B. Martelli

National ICT Services

Head: S. Longo

S. Antonelli, M. Pezzi, F. Rosso

Information System

Head: G. Guizzunti

S. Bovina, M. Canaparo, E. Capannini, F. Capannini, S. Cattabriga[7], C. Galli

Director Office

Head: A. Marchesi

Expenditure Centralization Office[8]

Head: M. Pischedda

[4] Until 7th Oct
[5] UniFE fellowship, partially funded by CNAF
[6] Until 4th Nov
[7] Until 14th Sep
[8] The office is under the INFN Director General.

Seminars

Jan. 12^{th}	Francesco Prelz **IPv6, crash-course**
Jan. 28^{th}	Fabio Bisi, Fiorenzo Degli Esposti **Introduzione alla stampa 3D**
Feb. 19^{th}	Stefano Bovina, Diego Michelotto, Giuseppe Misurelli **Da Nagios a Sensu solo andata**
Mar. 16^{th}	Enrico Fattibene, Matteo Panella **Report sul progetto !CHAOS**
Mar. 31^{st}	Gian Piero Siroli **La guerra informatica: il lato oscuro del mondo digitale**
Apr. 7^{th}	Massimo Donatelli **Introduzione alla programmazione di Arduino**
June 7^{th}	Alessandro Costantini, Riccardo Bucchi **Report da OpenStack Summit**
June 7^{th}	Marco Caberletti **Introduzione a Systemd**
June 15^{th}	Andrea Ferraro **Il progetto OPEN-NEXT**
June 28^{th}	Francesco Giacomini **Introduzione al C++ (1/3)**
June 29^{th}	Elisa Ercolessi **La seconda rivoluzione quantistica: quantum information and computation**
June 30^{th}	Francesco Giacomini **Introduzione al C++ (2/3)**
July 1^{st}	Francesco Giacomini **Introduzione al C++ (3/3)**
July 8^{th}	Paolo Giommi **Presente e futuro del Science Data Center dell'ASI**

Sept. 21^{st} Davide Salomoni, Cristina Aiftimiei, Andrea Ceccanti
La prima release software del progetto INDIGO-DataCloud

Oct. 17^{h} Riccardo Bucchi, Giovanni Zizzi
Report da MesosCon Europe 2016

Oct. 24^{th} Tim Mattson
The future of Big Data: Polystore, specialized storage engines, and embedded analytics.

Dec. 5^{th} Marco Bertogna
High-Performance Real-Time Laboratory (UniMORE)