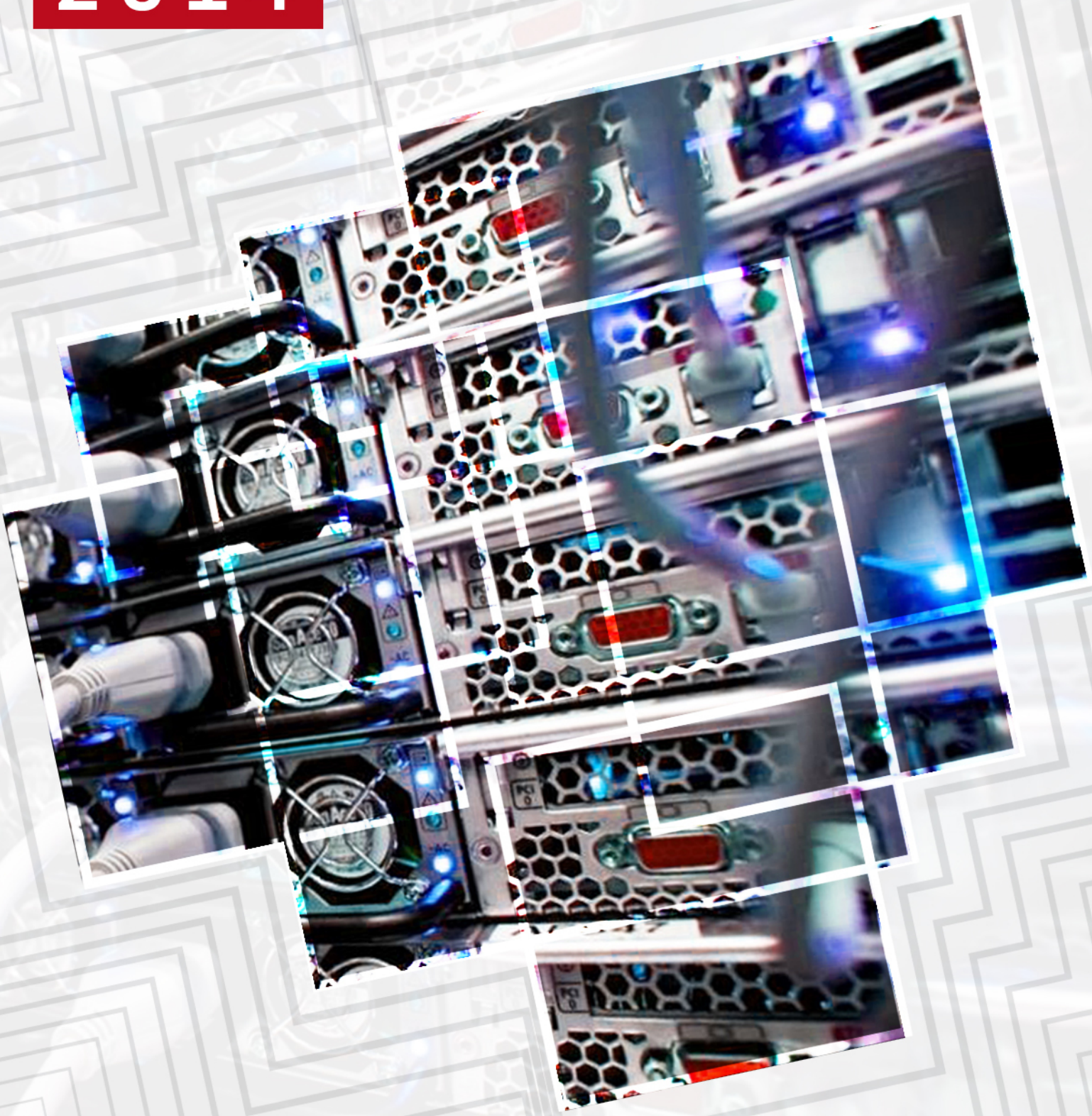
ANNUAL REPORT 2014



INFN CNAF

INFN-CNAF Annual Report 2014

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

Editors

Luca dell'Agnello
Francesco Giacomini
Lucia Morganti

Cover Design

Francesca Cuicchio

Address

INFN CNAF
Viale Berti Pichat, 6/2
I-40127 Bologna
Tel. +39 051 2095 475, Fax +39 051 2095 477
www.cnaf.infn.it

Contents

Scientific Exploitation of CNAF ICT Resources

The INFN-Tier1 Center and National ICT Services

Software Services and Distributed Systems

Knowledge Transfer

Additional Information

Introduction

During 2014 CNAF continued to pursue its main mission, namely supporting the scientific computing of the INFN activities, supporting the World-Wide LHC Computing Grid (WLCG) e-infrastructure, addressing the developments towards new distributed systems paradigms (e.g. Cloud) and developing software modules targeted to INFN experiments. A significant part of the activity has also been devoted to attracting external funds (basically proposing or participating to new H2020 projects) and to transferring knowledge to the industrial and public administration worlds.

The Tier-1 Data Center ran smoothly during the year, apart from an incident at one of our chillers that forced a one-day unexpected shutdown of the center, yet staying well inside the availability and reliability metrics requested to an LHC Tier-1 class data center. The year 2014 confirmed the previous years' trend of a steady increase in the number of particle and astroparticle physics experiments using the CNAF facilities; four new experiments (NA62, PANDA, CTA, Darkside and CUORE) in fact decided to rely on our center for their data storage and processing. Nevertheless the utilization ratio of the computing resources by the LHC experiments compared to other experiments remained at 75%.

All experimental groups present at CNAF are assisted by the User Support team, which is skilled both in proposing computing models for experiments and in the management of the daily operations. The team represents a well-defined interface between the users and the internal organization, from the operators of the data center to the Grid middleware developers.

During 2014 the computing power available at the Tier-1 reached 136000 HS06, the fast storage exceeded the capacity of 15 PB and the long-term storage the capacity of 24 PB. Some important objectives were achieved during the year: a thorough investigation into low-power CPUs; the introduction of the concept of "dynamic partitioning" in the CNAF batch system, to provide an acceptable resource-sharing mechanism also for jobs requiring a multicore environment; significant improvements to the monitoring and reporting system of the main operational parameters of the center. Important achievements were also accomplished in the domain of services of national utility for INFN, such as a highly-available DNS architecture, the introduction of an official INFN document management system, a Dropbox-like service and a disaster recovery system for the INFN Information System.

The Software Development and Distributed Systems (SDDS) group was involved in significant projects based on the OpenStack open source software to create private and public clouds. The Cloud@CNAF initiative has been the first example of this, a pilot Cloud opened to INFN internal users. !CHAOS and Open City Platform (OCP) are the other two projects present at CNAF and based on OpenStack, in this case funded by MIUR. The LHCb and KM3NeT experiments have profited of the skills available in the group to develop key parts of their data acquisition and trigger system software. The group is also involved in the management of the Grid-WLCG infrastructure and in the maintenance of some critical Grid middleware services, namely VOMS, StoRM and Argus.

Within the European H2020 framework program, CNAF participated to the 2014 calls related to the areas of development more suited for our center, especially on distributed systems (Grid and Cloud) and on infrastructures for our communities, but also on the development of new IT technology. Four projects were submitted: INDIGO Data Cloud (EINFRA-1-2014), on the development of a data and computing platform targeting scientific communities and provisioned over hybrid (private and public) e-infrastructures; ASTERICS (INFRADEV-4-2014-2015), to implement and operate cross-cutting services for a cluster of ESFRI projects focused on astronomy and astroparticle physics; ExaNeSt (FETHPC-1-2015), with the aim to develop and prototype solutions for some of the crucial problems on the way towards the production of exascale-level supercomputers; and EGI-Engage (EINFRA-1-2014), to engage

the Research Community towards an Open Science Commons. At the time of writing this annual report, we know that all four projects have been approved. The resources granted by these projects will allow CNAF to continue its development activities in the various fields of interest for the coming years.

Gaetano Maron
CNAF Director

Giulia Vita Finzi (1956–2015)

After a short illness, our friend and colleague Giulia Vita Finzi passed away on March 23^{th}, 2015. We have lost a very special colleague, always enthusiastically present in all initiatives that have made CNAF an important centre, at a national and international level.

Giulia joined CNAF as a scientific secretary. Later she became involved in the evolution of the networking infrastructure and in the founding of the ICT national services, historically an important asset for the whole INFN: the first INFN web site, the news system, the IXI network. Then came the experience with the CNTC (Commission for New Computing Technologies) and the participation to the development of the Grid, with a specific focus on the Information System and the management of X.509 certificates. Her last committment was the responsibility of the national service that guarantees proper technical support for hardware and software acquisitions. Her collaboration has been constant, constructive and precious.

For many of us she was also a dear friend, present in happy moments as well as in sad or difficult situations. And she always had a word of optimism and encouragement. Her great energy was contageous.

Working at CNAF has never been easy, but Giulia knew how to make it a little bit easier.

Scientific Exploitation of CNAF ICT Resources

The User Support unit at CNAF

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

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

1. Current status

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

The main activities of the team include:

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

Apart from these operational activities, the User Support staff is also involved in different projects of the Tier-1 site.

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

2. Supported experiments and resource usage

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

Recently, contacts have been taken for the support of new experiments such as KM3NET/NEMO and LHAASO.

Moreover, research groups belonging to the Computational Chemistry and Biomedical domains also access the center through Grid services.

The following figures show CPU usage (Figure 1), disk usage (Figure 2 and 3) and tape usage (Figure 4) by the supported experiments during 2014. The LHC experiments represent more than three quarters of the total resources funded at CNAF, but many other research groups use CNAF resources significantly.

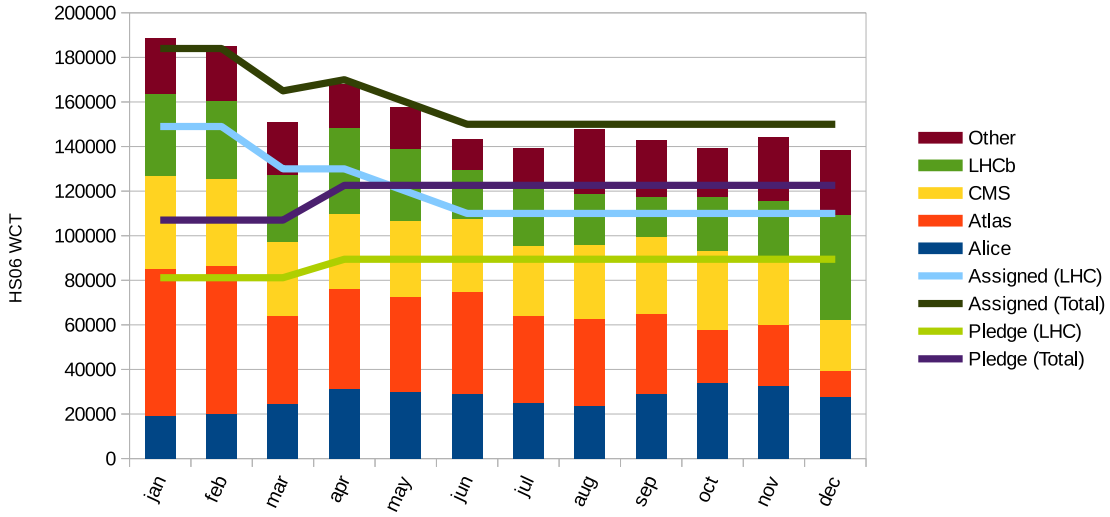


Figure 1. Average monthly CPU usage (HS06) during 2014. Non-LHC VOs are grouped together (*Other*). Lines show pledged and assigned resources for LHC experiments alone (*LHC*) and for LHC and non-LHC experiments together (*Total*). It is apparent that a large part of the overpledge was switched off in March.

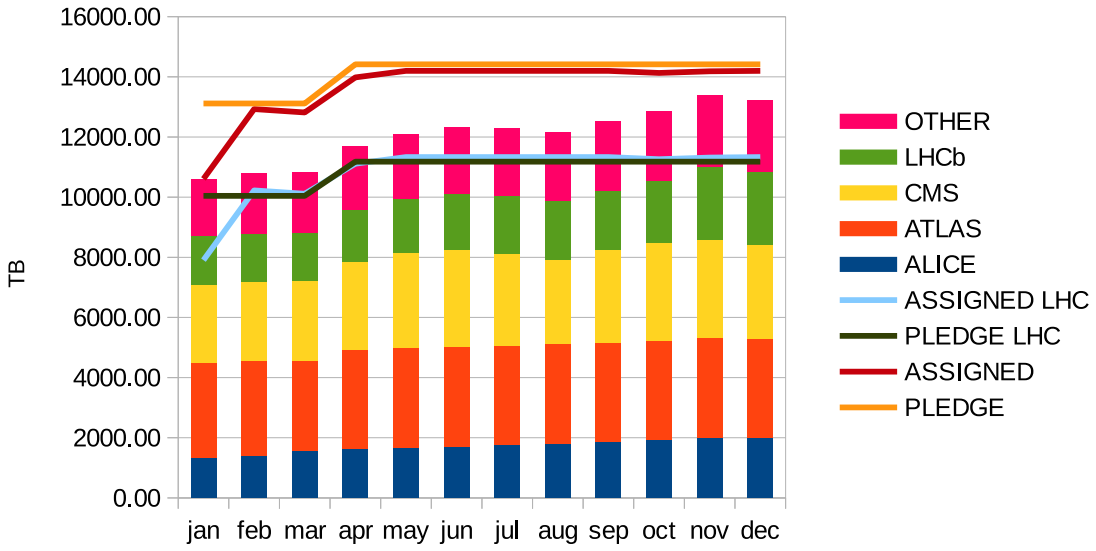


Figure 2. Disk usage (TB) for all the supported experiments during 2014. The non-LHC VOs are grouped together (*Other*). Lines show pledged and assigned resources.

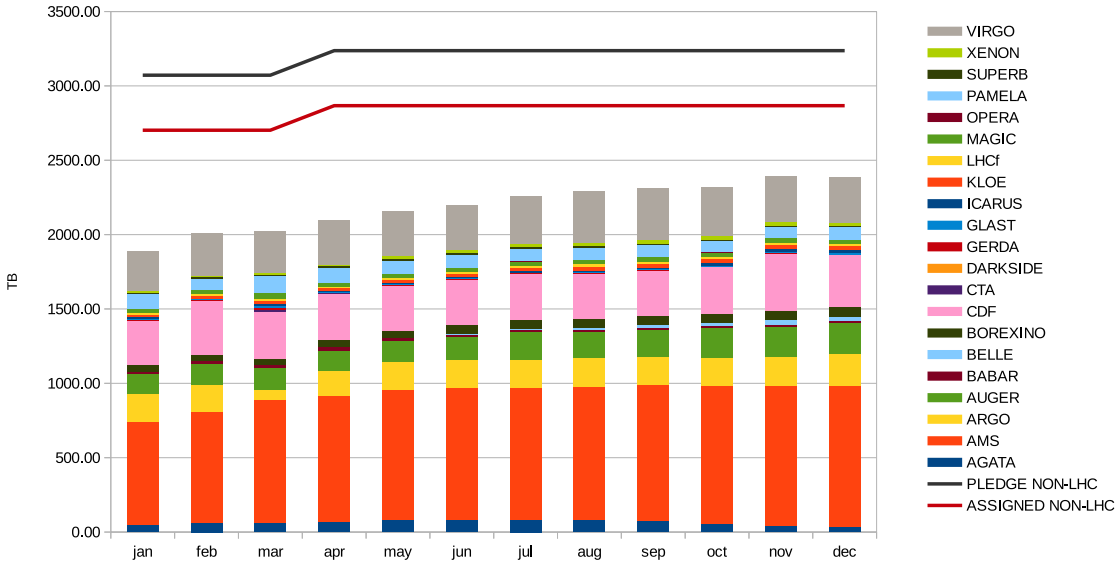


Figure 3. Disk usage (TB) for all the non-LHC experiments during 2014. Lines show pledged and assigned resources.

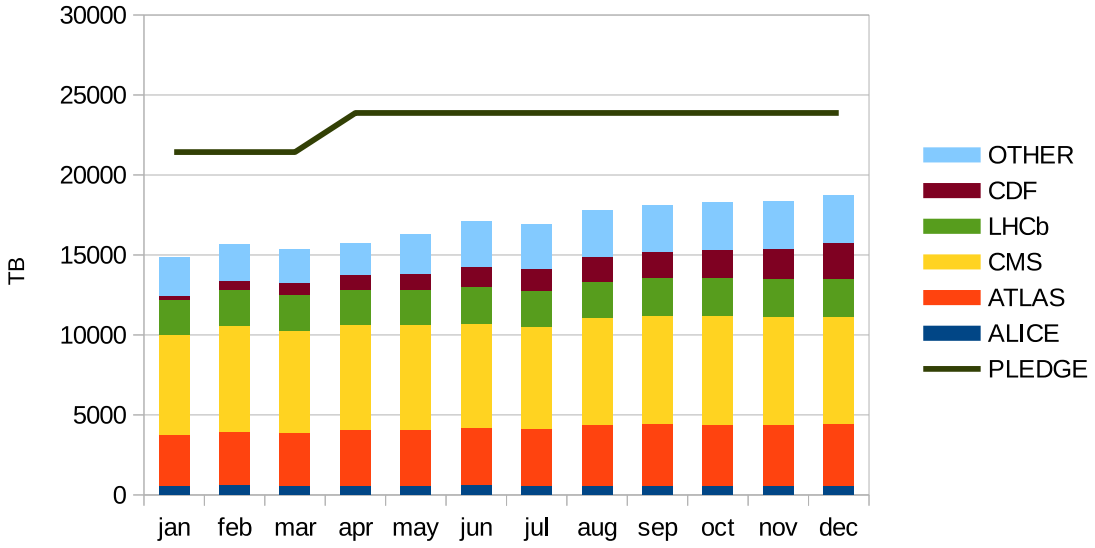


Figure 4. Tape usage (TB) for all the VOs during 2014. The dark line represents pledged resources. The increasing tape usage for the CDF Long Term Data Preservation activities is visible.

ALICE – A Large Ion Collider Experiment

Stefano Bagnasco[1], Domenico Elia[2], Stefano Piano[3]

[1] INFN Sezione di Torino
[2] INFN Sezione di Bari
[3] INFN Sezione di Trieste

Abstract. ALICE (A Large Ion Collider Experiment) is one of the four large CERN LHC experiments, specifically designed to exploit the heavy-ion special runs of the LHC to study the physics of strongly interacting matter and QGP. Since its beginning the ALICE Computing Model relied heavily on distributed computing; CNAF provided over the years a stable and reliable resource. In 2014 it provided 6.9% of the total CPU hours used by the collaboration for raw data reconstruction, Montecarlo simulation and analysis; it also hosts, on disk and tape, more than 2PB of experiment data.

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).

Its configuration has been upgraded 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 period of LHC. Other detectors were also upgraded or overhauled, with interventions like for example the installation of some extra or missing modules in the TRD and the PHOS, the refitting of the TPC with a different gas mixture and a new redesigned readout electronics, and many more. 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. Main physics results

While finalizing consolidation activities and new installations at the experiment, during 2014 the Collaboration has been continuing the analysis of the data taken in Run1. Many new physics results

have been obtained from the study of pp, p-Pb and Pb-Pb collisions, leading to 24 publications (109 in total since the start of the LHC operations) and about 100 conference presentations. In particular, no cold nuclear matter effects have been measured in p-Pb collisions [2], while several signals of collective effects have been unexpectedly observed in collisions of smaller systems (pp and p-Pb) [3], triggering a considerable interest among the theorists.

Among the other results extracted from the study of p-Pb collisions, the earlier observation of a long-range double structure in rapidity ("double-ridge") [4] has been complemented with the measurement of higher-order cumulants of the azimuthal correlations. Results indicate that the double-ridge arises from global (as opposed to few-particle) correlations [5]. Intriguing results have been also obtained from the study of pion Bose-Einstein correlations in p-Pb: at a given multiplicity value, the emission radii are found larger than in pp collisions, but not as large as those found in the Pb-Pb case [6].

Several other remarkable results have been extracted from the analysis of the heavy-ion collisions. Among them, the study of jet-quenching in Pb-Pb has been extended with the measurement of fully reconstructed charged jets down to transverse momenta of 30-40 GeV/c, where a suppression factor by as much as a factor three is observed with respect to expectation from pp (scaled by the number of binary nucleon-nucleon collisions) [7]. The measurement of the suppression of the Υ(1S) at forward rapidity in Pb-Pb has been also finalised [8]: this study has revealed a peculiar rapidity dependence pattern with larger suppression at higher rapidity, not reproduced by the current theoretical models. Finally, the study of strange and multi-strange baryon production in Pb-Pb collisions has been also completed: transverse momentum spectra agree reasonably well with predictions from hydrodynamics, while the enhancements relative to pp increase both with the strangeness content of the baryon and with centrality, but are less pronounced than at lower energies [9].

3. Computing model

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 [10]. According to this principle, the ALICE peculiarity has always been to operate its Grid as a "cloud" of computing resources (both CPU and storage) with no specific role assigned to any given centre, the only difference between them being the Tier to which they belong. All resources are to be made available to all ALICE members, according only to experiment policy and not on resource physical location, and data is distributed according to network topology and availability of resources and not in pre-defined datasets.

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

The Computing Model will not change significantly for Run2, except for scavenging of some extra computing power by opportunistically use the HLT farm when not needed for data taking. The much higher data rate foreseen for Run3, tough, will require a major rethinking of it in all its components, from the software framework to the algorithms to the distributed computing infrastructure. The proposed new design is mainly based on the concepts of Online-Offline integration and Cloud computing, and the O^2 Project has recently delivered to the LHCC its Technical Design report [11].

In Italy, computing R&D activities has been dominated by the STOA-LHC PRIN project, aiming at the deployment of federated Virtual Analysis Facilities leveraging on Cloud Computing technologies and the federation tools provided by PROOF (see for example [12]). 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 Torino, Bari, Catania and Padova/LNL, plus some extra resources in Cagliari, Bologna and Trieste.

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. Even though 2014 saw no LHC data taking, ongoing analyses, raw data reprocessing and Montecarlo simulation productions kept the ALICE computing resources, and CNAF among those, fully occupied.

Running at CNAF in 2014 has been remarkably stable: for example, both the disk and tae 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 during last year, matching and often exceeding the pledged amounts allowing access to resources unused by other collaborations. Overall, about 69% of the ALICE computing activity was Montecarlo simulation, 3% raw data processing (which takes place at the Tier-0 and Tier-1 centres only), 28% analysis activities.

In order to optimize the use of resources and enhance the "CPU efficiency" (the ratio of CPU to Wall Clock times), an effort was started in 2011 to move the analysis tasks from user-submitted "chaotic" jobs to organized, centrally managed "analysis trains". The effort went on in 2013 and 2014 with relative increases of the number of train jobs respectively of 47% (2013 over 2012) and 32% (2014 over 2013), whereas the number of individual analysis jobs was halved in 2013 and remained essentially stable in 2014. This leads up to a split of analysis activities, in terms of CPU hours, between 43% individual jobs, 57% organized trains (12% and 16% of the total, respectively).

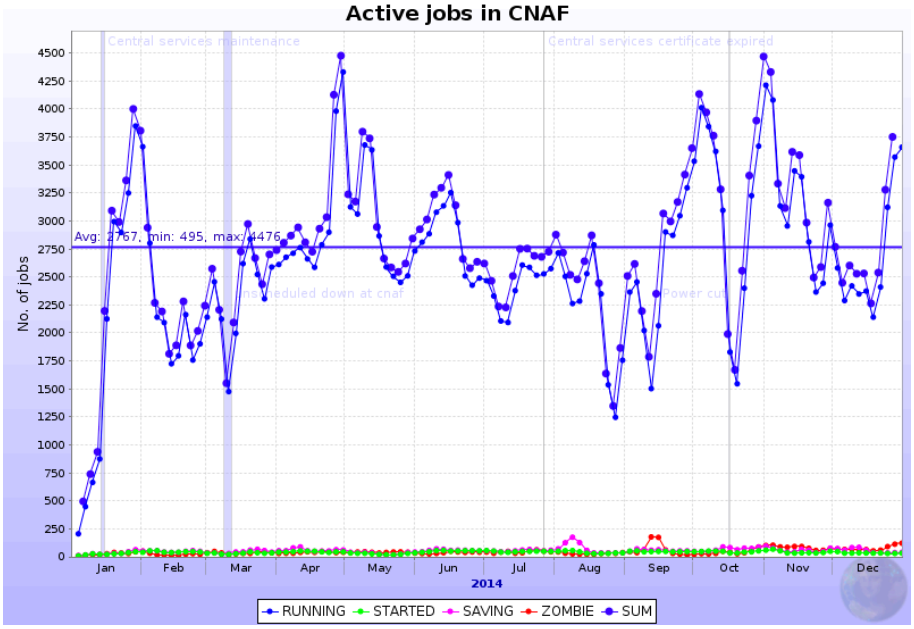


Figure 1: running jobs profile at CNAF in 2014.

In 2014, CNAF provided 6.9% of the total CPU hours used by ALICE, thus ranking second of the ALICE Tier-1 sites, following only FZK in Karlsruhe. This amounts to about 45% of the total INFN contribution: it successfully completed nearly 8.5 million jobs, for a total of more than 19 millions CPU hours.

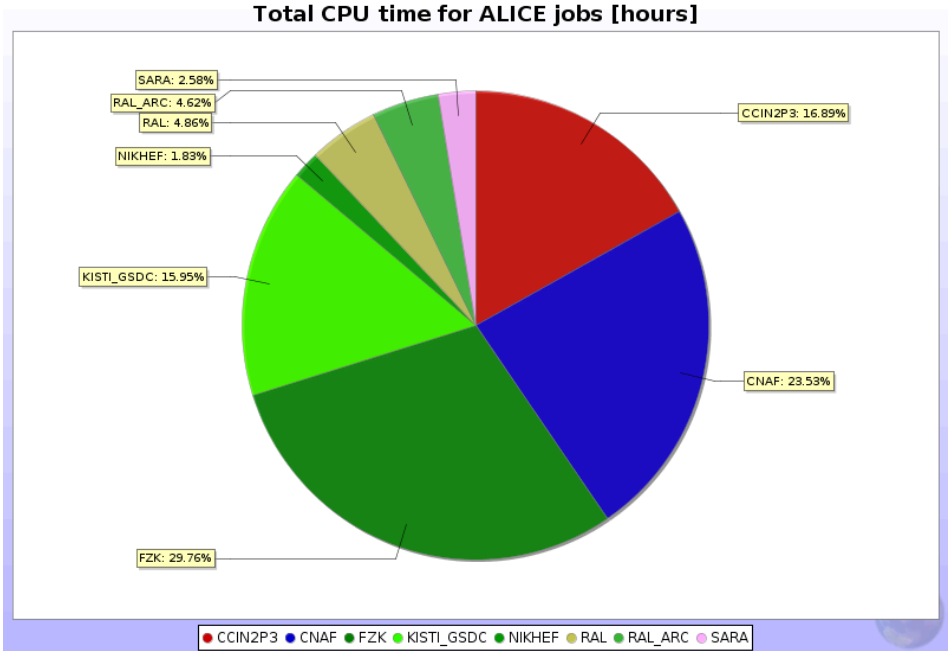


Figure 2: ranking of CNAF among ALICE Tier-1 centres in 2014

ALICE keeps on disk at CNAF about 1.4 PB of data in nearly 33 million files, plus about 700 TB of raw data on custodial tape storage; the reliability of the storage infrastructure is commendable, even taking into account the extra layer of complexity introduced by the xrootd interfaces. Also network connectivity has always been reliable; the 40Gb/s of the WAN links makes CNAF one of the better-connected sites in the ALICE Computing Grid.

References

[1] B. Abelev *et al.* (ALICE Collaboration), Int. J. Mod. Phys. A **29** 1430044 (2014).
[2] B. Abelev *et al.* (ALICE Collaboration), Eur. Phys. J. C **74** 3054 (2014).
[3] B. Abelev *et al.* (ALICE Collaboration), Physics Letters B **728** 25-38 (2014).
[4] B. Abelev *et al.* (ALICE Collaboration), Physics Letters B **719** 29-41 (2013).
[5] B. Abelev *et al.* (ALICE Collaboration), Phys. Rev. C **90** 054901 (2014).
[6] B. Abelev *et al.* (ALICE Collaboration), Physics Letters B **739** 139-151 (2014).
[7] B. Abelev *et al.* (ALICE Collaboration), J. High Energy Phys. **03** 013 (2014).
[8] B. Abelev *et al.* (ALICE Collaboration), Physics Letters B **738** 361-372 (2014).
[9] B. Abelev *et al.* (ALICE Collaboration), Physics Letters B **728** 216-227 (2014).
[10] P. Cortese *et al.* (ALICE Collaboration), CERN-LHCC-2005-018 (2005).
[11] J. Adam et al. (ALICE Collaboration), CERN-LHCC-2015-006 (2015).
[12] S. Bagnasco *et al.*, "Interoperating Cloud-based Virtual Farms", talk presented at the 21st International Conference on Computing in High Energy and Nuclear Physics (CHEP2015), Okinawa, Japan, April 13-17, 2015.

AMS data processing and analysis at CNAF

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

[1] Università di Perugia, I-06100 Perugia, Italy
[2] INFN, Sezione Perugia, I-06100 Perugia, Italy
[3] ASDC, I-00133 Roma, Italy
[4] INFN Sezione di Milano-Bicocca, I-20126 Milano, Italy
[5] CINECA, I-20090 Segrate, Milano, Italy
AMS experiment http://ams.cern.ch,http://www.ams02.org

E-mail: * bruna.bertucci@pg.infn.it

1. Introduction

AMS is a large acceptance instrument conceived to search for anti-particles (positrons, anti-protons, anti-deutons) coming from dark matter annihilations, 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.

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 microstrip detectors - the Tracker - with a spatial resolution of $\sim 10\mu m$ in the single point measurement along the track. Redundant measurements of the particles' 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's operating continuously since then, with a collected statistics of $\sim$ 60 billion events up to the end of 2014. 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 of $\sim$10 Mbit/s for the data transmission from space to ground, for a $\sim$ 100 GByte/day of raw data produced by the experiment.

Due to the rapidly changing environmental conditions along the $\sim$ 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. The first year of AMS data analysis has been mainly devoted to the development of calibration procedures which allow to achieve during space operations the same detector performances as measured during beam tests on ground. In 2013 the first physics publication was delivered, with a precision measurement of the positron fraction up to 350 GeV [1], followed in 2014 by three publications which extended the positron fraction measurement up to 500 GeV [2], gave the positron and electron flux measurements up to 500 e 700 GeV respectively [3] and the combined electron plus positron flux measurements [4] up to 1 TeV.

CNAF is one of the repositories of the full AMS data set, both raw and processed data are stored at CNAF which represents the central computing resource for the data analysis performed by Italian collaboration and contributes as well at the data production and MonteCarlo simulation in the international collaboration. In the following, the details on the AMS computing framework, data transfer to/from CNAF and use of the CNAF resources in 2014 will be given.

2. AMS Data structure 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. AMS data are organized in a custom format called AMSBlocks. AMSBlocks can assume many forms, such as special blocks for science data, or envelope blocks containing a set of sub blocks. Hence their size is not fixed.

To be transmitted on the NASA interfaces, AMSBlocks have to be divided into smaller pieces of fixed size: the frames. Framing process is the analogue of the one adopted in the normal ethernet networks. In this case, frame format is defined by the Consultative Committee for Space Data System (CCSDS), hence the name CCSDS frames.

AMS detector produce AMSBlocks that are split in pieces, one minute frames. Frames are transferred via NASA communication satellites to Marshall Space Flight Center in Alabama, through the White Sands Center antennas in New Mexico. From Marshall data are finally sent to the AMS Payload Operation Control Center (POCC) at CERN. At CERN, frames joined to rebuild up the AMSBlocks. The deframing process is similar to what happens to form IP packets when ethernet frames reach the network stack of an operating system. Both frames and AMSBlocks are saved to disk. AMSBlocks are then merged to build up runs (raw AMS data). Each run lasts almost 23 minutes, one quarter of the ISS orbit. Finally, raw data are reconstructed and a ROOT format is produced.

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 subdetectors 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 subdetector calibrations, alignments, ancillary data from ISS and slow control data to produce reconstructed data for the physics analysis. This **Second production** step is usually applied in an incremental way to the std data sample, every 3-6 months - the time needed to produce and certify the calibrations. A full reprocessing of all AMS data is carried out periodically in case of major software major updates, providing the so called "pass" production. Up to 2014 there were 5 full data reproductions done, published measurements were based on the pass4 data set.

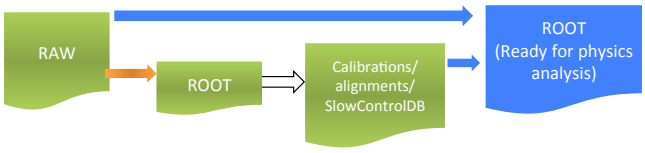


Figure 1. AMS02 Processing Data Flow

The First Production is processed on a local AMS farm at CERN 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).

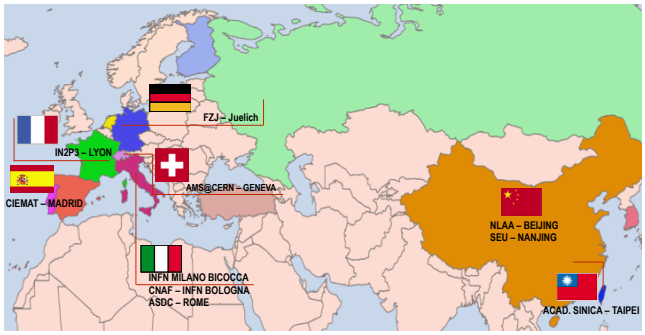


Figure 2. AMS02 Major Contributors to Computing Resources.

China and Taiwan centers are mostly devoted to Monte Carlo production, while CERN, CNAF and FTZ 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.

CNAF is the main computing resource for data analysis of the AMS Italian collaboration: a full copy of the AMS raw data is preserved on tape, the latest PASSxx production and part of the MonteCarlo sample are available on disk. A pool of ~ 30 italian users is routinely performing the bulk of their analysis at CNAF, transferring then to local sites (Rome, ASDC, Perugia, Pisa, Bologna, Milano Bicocca, Trento) just the final histograms and reduced data sets.

3. AMS02 Data Flow

AMS02 Data flow can be summarized in the following way: data are both recorded on board of the International Space Station (ISS) on AMS Laptop and directly transmitted, by means of satellites, to Marshall Space Flight Center (MSFC) in Alabama and therefrom, over the internet, to the Payload Operation Control Center (POCC). The CHD (critical health data) data are directly transmitted to the POCC in order to have an immediate overview of the detector status. The POCC is the control center of the AMS flight operations and where online data monitoring takes place to give a first look and evaluate the quality of data. From there, data will be sent to the Science Operation Center (SOC) where they will be processed and analyzed.

Finally data will be distributed from the SOC to Regional Sites that will act as storage of AMS Data samples and as Montecarlo production facilities to help the work of studying the detector response done by a simulation of real data at the SOC.

In this framework, a Data Transfer software has been setup and is maintained within the collaboration by the Milano Bicocca group, to efficiently transfer data from (to) CERN to (from) Regional Sites, and in particular CNAF, without interfering with SOC activities and keeping track of what has been moved and if successfully or not. The main core of the DT is a Multi-threaded finite state automa (written in Python) and the state transition jobs are written in Perl. It uses a database (Mysql/Oracle) for book-keeping and it relies on GRID's file transfer protocols [5].

The Data Transfer operates continuously since 2011 transfering data from CERN to CNAF, much in the fashion of a T2 of LHC's experiments, and from CNAF to CERN when the CNAF

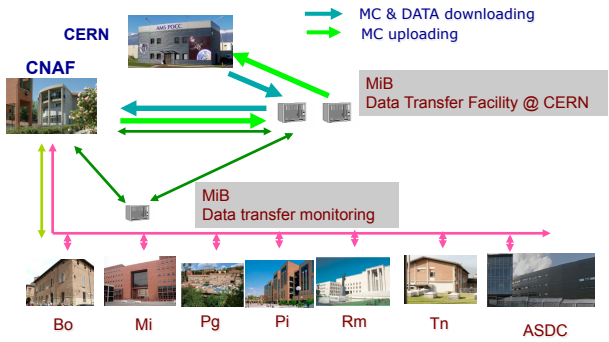


Figure 3. AMS02 Data from CERN to Italian users

resources are used in the MC production for the international collaboration. Thanks to the direct *srm* to *srm* protocol, 1.2Gbit/s throughput performance is achieved. In Table 1 is reported the amount of data moved between CERN and CNAF since the start of data taking in mid 2011.

Table 1. Data sets

Year	Data Set	Num. Files	Size (TB)
2011			
	`Reconstructed`	56102	65.1
	`MonteCarlo`	105480	66.5
	`RAW`	199747	23.2
	`DAQ`	890149	24.1
2012			
	`Reconstructed`	326463	755.6
	`MonteCarlo`	120337	70.8
	`RAW`	435851	45.3
	`DAQ`	1480325	36.4
2013			
	`Reconstructed`	38208	180.7
	`MonteCarlo`	379373	153.7
	`RAW`	275217	41.2
	`DAQ`	1879566	34.6
2014			
	`Reconstructed`	8260	55.5
	`MonteCarlo`	69700	14.5
	`RAW`	98800	39.4
	`DAQ`	1601080	39.1
Total			
	`Reconstructed`	429034	1056.8
	`MonteCarlo`	674882	305.5
	`RAW`	1009614	149.1
	`DAQ`	5851120	134.2

Remote Data Access Once data have been copied into the data repository at CNAF, they have to be accessible from all of italian institution involved in AMS. Most users are locally running at CNAF, directly accessing the AMS data set on disk, however, a local farm is available in ASDC for end point analysis, with reduced disk space but relevant CPU resources. In a such scenario, remote access to data constitutes one of the major challenges.

In 2014, 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) in Rome. The solution is based on the use of the General Parallel File System (GPFS) and of the Tivoli Storage Manager (TSM). Both products are developed by IBM. Owing to a new feature introduced in GPFS 3.5, so-called Active File Management (AFM), it is possible to define a single, geographically-distributed namespace, characterized by automated data flow management between different locations. This solution was developed in cooperation between CNAF staff and AMS physicists.

A database (based on ROOT TTree objects) with tags of events that have passed certain preselection requirements has been locally created in ASDC. Each data processing job at ASDC queries the preselection database to look for the tags of interesting events, in order to access them (and only them) from a remote file. In this scheme, AFM Prefetch Threshold has been tuned to manage 10 GB files (average size of AMS run file) accessed randomly. The configuration allows to process the same file remotely paying only a fraction of 15% in execution time.

4. Activities in 2014

AMS activities at CNAF in 2014 have been mainly related to the data analysis based on the PASS4 reconstruction, which covers the data taking period May 2011-Nov.2013, and Monte Carlo production in a coordinated manner with the other regional centres of the Collaboration. No major effort on a new data production has been performed at CNAF in 2014, since most of the PASS4 reconstruction - which has been used for the publications [2, 3, 4] was performed in 2013 and just few months (July-November 2013) have been processed during 2014. Only a test reproduction on few months of the AMS data, the PASS5, was performed by the international collaboration and CNAF resources were only partially used for this in November.

Out of the 7kHS06 pledged to AMS in 2014, 10.6 kHS06 have been used for these activities with a regular profile along the year. In Fig.4 the number of jobs along the year are reported for the different AMS queues. Two local queues are available for the $\sim$ 30 AMS users: the default running is on the AMS queue, with a CPU limit of 3300 minutes and a maximum of 600 job running simultaneously, where as for test runs the AMS_short, with high priority but a CPU limit of 360 minutes and a max 100 jobs running limit is used. For data reprocessing or MC production the AMS_prod queue, with a CPU limit of 5760 minutes and 2000 jobs limit, is available and accessible only to data production team of the international collaboration and few experts users of the italian team. In fact, the AMS_prod queue is used within the data analysis process to produce data streams of pre-selected events and lightweight data files with a custom format, 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.

The disk resources pledged in 2014, $\sim 1PB$, were mostly devoted to the PASS4/PASS5 data sample ($\sim$500 TB), MC data sample ($\sim$ 200 TB), selected data streams ($\sim$ 30 TB of pre-selected data used for common electron/positron analysis) and scratch area for users.

Different analysis are carried on by the Italian collaboration, in 2014 most of the CNAF resources where devoted to the electron/positron analysis, both in terms of flux measurement and anysotropies, the evaluation of the geomagnetic cutoff for all the on-going analyses in the collaboration and the study of time dependence of electron/positron and proton fluxes.

Figure 4. Usage of AMS queues in 2014. AMS (top left), AMS_prod top right), AMS_short (bottom)

References

[1] M.Aguilar *et al.*, AMS-02 Collaboration, Phys.Rev. Lett,110 (2013) ,141102.1-10
[2] L.Accardo *et al.*, AMS-02 Collaboration, Phys.Rev. Lett,113 (2014) ,121101.1-9
[3] M.Aguilar *et al.*, AMS-02 Collaboration, Phys.Rev. Lett,110 (2014) ,121102.1-9
[4] M.Aguilar *et al.*, AMS-02 Collaboration, Phys.Rev. Lett,110 (2014) ,221102.1-7
[5] M. Boschini et al., New generation Data Transfer for AMS02, 10th International Conference, ICATPP 2005: Astroparticle, Particle, Space Physics, Detectors and Medical Physics Applications, World Scientific, 2008.

ATLAS activities

A De Salvo[1]

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

Abstract. In this paper we describe the computing activities of the ATLAS experiment at LHC, CERN, in relation to the Italian Tier-1 located at CNAF, Bologna. The major achievements in terms of computing 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.

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

[1] INFN, Sez. Roma1

Its most notable result up to now has been the discovery of a new resonance at a mass of about 125 GeV, followed by the measurement of its properties (mass, production cross sections in various channels and couplings). These measurements have confirmed the compatibility of the new resonance with the Higgs boson, foreseen by the SM but never observed before.

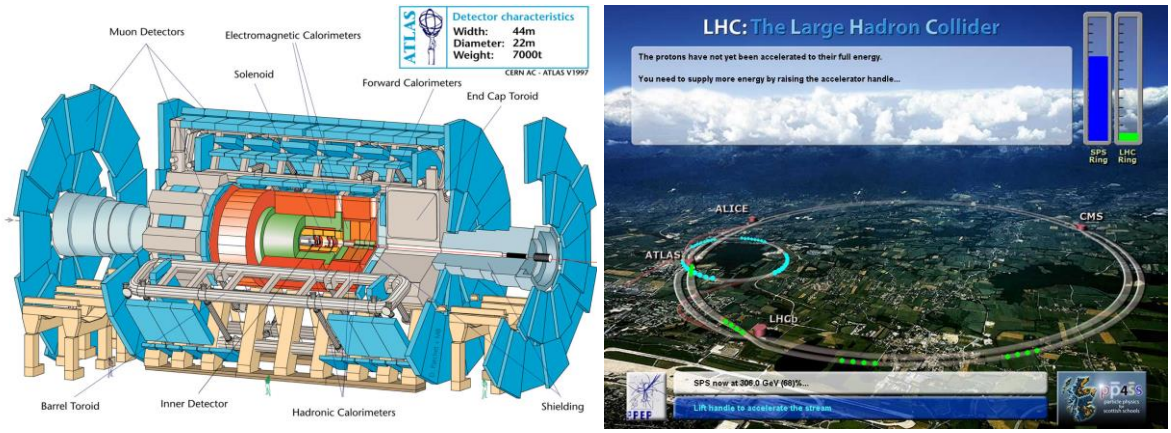


Figure 1 - The ATLAS experiment at LHC

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 an higher importance with respect to normal Tier-2s and can have primary data too. The total of the T2 sites corresponds to more than the total ATLAS size at the T1, for what concerns disk and CPUs; tape is not available in the T2 sites.
A third category of sites is the so-called Tier-3 centres. Those are smaller centres, scattered in different places in Italy, that nevertheless contributes in a consistent way to the overall computing power, in terms of disk and CPUs. The overall size of the Tier-3 sites corresponds roughly to the size of a Tier-2 site. The Tier-1 and Tier-2 sites have pledged resources, while the Tier-3 sites do not have any pledge resource available.
In terms of pledged resources, Italy contributes to the ATLAS computing as 9% of both CPU and disk for the Tier-1. The share of the T2 facilities corresponds to 7% of disk and 9% of CPU of the whole ATLAS computing infrastructure.
The Italian Tier-1, together with the other Italian centres, provides both resources and expertise to the ATLAS computing community, and manages the so-called Italian Cloud of computing. Up to 2014 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 2014.
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 2014, for some specific activities:

1) test and fine tuning of the Xrootd federation using the StoRM storage system, completely developed by CNAF within the LCG and related projects, funded by EU;
2) 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

In 2014 the Italian centers have mainly contributed to the upgrade of the Computing Model both from the testing side and the development of specific working groups, as described later.
Several improvements in the Computing Model has been achieved in 2014, more precisely in the software domain and the infrastructure.
The software have been improved to cope with the new data of the run 2, reaching a speed factor of 3 with respect to the old software used for the previous run. The fast simulation is expected to be released in the next months, leading to an additional improvement in the global chain.
On the infrastructure side, all the main subsystems have been redesigned, namely the Production System and the Data Management System (Rucio). The improvements are both for performance and flexibility, as well as efficiency. The users have seen a smooth transition from the old scenario to the new infrastructure, and the workflow of the analysis has not been stopped or heavily perturbed in the transition phases.
On the network side, the increase in performance of the data backbones has been also helped a lot by the new generation of network connections and the dedicated overlay network LHCONE, designed to connect the Tier2 sites of LHC. Already since mid 2012 almost all the Tier2 centers have been connected at >= 10 Gbps via the NREN networks, while the T1s were already connected at 10 Gbps via the dedicated infrastructure LHCOPN. The Italian T1 has been upgraded to a bandwidth of 40 Gbps, allowing for better performance and increased capacity, suitable for direct WAN access and similar techniques, very important, for example, in the storage federation domains.
In order to improve the reliability and efficiency of the whole system, ATLAS introduced in 2013 the so-called Federation of Xrootd storage systems (FAX), on top of the existing infrastructure. FAX has now reached its full deployment, reaching more than 90% of the total number of sites. Using FAX, the users now have the possibility to access remote files via the XRootd protocol in a transparent way, using a global namespace and a hierarchy of redirectors, thus reducing the number of failures due to missing or not accessible local files, while also giving the possibility to relax the data management and storage requirements in the sites. The testing of the FAX federation started in mid 2012 and Italy joined it with 3 Tier2 sites in November 2012. At the moment FAX is in pre-production mode.
The contribution of the Italian sites in to the computing activities of 2014, 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.
In addition to those activities, the Italian Community contributed in the upgrade of the ATLAS Computing Model for what concerns the data selection (Event Index), the data access (rooted via Proof) and the storage systems (testing of DPM facilities and SAM monitoring). The Italian Computing team is involved in various groups, and in particular the Database, the VO management, the grid software installation, the monitoring, the network and the security team areas.
Besides the Tier1 and Tier2s, in 2014 also the Tier3s gave a significant contribution to the Italian physicists community for the data analysis. The Tier3s are local farms dedicated to the interactive data analysis, the last step of the analysis workflow, and to the grid analysis over small data sample. Many Italian groups set up a farm for such a purpose in their universities and, after a testing and validation process performed by the distributed computing team of the collaboration, all have been recognized as official Tier3s of the collaboration.

5. References

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

Pierre Auger Observatory Data Simulation and Analysis at CNAF.

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

[1] Istituto Nazionale Fisica Nucleare, sezione di Lecce, Italy.
[2] Observatorio Pierre Auger, Av. San Martìn Norte 304, 5613 Malargüe, Argentina
(Full author list : http://www.auger.org/archive/authors_2014_12.html)

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

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

1. Introduction

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

2. Organization of the Auger analysis.

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

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

3. Organization of the Italian Auger Computing

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

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

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

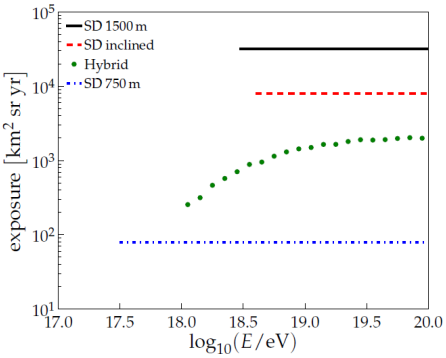


Figure 1. The integrated exposure of the different detectors at the Pierre Auger Observatory as a function of energy. The SD exposure in the three cases is at above the energy corresponding to full trigger efficiency for the surface arrays.

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

Unfortunatly the updated flux of cosmic rays above 3×10^{17}eV that has been measured combining data from surface and fluorescence detectors can not be included in this report since it is foreseen his publication in the next ICRC conference (http://icrc2015.nl).

5. The upgrade program of the experiment

The upgrade program of the Auger experiment will be presented at the april meeting of the CSN2, and subsequently at the Finance Board at the end of May. At CNAF several mass production have been runned in order to finalize and evaluate the impact of the new hardware on the future detector. Among the experimental improvement there is the possibility to use a new small sized photomultiplier[6]. The motivation for this study is to extend the dynamic range of the surface detector, reducing the impact of saturation for high energy events with shower axis close to a station. In order to do so an implementation of the small area PMT in the Simulation-Reconstruction framework has been performed and a mass CORSIKA proton shower production ad hoc has been realized[7].

References

[1] The Pierre Auger Collaboration 2010 *Nucl. Instr. and Methods in Physics Research* A **613** 29
[2] J. Knapp and D. Heck 1993 *Extensive Air Shower Simulation with CORSIKA, KFZ Karlsruhe* **KfK 5195B**
[3] The Pierre Auger Collaboration 2011 *Astropart. Phys.* **34** 368
[4] M. Tueros for the Pierre Auger Collaboration 2013 *Proc. 33rd ICRC, Rio de Janeiro, Brazil* **arXiv:1307.5059**
[5] V. S. Berezinsky and S. I. Grigorieva 1988 *Astron. and Astrophys.* **199** 1
[6] M. Aglietta et al., **GAP**2013 – 021, *Small PMT A proposal to extend the dynamic range of the Auger Surface Detector for operations beyond 2015*
[7] V. Scherini et al., **GAP**2014 – 089, *Implementation of a small PMT in the Offline simulation of the Surface Detector*

The Belle II Experiment at the INFN CNAF Tier1

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

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

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

1. Introduction

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

2. Computing Model

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

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

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

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

3. The Role of CNAF

INFN has officially joined the Belle II collaboration only in July 2013. Since then the Belle virtual organisation has already access to data centers in CNAF, Frascati, Legnaro, Napoli, Pisa, and Torino.
The INFN contribution to Belle II computing will consist of pledged resources located at CNAF and in Napoli (taking advantage of the ReCaS infrastructure) and eventually in other sites. CNAF will have the role of Regional Data Center while the other sites will be MC Production sites.
Hosting micro-DST format, detector and Monte Carlo data and ntuple-level user data at CNAF is already planned. In addition the possibility of having at CNAF also a fraction of the raw data and of the reprocessing is under discussion and a decision will be taken by the end of 2015.

4. Conclusions

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

The Borexino experiment at the INFN CNAF Tier1

Alessandra Carlotta Re
on behalf of the BOREXINO collaboration

Università degli Studi e INFN di Milano, via Celoria 16, 20133 Milano, Italy

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 with resulting in a shielding capacity against cosmic rays of 3,800 meter water equivalent (m.w.e.): at the LNGS, the muon flux is reduced of a factor 10^6 respect to the surface.

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

2. The Borexino recent result and future perspectives

The Borexino experiment started taking data in 2007. Since then, it has produced a considerable amount of interesting results which include the first direct measurement of proton-proton solar neutrino interaction rate, the precision measurement of the ^{7}Be solar neutrino rate (with a total error of less than 5%), the first direct measurement of the so-called pep solar neutrinos and the measurement of the ^{8}B solar neutrino rate with an unprecedented low energy threshold. Borexino has also published significant results on non-solar neutrino physics, such as the first observation of anti-neutrinos from the Earth (the geoneutrinos) and several limits on rare or forbidden processes. Among the most important scientific results obtained by Borexino during 2014 we recall especially the the first direct spectroscopy of proton-proton solar neutrinos [2] and

the extensive review [3] of all the Borexino phase-I measurements, including the identification of the annual modulation signature in the ^{7}Be neutrino signal.

Besides its application in the solar physics and geophysics fields, the Borexino detector offers a unique opportunity to perform a short-baseline neutrino oscillation study. This is the idea of SOX (Short distance neutrino Oscillations with boreXino). The SOX experiment[4] aims at the complete confirmation or at a clear disproof of the so-called neutrino anomalies, a set of circumstantial evidences of electron neutrino disappearance observed at LSND, MiniBoone, with nuclear reactors and with solar neutrino Gallium detectors. If successful, SOX will demonstrate the existence of sterile neutrino components and will open a brand new era in fundamental particle physics and cosmology. A solid signal would mean the discovery of the first particles beyond the Standard Electroweak Model and would have profound implications in our understanding of the Universe and of fundamental particle physics. In case of a negative result, SOX would be able to close a long-standing debate about the reality of the neutrino anomalies, would probe the existence of new physics in low energy neutrino interactions, would provide a measurement of the neutrino magnetic moment, and would yield a superb energy calibration for Borexino which will be very beneficial for future high-precision solar neutrino measurements. The SOX experiment will use two powerful and innovative neutrino and antineutrino generators made of ^{51}Cr and ^{144}Ce respectively. These generators will be located at a short distance from the Borexino detector and will yield tens of thousands of clean neutrino and antineutrino interactions in the internal volume of the Borexino detector. The SOX experiment is expected to start in 2016 and will take data for about two years.

3. Borexino computing at CNAF

At present, the whole Borexino data statistics and the user areas for physics studies are hosted at CNAF. The Borexino data are classified into three types: raw data, root files and DSTs. Raw data are compressed binary files with a typical size of about 600 Mb corresponding to a data taking time of ∼6h. Root files are reconstructed events files each organized in a number of `ROOT TTree`: their typical dimension is ∼1Gb. A DST file contains only selected events for high level analyses. Borexino standard data taking requires a disk space increase of about 10 Tb/year and a similar disk space is required for the Monte Carlo simulations. 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 a few weeks a year, an extraordinary *peak usage* (up to 500 cnodes at least) is needed in order to perform a full reprocessing on the whole data statistics with an updated version of the reconstruction code.

4. Conclusions

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

References

[1] Alimonti G et al. 2009 *Nucl. Instrum. Methods* A **600** 568
[2] Bellini G. et al. 2014 *Nature* **512** 383
[3] Bellini G. et al. 2014 *Phys. Rev.* D **89** 112007
[4] Bellini G et al. 2013 *JHEP* **8** 038

The CMS Experiment at the INFN CNAF Tier1

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

E-mail: Tommaso.Boccali@cern.ch

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

1. Introduction

The CMS Experiment at CERN collects and analyses data from the pp collisions in the LHC Collider. The first physics Run, at centre of mass energy of 7-8TeV, started in late March 2010, and ended in February 2013; more than 25 fb^{-1} of collisions were collected during the Run.
The CMS Experiment is designed as a general purpose detector, and hence is interested in a huge list of physics subjects; however, given the new energy regime the LHC can probe, the main expectations were on one side on the completion of the Standard Model, with the discovery of a Higgs-like boson, on the other side on the discovery if physics beyond the Standard Model, where multiple models were to be probed (Super-symmetry in all the possible incarnations, Extra dimensions, and all the sorts of more exotic models).
More than 300 physics papers were produced from Run I data, including the now renowned paper on the Observation of a 126 GeV Higgs Boson, which sets the final cornerstone to the Standard Model.
Year 2014 has been a year without collisions from the LHC; nevertheless computing activities have been frenetic: on one hand, resources have been used to complete Run I analyses, on the other the centres have undergone tests and new service deployment for the start on Run II in 2015.

2. The CMS Computing Model

CMS trigger rates, exceeding 1 kHz in the last months of 2012 (less than 500 Hz averaged on the Run, though), combined with large event sizes and computational needs, have a big impact on CMS Computing Model. CMS uses a derivative of the MONARC Hierarchical Model, based on GRID Middleware, where a Tier0, 7 Tier1 and roughly 50 Tier2 sites share the computational load. One of the Tier1s resides at CNAF, in Bologna, Italy.
The CNAF Tier1 has been used during Run1 to fulfil a series of tasks:

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

The resources CMS has deployed at CNAF amount to the 13% of the total Tier1 resources, the fraction being equal to the fraction of the Italian component in CMS; they amount (2014 numbers) to

- 22.75 kHS06 computational power;

- 7150 TB of tape;
- 3400 TB of disk.

Due to the very specific nature of CNAF, which serves all the LHC Collaborations and other less demanding experiments, CMS has actually been able to use large CPU over pledges quite constantly over time, consistently resulting as the second Tier1 as number of processed hours after the US Tier1.
The tape resource has been used at levels exceeding 90%, resulting in CNAF as the Tier1 holding more custodial data, again after the US Tier1. The disk resource has been used up 98%, at which point a cleaning was requested to CMS. Now, at the start of Run II, CNAF disk has about 1.5 PB free for data to arrive.

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

Starting from 2013, the CMS storage setup has evolved at CNAF.
Access to the files has been granted from remote locations via the Xrootd, and later in the year the disk has been split into a smaller tape cache, and a proper disk area directly managed by the experiment. The Xrootd servers have been directed only to this latter resource, protecting the tape area.
The new setup for the CPU + the disk area reduced significantly the differences between a Tier1 and a Tier2; indeed, during 2014 CNAF has opened the batch queues also for the standard analysis jobs, ramped from virtually zero at the end of 2013, to O(30%) level, limited just by the higher priority of production.

During 2014, CNAF has again been the second Tier1 in CMS as number of processed jobs, as already in 2013 (see Fig. 2).

Figure 2. Number of jobs processed by each CMS Tier1 during 2014.

3. New developments

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

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

- Usually, when CNAF storage is down for scheduled maintenance, all local CMS activity are stopped. In principle, local analysis activities accessing remote data via Xrootd can still work. This was tested on a 4 day storage downtime, where more 5000 analysis jobs were running simultaneously at CNAF. The overall job CPU efficiency has been exceeding 80%, comparable with local access, and additional failures due to the remote operation mode have been at the % level. This has resulted in a near saturation of the CNAF LHCOPN 40 Gbps line.

Furthermore CNAF provides experiments with testing environments for new technologies such as multi-core queues and many-cores systems, which CMS is actively using to evolve its computing and software frameworks. Indeed, multi-core jobs are going to become the standard in the next years, and CNAF is well prepared to handle them.

4. Expected resource growth

The LHC collider is at the moment (March 2015) off for upgrade, and is expected to go back in operations within the Spring 2015, with a new centre of mass energy of 13 TeV. The new Run II will last up to 2018, with an instantaneous luminosity more than doubled with respect to Run I. CMS expects to carry on an extensive study of the Higgs boson properties during Run II, while performing searches for new physics at the newly available energy.
The former aspect needs an increased trigger, which should stably collect 1 kHz of more complex events. Present estimates require, even in presence of drastic optimizations, roughly a factor 2 in Tier1 CPU resources. For 2015 CPU pledges at CNAF have increased by 70%, to 39 kHS06; disk pledges have not changed, and tape requests have skyrocketed to almost 10 PB. Expectations for 2016 prospect a milder resource increase, with CPU reaching over 50 kHS06, disk 4 PB and tape 13 PB, and are currently under scrutiny by the RRB-CRSG scrutiny group.

5. Conclusions

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

The COKA Project

R Alfieri[1], M Brambilla[1], E Calore[2], R De Pietri[1], F Di Renzo[1], A Feo[1], S F Schifano[2] and R Tripiccione[2]

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

E-mail: alfieri@pr.infn.it, brambilla@pr.infn.it, calore@fe.infn.it, depietri@pr.infn.it, direnzo@pr.infn.it, feo@pr.infn.it, schifano@fe.infn.it, tripiccione@fe.infn.it,

Abstract.

This document describes the work carried out by the COKA project in 2014. In summary, the project has focused on the use of MIC-based accelerator boards as working horses for several computational applications in theoretical physics.

These accelerators were released by Intel at the beginning of 2013 under the name Xeon-Phi. We have worked on the development of the best optimization strategies to use these processor efficiently, including an analysis of the available programming environments, and have assessed the performance and usability of these systems with respect to more standard processors. We have also started to consider the best options to use parallel systems containing several MIC accelerators.

1. Introduction

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

A large fraction of our work has been focused on computational physics: we have completed the porting of a production-grade lattice-Boltzmann code for computational fluid-dynamics, and investigated scalability over a multi-accelerator system. We have also tested performance of two widely used application framework, Einstein Toolkit and CHROMA, and investigated programming methodologies that allow to easily port codes across accelerators.

2. Main Results

In the following we highlight some interesting results obtained during the year 2014.

2.1. Experimental Hardware

We have used the funds assigned to the project to procure, install and configure part of an R&D testbed at CNAF. The COKA part of this testbed currently contains two servers, with a total of 3 Intel Xeon-Phi boards, 2 NVIDIA K20 boards and 24 x86-cores processors. These machines are accessible by COKA users through a user interface and a batch system, configured with a number of queues serving different purposes. For part of our work (mainly for the analysis of the

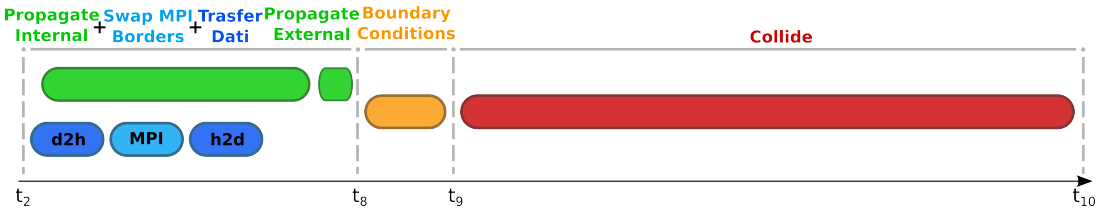


Figure 1. Program schedule allowing to overlap communications and processing of the `propagate` kernel.

scaling properties of our systems on massively parallel systems) we have also used accelerator based systems installed at CINECA and at the Jülich Supercomputer Centre.

2.2. Computational Fluid-dynamics

We have completed the porting of our fluid-dynamics code based on Lattice Boltzmann methods on the Xeon-Phi architecture in 2013 [1]. In 2014 we have fine-tuned the code to run it efficiently on multi-accelerator systems; we have also compared the performance of the same physics code, optimized for several architectures, including GPU accelerators [2, 3, 4] and "classic" multi-core CPUs [5, 6, 7].

On multi-accelerator systems, accelerators are installed on the same host (typically up to 4 or 8 accelerators) or on multiple hosts that exchange data using commodity network such as Infiniband. Here we discuss an implementation for one host attached to four accelerators.

The computationally relevant core of a Lattice Boltzmann program is basically a stencil code working on a discrete and regular spatial lattice. Two routines are associated to most of the computational load: propagate is dominated by memory moves at regular but sparse memory addresses, while collide applies a large set of mathematical operations to the data items associated to each lattice site; finally pbc contains all the communications steps across different processing elements.

Our implementation splits a lattice of size $L_x \times L_y$ on N_p accelerators along the X dimension, each handling a *sub-lattice* of size $L_x/N_p \times L_y$. This splitting is necessary to have continuous halo-columns allocated in memory, and avoid a further gather-step to collect halos on a contiguous buffer before doing communications. This allocation scheme implies a virtual ordering of the accelerators along a ring, so each one is connected with a previous and a next companion; at the beginning of each time-step, before starting `propagate`, accelerators must exchange data, since cells close to the right and left edges of the sub-lattice needs data allocated on the logically previous and next nodes.

When using multi-accelerator systems, the overlap of communications with computation is a key approach to scalability. In our case, the key point to consider is that propagate for the bulk of the lattice (all lattice points except for three columns at right and left) has no data dependency with pbc (while propagate on the edges depend on fresh data moved to the halos by pbc). Our strategy therefore tries to overlap as much as possible data transfers with the execution of propagate on the bulk.

The details of the program schedule that optimize this overlap are shown in Figure 1. The overall sequence of operations is the same for our MIC and GPU implementations. A conceptually irrelevant but technically important difference between the two cases is that on MIC systems MPI functions cannot access data allocated on the accelerator, so the code must explicitly move data between host and accelerator. This require additional care: details are shown in Figure 2.

In Table 1 we compare performance figures of our code optimized for the Xeon Phi, the

```
// launch asynchronous transfer from device to host (D2H)
#pragma offload_transfer: out( cf2[LEFT_HALO]  : REUSE into( send_L_buf ) ) signal( &send_L_buf )
#pragma offload_transfer: out( cf2[RIGHT_HALO] : REUSE into( send_R_buf ) ) signal( &send_R_buf )

// launch asynchronous execution of propagate kernel over BULK
#pragma offload: signal( &internal_prop_signal ){ propagate_m ( ... ); }

// wait end of d2h transfer
#pragma offload_wait: wait( &send_L_buf )
#pragma offload_wait: wait( &send_R_buf )

// execute halos SWAP
MPI_Sendrecv(send_R_buf to mpi_rank_R, TAG_RIGHT, recv_L_buf to mpi_rank_L, TAG_RIGHT);
MPI_Sendrecv(send_L_buf to mpi_rank_L, TAG_LEFT,  recv_R_buf to mpi_rank_R, TAG_LEFT);

// launch asynchrouns transfer from host to device (H2D)
#pragma offload_transfer: in( recv_L_buf : REUSE into(cf2[LEFT_HALO] )) signal( &recv_L_buf )
#pragma offload_transfer: in( recv_R_buf : REUSE into(cf2[RIGHT_HALO])) signal( &recv_R_buf )

// wait end of h2d transfer
#pragma offload_wait: wait( &recv_L_buf )
#pragma offload_wait: wait( &recv_R_buf )

// launch asynchronous execution of propagate over left- and right-columns
#pragma offload { propagate_m ( ... ); } signal(&external_prop_signal)

// wait end of propagate kernels
#pragma offload_wait: wait( &internal_prop_signal )
#pragma offload_wait: wait( &external_prop_signal )
```

Figure 2. Scheduling of operations on the MIC board to overlap communications and execution of `propagate`.

Kepler K80 GPU and for a dual-processor commodity systems [7] (dual Intel E5-2630), based on the Haswell micro-architecture.

We first focus on comparing performances with just one accelerator. The `propagate` kernel is a memory-bound step corresponding to memory copies with sparse address patterns. On the *Kepler* architecture we reach $\approx 64\%$ of the available peak bandwidth. The *Xeon-Phi*, that uses the same class of memories, obtains a lower bandwidth, ≈ 85 GB/s, that is $\approx 24\%$ of peak. This is mainly due to the limited bandwidth (≈ 220 GB/s) of the internal ring, connecting cores and memory controllers. On the Haswell processor we measure ≈ 40 GB/s corresponding to $\approx 67\%$ of the available peak.

The `collide` kernel is a strongly compute-bound step, requiring approximately 20 double-precision floating-point operations per byte. On *Kepler*, this kernel reaches a sustained performance of $\approx$ 46% of the available peak. The *Xeon-Phi* performance is lower, only approximately $\approx 30\%$ of the available peak, while on the Haswell CPU we measure an efficiency of $\approx$ 29%. The last section (Global P) of the table shows the performance of the full code, measured in *Millions Lattice Update Per Second* (MLUPS). Comparing with the traditional

Table 1. Performance comparison of the `propagate` and `collide` kernels on KNC and Kepler accelerators and on a dual 8-core E5-2630 (Intel Haswell micro-architecture) processor running at 2.4 GHz.

	Intel Xeon 7120				NVIDIA K80				E5-2630 v3	
# devices(s)	1	2	3	4	1	2	3	4	1	2
Propagate (GB/s)	85	161	225	274	154	266	393	520	40	88
S_r	1.0X	1.90X	2.65X	3.22X	1.0X	1.72X	2.55X	3.37X	1.0X	2.2X
Collide (GFs)	358	709	1056	1383	667	1359	2029	2706	111	222
S_r	1.0X	1.98X	2.95X	3.86X	1.0X	2.03X	3.07X	4.05X	1.0X	2.0X
Global P (MLUPS)	39	73	110	139	72	140	209	277	12	23
S_r	1.0X	1.95X	2.82X	3.56X	1.0X	1.93X	2.88X	3.84X	1.0X	1.98X

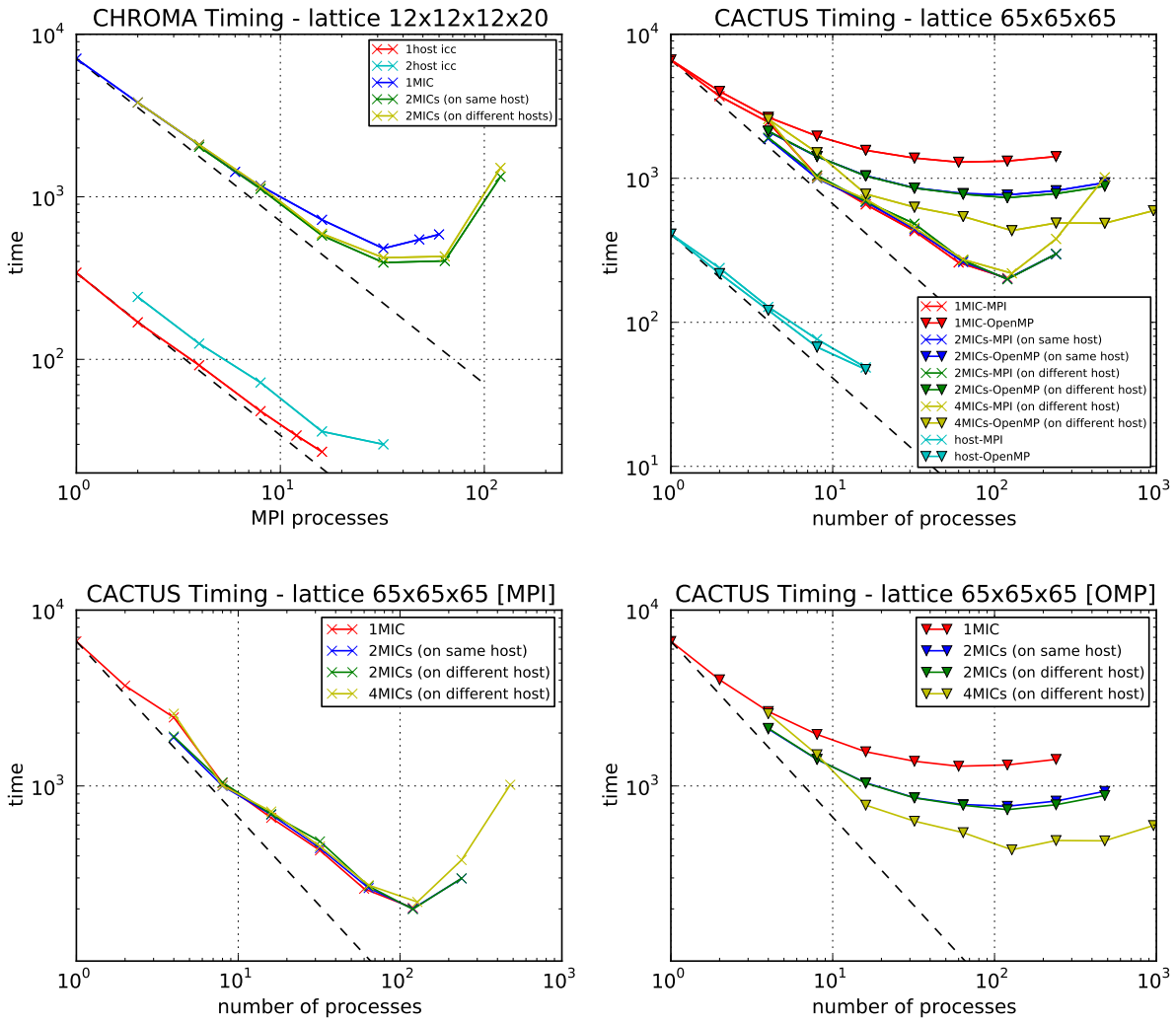


Figure 3. The top-left panel shows the time needed to perform 200 Monte Carlo sweeps on a $12^3 \times 20$ lattice for pure Gauge SU(3) using CHROMA. The execution time is 341 seconds on a host core while it is 7033 seconds on a PHI core. The top-right panel shows the time needed to compute 32 time-evolution steps of a single General Relativistic Star using the EinsteinToolkit on a 65^3 grid (0.6 total GBytes allocated memory). On a single host core the requested time is 410 seconds while on a single PHI core the requested time is 6857 seconds. The bottom panels disaggregate MPI (left) and OpenMP (right) scaling data for the EinsteinToolkit.

eight-core CPU, the *Xeon-Phi* is $\approx$ 3X faster, while on one GPU of the K80 system, the speed-up is 6X.

Table 1 also shows scalability results (S_r). We see that individual steps and the full code scale quite well meaning that communications have been successfully hidden with computation; running with 4 GPUs the sustained performance of the full code is $\approx$ 1.7 TFLOPS.

In conclusion, our application enjoys up to a 6X performance increase running on accelerators, and also a good scaling on multi-accelerator systems. While these are valuable results, we underline that they were obtained with handcrafted optimizations tailored for each target accelerator. This is mainly due to the lack of programing methodologies able to support execution of the same code on different accelerator architectures. Portability of codes and also

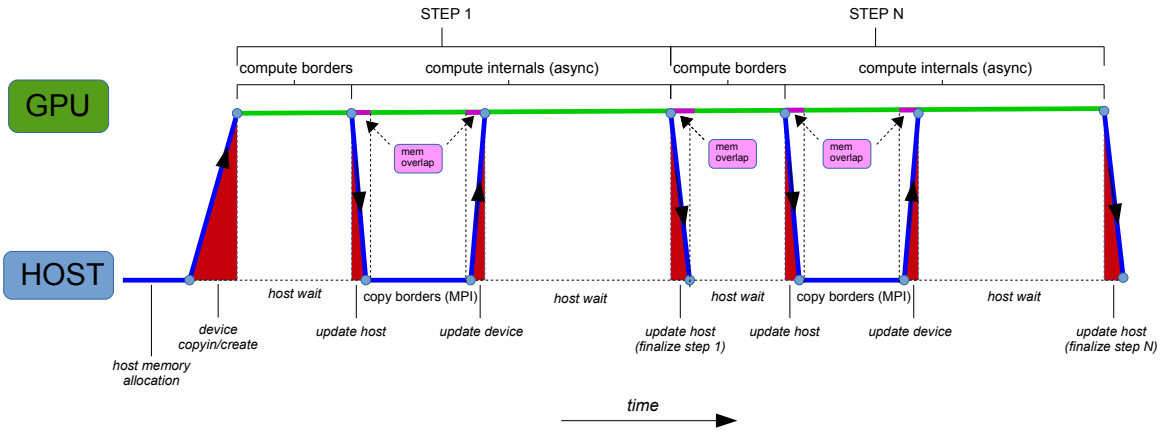


Figure 4. Communication-computation overlap between host system and GPU implemented.

of performances is today a real issue that needs to be solved to make accelerators a standard solution for HPC computing.

2.3. Chroma and Cactus

We started a test work on two widely used and freely available application frameworks: the **CHROMA** lattice QCD application [8] and the **Einstein Toolkit** Astrophysical Application suite [9]. The main problem we had to deal with was the compilation (for native mode execution) of the auxiliary libraries needed by these applications. The strategy used in testing these applications was to analyze the performance that can be obtained by just recompiling on the MIC environment without any code customizations, that is, using the potential key advantage of the MIC architecture with respect to other accelerators like GPUs. The key result of this simple procedure is that it is indeed possible – albeit with a non negligible effort – to run an unmodified large scientific package on the MIC, but performances are dramatically unsatisfactory if compared with the execution on the host (2x Sandy Bridge, E5-2687W 3.10 GHz, 8 cores) so a significant tuning and optimization work is definitely necessary. The results of porting are summarized in Figure 3 where in the top panels are expressed the overall scaling result (the dashed lines represent perfect scaling) while the bottom panels show the additional problems of the OpenMP bad scaling of the use of a standard OpenMP porting when the number of threads is large.

2.4. Test of portable programming techniques

A further contribution to the project has been the study of portable programing techniques that can be exploited on GPUs on many-core systems. This work was carried out in the framework of a Thesis for the "laurea" degree in Physics and was based on the porting of the classical "game-of-life" with an addition of a highly vectorizable computational core. The programming environments used were OpenMP for the *Xeon-Phi* (MIC) and openACC for the *Kepler* GPU, the common idea in both cases was to use "annotated" standard C code. On the GPU side the possibility to overlap communication (GPU-host) and computation (see Figure 4) is extremely relevant. Using this technique a *Kepler* GPU has a speedup of 77.5 with respect to a serial version of the same code on a dual 8-core E5-2630 Xeon processor. For the *Xeon-Phi* (MIC) the speedup is much smaller, 19.1, using the maximum allowed number of threads (240). To explain this difference, one remarks that on the *Kepler* GPU it was essential to overlap computation

and memory access and to ensure a reasonable level of vectorization of the main computational kernel; these were not possible using the OpenMP 2.0 directives used for the *Xeon-Phi* porting.

Acknowledgments
This work was done in the framework of the COKA, and SUMA projects of INFN. We thank CINECA (Bologna, Italy), and the NVIDIA Jülich Application Lab (Jülich Supercomputer Center, Germany) for allowing us to use their computing systems.

References

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

[2] L. Biferale, F. Mantovani, M. Pivanti, F. Pozzati, M. Sbragaglia, A. Scagliarini, S. F. Schifano, F. Toschi, R. Tripiccione, *A multi-GPU implementation of a D2Q37 Lattice Boltzmann Code*, 9a International Conference on Parallel Processing and Applied Mathematics (PPAM11), September 11-14, 2011, Torun (Poland). R. Wyrzykowski et al. (Eds.): PPAM 2011, Part I, LNCS 7203, pp. 640-650. Springer, Heidelberg (2012), doi:10.1007/978-3-642-31464-3_65.

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

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

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

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

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

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

[9] Frank Löffler, Joshua Faber, Eloisa Bentivegna, Tanja Bode, Peter Diener, Roland Haas, Ian Hinder, Bruno C. Mundim, Christian D. Ott, Erik Schnetter, Gabrielle Allen, Manuela Campanelli, and Pablo Laguna. The Einstein Toolkit: A Community Computational Infrastructure for Relativistic Astrophysic s. Classical and Quantum Gravity, 29(11):115001, 2012. (doi:10.1088/0264-9381/29/11/115001)

The Cherenkov Telescope Array

Ciro Bigongiari

INAF - Osservatorio Astrofisico di Torino,
Strada Osservatorio 20 - 10025 Pino Torinese, Torino, IT

E-mail: bigongiari@oato.inaf.it

Abstract. The Cherenkov Telescope Array (CTA) is an ongoing international project to build a new generation ground-based gamma-ray detector composed by tens of Cherenkov telescopes of different size which will cover a very wide energy range, between few teens of GeV to some hundreds of TeV. It will be operated as an open observatory and will provide a deep insight into the non-thermal processes which are responsible of the high energy emission by many astrophysical sources, like Supernova Remnants, Pulsar Wind Nebulae, Micro-quasars, Active Galactic Nuclei and Gamma Ray Bursts.

1. Introduction

Very High Energy gamma-rays can be produced in the collision of highly relativistic particles with surrounding gas clouds or in their interaction with low energy photons or magnetic fields. Possible sources of such energetic particles include jets emerging from active galactic nuclei, special radio galaxies, 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 therefore also for fundamental physics measurements, such as the indirect search of dark matter, searches for high energy violation of Lorentz invariance by detection of delays between the arrival times of photons with different energies and searches for axion-like particles which, besides being possible candidates for dark matter, may also explain the unexpectedly low absorption by extra-galactic background light of gamma rays from very distant blazars. High-energy gamma-rays can be used moreover to trace the populations of high-energy particles, thus providing insightful information about the sources of cosmic rays. Close cooperation with observatories of other wavelength ranges of the electromagnetic spectrum, and those using cosmic rays, neutrinos and gravitational waves are foreseen. To ensure a full coverage of the sky the CTA detector will be composed actually by two arrays, one placed in the Southern hemisphere and one in the Northern one. The selection of the two sites is still ongoing but the number of candidates has been reduced to four, two for the south site and two for the north one. The final decision about the south site is expected by August 2015, while for Northern site the deadline is November 2015. The construction of CTA will start shortly after the decision and the whole CTA Observatory is expected to become fully operational in 2020. A detailed description of the project and its expected performance can be found in a dedicated volume of the Astroparticle Physics journal [1].

2. Computing needs

The CTA project is presently in the pre-production phase when many innovative technologies are being tested and developed for the construction of the various classes of telescopes. Meanwhile detailed Monte Carlo simulation of the entire array are ongoing to estimate its overall performance and to optimize many parameters like the telescope layout and the trigger strategy. A huge effort has been dedicated so far to the evaluation of the expected performance at many different site candidates to provide the information needed for a fair comparison to the site selection committee. Each site requires a full simulation because the performance of a Cherenkov telescope depends on many site-dependent parameters like its altitude, atmospheric conditions, geomagnetic field and night-sky brightness. Due to the very effective hadronic background rejection achieved with the imaging air Cherenkov technique a huge amount of simulated background events is needed to achieve reliable estimates of the array performance. About 10^{10} cosmic ray induced atmospheric showers for each site are needed to properly estimate the array sensitivity, energy and angular resolution requiring extensive computing needs in term of both disk space and CPU power. About 2.6 million of GRID jobs have been executed in 2014 for such task corresponding to about 110 millions of HS06 hours of CPU power and 630 TB of disk space. CNAF contributed to this effort with about 7 millions of HS06 hours, corresponding to 6.4% of the overall CPU power used, resulting the fourth contributor in terms of CPU time and the sixth in terms of disk space. For 2015 even larger needs of CPU power and disk space are foreseen due to the ongoing simulations for the array layout optimization of the selected sites. CNAF will surely keep its role among the main contributors to the CTA virtual organization thanks to the greatly increased dedicated resources.

References

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

CUORE experiment

CUORE collaboration

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

Abstract. CUORE is a ton scale bolometric experiment for the search of neutrinoless double beta decay in ^{130}Te. The detector is in the advanced commissioning phase at the Laboratori Nazionali del Gran Sasso of INFN, in Italy. It is composed by an array of 988 TeO_2 bolometers, for a total mass of 0.75 ton. Based on the performance of CUORE-0, a pilot experiment which is running since march 2013, 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 migration of the CUORE data processing to the CNAF computing cluster has started 2014, and a more intense use of resources is expected in 2015.

1. The experiment

The main goal of the CUORE experiment [1] is to search for neutrinoless double beta decay (0νDBD) of the isotope ^{130}Te. In this spontaneous decay a nucleus changes its atomic number by two units, and two electrons are emitted. Its observation would imply the Majorana nature of the neutrino mass, and could give information on the neutrino mass hierarchy and absolute scale. To date there is no experimental evidence for this decay, and the half life limits lie in the range $10^{22} \div 10^{25}$ y, depending on the isotope that is being considered. In a calorimetric detector, the sum energy of the two electrons emitted in 0νDBD produces a sharp peak in the spectrum, centered at the Q-value of the decay. Typical 0νDBD Q-values are in the few MeV range, therefore the tiny signal is submerged by background from natural radioactive decays. The CUORE detector is an array of 988 $^{nat}TeO_2$ bolometers, with a total mass of 741 kg (201 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, equipped with a neutron transmutation doped (NTD) germanium sensor for signal readout. The TeO_2 crystals are coupled to a copper support structure by mean of small PTFE supports. The bolometer array is enclosed in a dilution refrigerator whose mixing chamber is cooled to $\sim$10 mK and thermally coupled to the copper support structure of the detectors. The CUORE bolometers act at the same time as source and detectors for the sought signal. When a particle interacts in a CUORE crystal, it produces a sizable temperature rise, ΔT=E/C, that can be read by the NTD sensor. The CUORE collaboration aims at reaching a background of 10^{-2} counts/(keV·kg·y) in the region of the energy spectrum where the 0νDBD signal is expected ($Q_{\beta\beta} \simeq 2528$ keV), and a FWHM energy resolution of 5 keV. With these parameters, the experiment will reach a ^{130}Te half-life sensitivity of about 10^{26} y in five years of live time.

2. Status of CUORE and CUORE-0

The CUORE experiment is currently in the advanced commissioning phase at the Laboratori Nazionali del Gran Sasso of the INFN, Italy. All the 19 bolometer towers were successfully built

and are now stored underground in nitrogen overpressure. The commissioning of the CUORE cryostat is ongoing. In 2014 the cryostat reached a stable base temperature of 6 mK. After this test the detector wires running from the mixing chamber to room temperature were installed, and a mini-tower of 8 bolometers was operated for the first time in the CUORE cryostat at the end of 2014. The commissioning of the cryostat will continue during the first half of 2015. In autumn the CUORE towers will be installed in the cryostat and the detector cool down will start by the end of the year.

To check the effectiveness of the CUORE detector assembly procedure, a first CUORE-like tower made of 52 bolometers, named CUORE-0 [2], was built and is taking data in the former Cuoricino [3] cryostat since April 2013. From a study of the CUORE-0 background spectrum and with the aid of Monte Carlo simulations, it could be possible to evince that the CUORE background goal of 10^{-2} counts/(keV·kg·y) is within reach. Moreover CUORE-0 measured an average energy resolution slightly better than 5 keV FWHM on the 2615 keV photoelectric peak from ^{208}Tl, in perfect agreement with the energy resolution goal of CUORE.

3. CUORE computing model and the role of CNAF

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

The CUORE-0 and CUORE data analysis flow consists in two steps. In the first level analysis the event-based quantities are evaluated, while in the second level analysis the energy spectra are produced and studied. The analysis software is organized in sequences. Each sequence consists in a collection of modules that scan the events in the Root files sequentially, evaluate some relevant quantities and store them back in the events. The analysis flow consists in several fundamental steps that can be summarized in pulse amplitude estimation, detector gain correction, energy calibration and search for events in coincidence among multiple bolometers.

Most of the CUORE-0 data analysis and simulations were run on a dedicated computing cluster located at the Roma1 division of INFN. Since 2014 a transition phase has started to move the CUORE-0 and CUORE analysis and simulation software to CNAF. In 2014 the CUORE data analysis software was installed, configured and tested at CNAF, and some CUORE and CUORE-0 Monte Carlo simulations were run. In 2015 a much more intense usage of the CNAF computing resources is expected, both in terms of data analysis (reprocessing of the CUORE-0 data) and simulations. When the CUORE data taking will start, it is expected that the primary data storage and a computing cluster for basic analysis will be located at LNGS. The data processing and the simulations will be run at the CNAF computing cluster, which will also host a copy of the raw data and of the data analysis root files.

References

[1] Artusa D *et al.* (CUORE) 2015 *Adv.High Energy Phys.* **2015** 879871 (*Preprint* 1402.6072)
[2] Artusa D *et al.* (CUORE) 2014 *Eur.Phys.J.* **C74** 2956 (*Preprint* 1402.0922)
[3] Andreotti E *et al.* 2011 *Astropart.Phys.* **34** 822–831 (*Preprint* 1012.3266)

The *EEE Project* activity at CNAF

E. Fattibene, A. Ferraro, B. Martelli, F. Noferini, V. Sapunenko, C. Vistoli, S. Zani

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

E-mail: enrico.fattibene@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 Trasnfers

The EEE activity at CNAF started with the data collected during the EEE pilot run in 2014, involving 21 schools (+ two INFN telescopes) in a coordinated data aquisition. During that run[1] all the schools were connected/authenticated at CNAF in order to transfer data. To realize this goal in the initial phase of the project a SL5 virtual machine was installed in each data acquistion system running a rsync client. Then, after a test activity, this system based on rsync has been replaced by the BitTorrent Sync [2] software, that is able to automatically syncronize among different peers. There are many advantages to move to BitTorrent Sync: the client is

[1] Pilot run from 27-10-2014 to 14-11-2014.

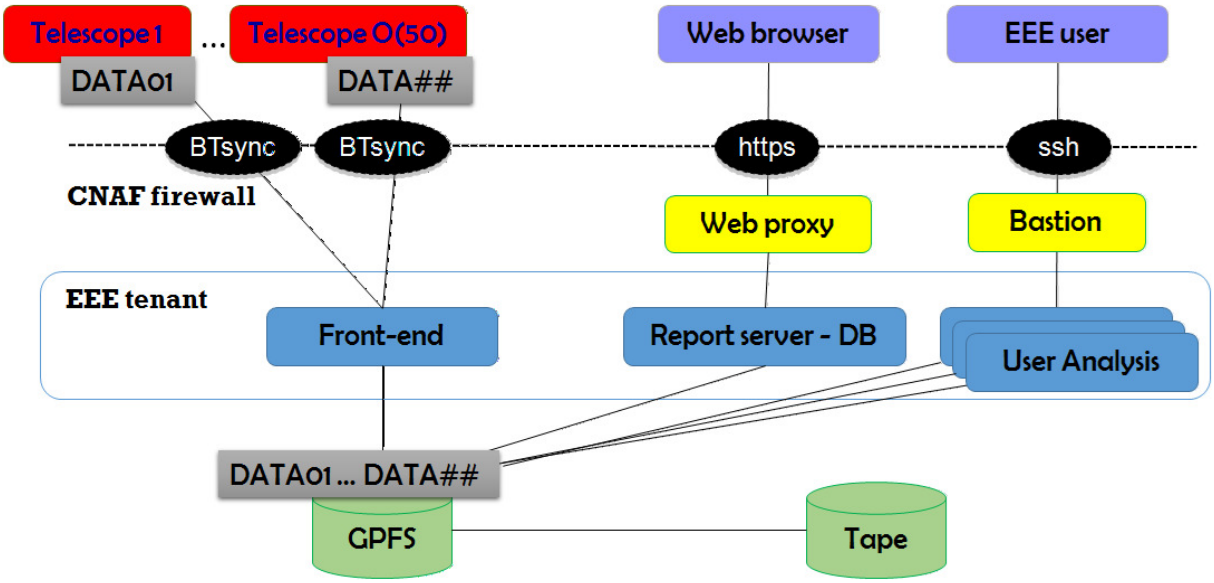


Figure 1. Architecture of the EEE IT infrastructure.

deployed directly on the same system of the data aquisition (Win OS); it is easy to add one or more peers to replicate data in other geographical distributed sites; a web interface is provided to the adminstrators to control the status of all transfers. In the CNAF cloud infrastructure a front-end was dedicated to receive all the data with a total required bandwidth of 300 kB/s, to collect the expected 510 TB per year. All the data collected are considered as custodial and for this reason the storage was configured to allow the migration of the data to a tape system which is expected to be ready by 2015. In Fig. 1 the general architecture for the EEE data flow is reported.

In the period including the pilot run we collected about 1 billion cosmic rays, corresponding to 300 GB data transferred at CNAF. In Fig. 2 a summary of the data flow performances during the pilot run is reported.

3. Data Reconstruction/Monitor/Analysis

The chain to reconstruct data at CNAF is fully automated. 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 [3] (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 [4] 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.

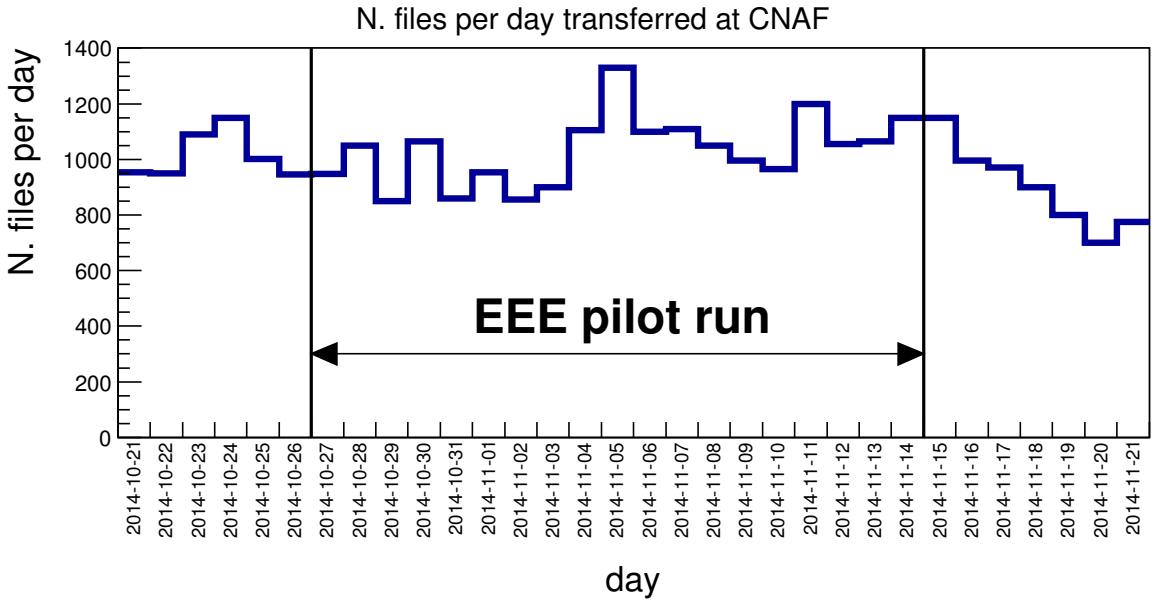


Figure 2. Statistics for the EEE pilot run in 2014. For each day the number of files transferred at CNAF is reported.

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

Extreme Energy Events Monitor

Ultimo aggiornamento: ore 17:14 - mer 10 dicembre 2014

ELOGBOOK delle SCUOLE | ELOGBOOK dello SHIFTER | Stato trasmissione CNAF

EEE Main Monitoring Table

Questa tabella mostra la situazione dei telescopi in acquisizione
In **verde** sono indicati i telescopi in presa dati e trasferimento nelle ultime 4 ore.
In **giallo** sono indicati i telescopi in cui trasferimento e/o acquisizione sono sospesi da più di 4 ore.
In **rosso** sono indicati i telescopi in cui trasferimento e/o acquisizione sono sospesi da più di un giorno.

Scuola	Giorno	Ora	Nome dell'ultimo File trasferito	Numero Files trasferiti oggi	Ultima Entry nell'e-logbook delle Scuole	Report giornaliero DQM	RATE of Triggers for the last Run in DQM	RATE of Tracks for the last Run in DQM	Link DQM
ALTA-01	mer 10 dicembre	08:06	ALTA-01-2014-12-10-00020.bin	20 [History]	10:56 21/11/2014	10/12 [History]	27.4	21.0	ALTA-01
BARI-01	mer 10 dicembre	05:30	BARI-01-2014-12-10-00008.bin	8 [History]	09:59 06/12/2014	10/12 [History]	21.5	18.7	BARI-01
BOLO-01	mer 10 dicembre	17:00	BOLO-01-2014-12-10-00060.bin	62 [History]		10/12 [History]	47.4	28.3	BOLO-01
BOLO-03	mer 10 dicembre	17:01	BOLO-03-2014-12-10-00048.bin	48 [History]	13:19 10/12/2014	10/12 [History]			BOLO-03
CAGL-01	mer 10 dicembre	17:06	CAGL-01-2014-12-10-00016.bin	16 [History]	08:22 09/12/2014	10/12 [History]	17.4	14.5	CAGL-01
CAGL-02	mer 10 dicembre	16:54	CAGL-02-2014-12-10-00051.bin	51 [History]	12:01 03/12/2014	10/12 [History]	39.9	33.8	CAGL-02

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

4. Future work

In 2014 all the EEE virtual services were depoyed and managed by CNAF people. In 2015 the EEE users will be able to instanciate the analysis virtual machines in a self service mode. We are also investigating the possibility to allow EEE users to submit batch jobs on the CNAF Tier1 facility. 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] INFN-CNAF, EEE monitor, *https://www.centrofermi.it/monitor/*.
[4] M. Abbrescia et al., The EEE Project: Cosmic rays, multigap resistive plate chambers and high school students, JINST 7 (2012) 11011.

The GERDA experiment

E. Medinaceli on behalf of the GERDA collaboration
Università di Padova, via Marzolo 8, 35100 Padova, Italy
E-mail: medinaceli@pd.infn.it

Abstract. GERDA (GERmanium Detector Array) is an experiment searching for the neutrino-less double beta decay ($0\nu\beta\beta$) of the isotope ^{76}Ge. The first part of the experiment was carried on between 2011 and 2013. The second phase is presently under preparation at the Laboratori Nazionali del Gran Sasso of INFN. This contribution briefly reviews the basic facts and results achieved by GERDA, with a particular focus given on computing.

1. The experiment, latest results and on-going upgrade

The GERmanium Detector Array (GERDA), located at the INFN underground laboratory of Gran Sasso, is an experiment searching for the neutrino-less double beta decay ($0\nu\beta\beta$) of ^{76}Ge. This type of decay, which is predicted by several extensions of the Standard Model of particle physics, is a process that violates the lepton number conservation by two units. The decay is possible only if neutrinos have a Majorana mass component. In the assumption that the decay is mediated by the Majorana neutrino mass, its half-life is directly connected to the neutrino absolute mass, which is presently unknown: neutrino oscillation experiments provide a measurement of the mass splitting of the neutrino eigenstates, but cannot pinpoint the absolute mass scale. The $0\nu\beta\beta$ decay, if exists at all, is a very rare process, with half-life exceeding 10^{25} yr. Therefore, experiments aiming to search for it must feature a ultra-low background.

The experiment uses an array of high-purity germanium (HPGe) detectors immersed in a new shielding concept of liquid argon and water [1, 2]. The HPGe detectors are made from germanium isotopically enriched to about 86% in the isotope of interest, ^{76}Ge. An array of HPGe detectors is deployed in a 64 m^3 cryostat filled with liquid argon (LAr). The LAr serves as cooling medium and shielding against external backgrounds. The shielding is complemented by 3 m of ultra-pure water instrumented with photo-multipliers to detect Cherenkov light emitted by cosmic ray muons [1]. The LAr volume was used as a passive shielding during the first phase of GERDA, but will be operated as an active veto in the Phase II, to further suppress the residual background.

The signature for $0\nu\beta\beta$ decay is a single peak at $Q_{\beta\beta}$-value of the decay (2039 keV for ^{76}Ge). Pulse shape discrimination techniques of the charge signal from the HPGe detectors are employed to improve the experiment sensitivity. The so-called "GERDA Phase I" took data between November 2011 and May 2013 with eight enriched HPGe detectors from the predecessor experiments Heidelberg-Moscow [3] and IGEX [4], and five new-generation custom-made enriched detectors of BEGe type. A total total exposure of 21.6 kg·yr was collected with a mean background index at $Q_{\beta\beta}$ of 10^{-2} counts/(keV kg yr).

A blind analysis approach was pursued: events in the region of interest around $Q_{\beta\beta}$ were initially

not made available for analysis. The blinded region was opened only when all cuts and algorithms had been frozen. No signal was observed and a lower limit was derived for the half-life of $0\nu\beta\beta$ decay of ^{76}Ge, $T^{0\nu}_{1/2} > 2.1 \times 10^{25}$ yr (at 90% C.L.) [2]. These results show no indication of a peak at $Q_{\beta\beta}$, i.e. the claim for the observation of $0\nu\beta\beta$ decay in ^{76}Ge based on the Heidelberg-Moscow data was not confirmed [5]. A background model has been developed in GERDA [6] to describe the observed energy spectrum outside the blinded region at $Q_{\beta\beta}$. The model contains several contributions, that are expected on the basis of material screening or that are established by the observation of characteristic structures in the energy spectrum. The model was used to predict the intensity and the spectral shape of the background in the region of interest around the $Q_{\beta\beta}$, before the actual unblinding.
GERDA is currently in the transition to the so-called "Phase II" aiming to significantly increase the experimental sensitivity by collecting a larger exposure (about 100 kg·yr) and by further suppressing the residual background. About 30 new-generation BEGe type enriched HPGe detectors [7] will be deployed, which will more than double the total available germanium mass. A major upgrade is the instrumentation of the LAr volume surrounding the detector array as an active veto system. The goal of GERDA Phase II is to achieve a background index of the order of 10^{-3} counts/(keV kg yr) at $Q_{\beta\beta}$. Given an exposure of 100 kg·yr, this would yield a sensitivity on $T^{0\nu}_{1/2}$ of about 10^{26} yr. Currently the experiment is taking the first commissioning data for Phase II.

2. GERDA computing

The two key paradigms adopted for the offline data processing are: (1) blinding and (2) hierarchical structure. Due to the requirement of blinding events in the region of interest at $Q_{\beta\beta}$, the raw data files produced by the DAQ system cannot be immediately distributed to the Collaboration, but have to be pre-processed to remove "sensitive" events. The original data files must be stored on a protected disk area and then re-processed for the unblinding.
The analysis chain is performed in a layer of storage file structures, whose level corresponds to a different stage of the reconstruction of the event and to a different kind of information. The first level of the data (raw data), which contain pulse shapes and other digitizer information (time stamps, etc.), is produced directly by the DAQ and has its own specific binary format. Data at this level is referred as Tier0. Raw data are hence converted into a standardized format, which is based on the MGDO libraries [8], and stored as ROOT files [9]. This second level of the data structure is referred as Tier1. The Tier1 files basically contain the same information as the Tier0 (waveforms and digitizer data), except for the events falling into the region of interest: they are removed from the Tier1 file in the process and stored separately. Having a standardized exportable and stable format is functional to the decoupling of the higher-level analysis from the specific binary format of the raw data: this allows the same algorithms to be used plug-and-play, irrespectively of the binary format of the parent DAQ system.
Afterward, waveforms are analyzed individually by the application of a chain of digital signal processing algorithms. The algorithms are coded as nearly-independent modules within the software framework GELATIO [10], thus to provide the flexibility to produce many user-customized chains of signal processing. The output of the GELATIO modules (e.g. rise time, pulse amplitude, etc.) are stored as a ROOT file (labeled Tier2) and used for higher-level analysis. Finally, quality cuts, energy calibration and pulse shape discrimination are performed, whose results are stored in new ROOT files denoted as Tier3 and Tier4. In the path from the Tier1 to Tier4, the physics information is extracted and digested for the final analysis and the size of the files shrinks by a factor of ~ 1000. The final physics analysis is performed on the Tier4 files.

Being GERDA a low-background experiment, the data throughput is modest. In Phase I, it was about 4 GB/day for the physics data taking, plus 20 GB/week for calibrations. The data

rate, including the pre-processing and the basic reference offline analysis, could be handled by using the GERDA computing resources at Gran Sasso.
The Collaboration policy requires three backup copies of the raw data to be made, in addition to the master version which is stored at the Gran Sasso laboratory. The copies must be kept in Italy, in Germany and in Russia. The CNAF center in Bologna is acting as the Italian backup center for the GERDA data, which are kept on disk and on tape. All raw data of GERDA Phase I, plus different sets of data coming from the GERDA Commissioning and from detector characterization campaigns , are stored at CNAF. The virtual organization (VO) `gerda.mpg.de` has been made available for GERDA to access the GRID resources at CNAF and all files are registered in the corresponding catalog. Currently the amount of disk usage of these data is around 9 TB; Phase I data account for about 4 TB. Higher-level TierX files are not stored at the moment because they are always reproducible by re-processing the initial raw files.

The complete suite of GERDA analysis software is also installed at CNAF. Several test were performed using data from GERDA Phase I, and new tools are being developed for the analysis of Phase II data. It is expected that data from GERDA Phase II, including commissioning data and calibration results will also be store at CNAF. The anticipated amount of data for the entire GERDA Phase II is about 20÷30 TB. The increase with respect to the Phase I is mainly due to the read-out of the LAr instrumentation and to the higher number of active HPGe detectors. All files will be registered to the catalog of the GERDA VO.
In the next few months, CNAF resources other than storage will be employed, and specifically CPU. It is foreseen to run and store on the CNAF facilities dedicated GERDA Monte Carlo simulations based on the software MaGe [11]. Monte Carlo simulations will be required to build a background model for Phase II and to benchmark the new pulse shape discrimination algorithms that are being developed.

References

[1] K.-H. Ackermann *et al.* (GERDA Collaboration), Eur. Phys. J. C (2013) 73:2330
[2] M. Agostini *et al.* (GERDA Collaboration), Phys. Rev. Lett. 111 (2013) 122503
[3] H.V. Klapdor-Kleingrothaus *et al.* (Heidelberg-Moscow Collaboration), Eur. Phys. J. A (2001) 12:147
[4] C.E. Aalseth *et al.* (IGEX Collaboration), Phys. Rev. D 65 (2002) 092007
[5] H.V. Klapdor-Kleingrothaus *et al.*, Phys. Lett. B 586 (2004) 198
[6] M. Agostini *et al.* (GERDA Collaboration), Eur. Phys. J. C (2014) 74:2764
[7] M. Agostini *et al.* (GERDA Collaboration), Eur. Phys. J. C (2015) 75:39
[8] M. Agostini, *et al.*, J. of Phys.: Conf. Ser. 375 (2012) 042027
[9] R. Brun and F. Rademakers, Nucl. Inst. & Meth. in Phys. Res. A 389 (1997) 81
[10] M. Agostini, *et al.*, J. of Phys.: Conf. Ser. 368 (2012) 012047.
[11] M. Boswell *et al.*, IEEE Trans. Nucl. Sci. 58 (2011) 1212

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

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

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

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

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

1. The *Fermi* LAT Experiment

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

So far, the LAT has registered 400 billion trigger (1800 Hz average trigger rate). An on-board filter analyses the event topology and discards about 80%. Of the 80 billion events that were transferred to ground 400 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 2014

In 2014, 54 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 17. Also external scientists are able to analyse the *Fermi* public data, resulting in 192 external publications.

The major scientific achievements in 2014 were the observation of the extremely bright gamma-ray burst (GRB) 130427A [2, 3] (ref. [2] was chosen for the title page of the Science issue), the observations in gamma-rays of four novae establishing them as a new class of gamma-ray sources [4] which triggered a NASA Press release [5], and several papers on constraints for the annihilation of hypothetical dark matter particles [6, 7, 8]. Another discovery which triggered a NASA press release [9] was the observation in gamma rays of the multiple images of a gravitationally lensed blazar [10].

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 [11] 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 [12] are accessed through a DIRAC (Distributed Infrastructure with Remote Agent Control) [13] interface to the LAT data pipeline [14]. This setup was extensively tested in 2013 and the first months of 2014 and is in ordinary production mode since april 2014.

The jobs submitted through DIRAC constitute a substantial fraction of the submitted jobs. However, we also exploit the possibility to submit jobs directly using the grid middleware. Figure 1 shows the usage of grid resources in 2014. About 22% of the jobs were run at the INFN Tier 1 at CNAF as shown by Fig. 2. The total usage in 2014 was 2969 HS06. Assuming 10 HS06 per core, this is equivalent to about 300 CPU-years.

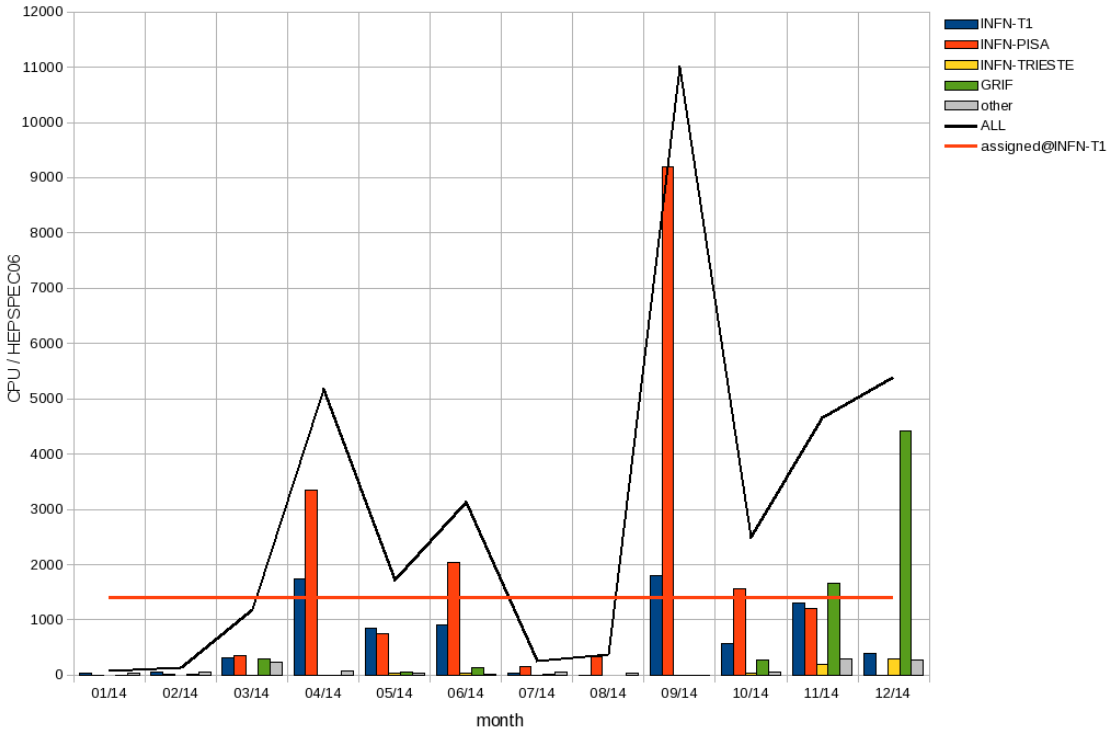


Figure 1. Usage of grid sites by the VO glast.org in 2014

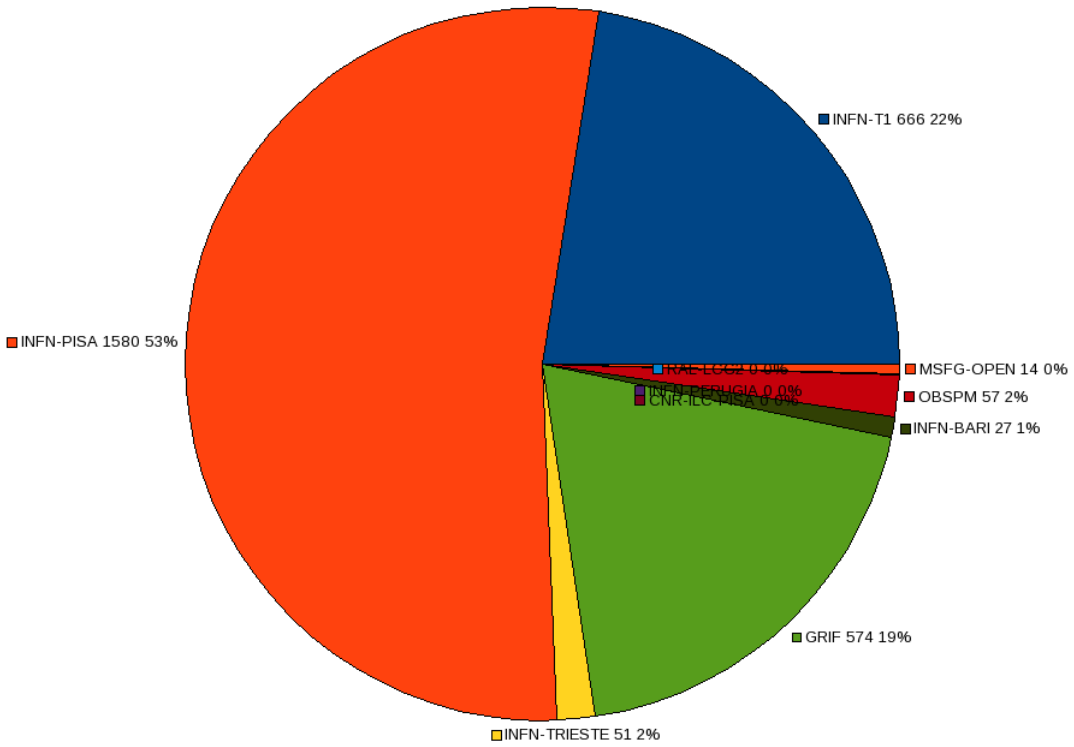


Figure 2. Cumulative usage (in HEP-SPEC06) of grid sites by the VO glast.org in 2014

4. Conclusions and Perspectives

The prototype setup based on the DIRAC framework described in the INFN-CNAF Annual Report 2013 [15] proved to be successful. In 2014 we transitioned into production mode, and several large MC tasks were run using grid resources at the INFN-CNAF Tier 1 and elsewhere. In 2015 we anticipate an increased demand for computing power in view of the massive simulation of "Pass 8" [16] backgrounds now scheduled for summer 2015.

References

[1] Atwood W B et al. 2009 *The Astrophysical Journal* **697** 1071
[2] Ackermann M et al. 2014 *Science* **343** 42
[3] Preece R et al. 2014 *Science* **343** 51
[4] Ackermann M et al. 2014 *Science* **345** 554
[5] NASA Press release 14-209 2014 July 31 http://www.nasa.gov/press/2014/july/nasas-fermi-space-telescope-reveals-new-source-of-gamma-rays/
[6] Ackermann M et al. 2014 *Phys. Rev. D* **89** 042001
[7] Drlica-Wagner A et al. 2014 *ApJ* **790** 24
[8] Albert A et al. 2014 *JCAP* **10** 023
[9] NASA Press release 14-005 2014 January 6 http://www.nasa.gov/press/2014/january/nasas-fermi-makes-first-gamma-ray-study-of-a-gravitational-lens/
[10] Abdo A A et al. 2015 *ApJ* **799** 143
[11] Dubois R 2009 *ASP Conference Series* **411** 189
[12] Arrabito L et al. 2013 CHEP 2013 conference proceedings arXiv:1403.7221
[13] Tsaregorodtsev A et al. 2008 *Journal of Physics: Conference Series* **119** 062048
[14] Zimmer S et al. 2012 *Journal of Physics: Conference Series* **396** 032121
[15] Arrabito L et al. 2014 *INFN-CNAF Annual Report 2013*, edited by L. dell'Agnello, F. Giacomini, and C. Grandi, pp. 46
[16] Atwood W B et al. 2013 *2012 Fermi Symposium: eConf Proceedings C121028* arXiv:1303.3514

LHCb Computing at CNAF

C. Bozzi
INFN Sezione di Ferrara, via Saragat 1, 44122 Ferrara, Italy
E-mail: Concezio.Bozzi@fe.infn.it

V. Vagnoni
INFN Sezione di Bologna, via Irnerio 46, 40126 Bologna, Italy
E-mail: Vincenzo.Vagnoni@bo.infn.it

Abstract. A quick overview of the LHCb computing activities is given, including the latest evolutions of the computing model. An analysis of the usage of CPU, tape and disk resources in 2014 is presented, emphasising the achievements of the INFN Tier-1 at CNAF. The expected growth of computing resources in the years to come is also briefly discussed.

1. Introduction

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

The experiment has a wide physics programme covering many important aspects of heavy flavour, electroweak and QCD physics. The core LHCb physics measurements notably include the branching ratio of the rare $B_s \to \mu^+\mu^-$ decay [2], the forward-backward asymmetry of the muon pair in the flavour changing neutral current $B^0 \to K^{*0}\mu^+\mu^-$ decay [3], the *CP* violating phase in the decay $B_s \to J/\psi\phi$ [4], the properties of radiative *B* decays [5], the determination of the unitarity triangle angle γ [6], and charmless charged two-body *B* decays [7]. More than 250 physics papers have been heretofore produced.

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

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

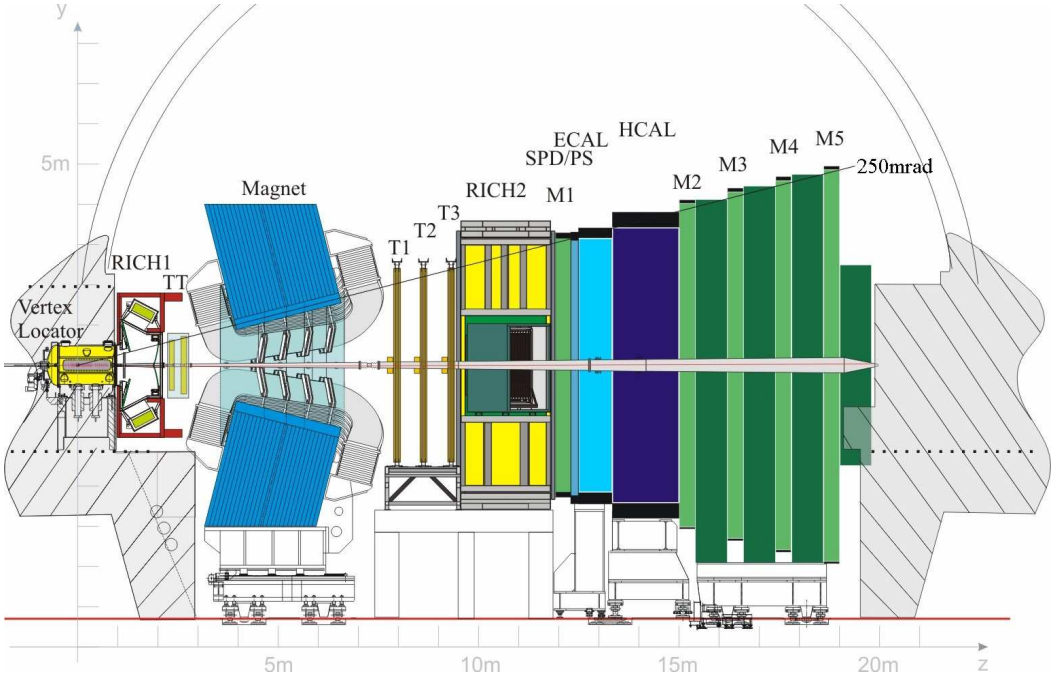


Fig. 1: Sketch of the LHCb detector.

2. Recent evolutions of the LHCb computing model

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

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

The DIRAC framework [9] for distributed computing has allowed to easily integrate non WLCG resources in the LHCb production system. In 2013, 20% of CPU resources were routinely provided by the LHCb HLT farm, and a further 6.5% by the Yandex company. This trend continued in 2014. Small scale production use has been made of virtual machines on clouds and other infrastructures, including volunteer computing through the BOINC framework. LHCb is therefore in a good position to make production use of opportunistic resources.

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

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

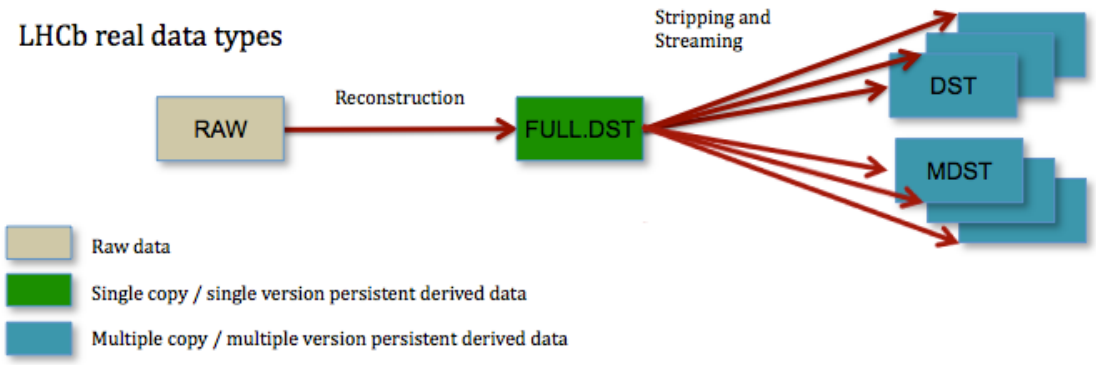


Fig. 2: The LHCb data model.

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

Tab. 1: LHCb data formats.

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

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

As a file catalog to pinpoint active file replicas, LHCb recently replaced the central LCG File Catalog (LFC) instance located at CERN, by the integrated DIRAC File Catalog (DFC) that is better adapted within the DIRAC Data Management System, as it provides natively functionalities that are not available in the LFC.

3. Resource usage in 2014

Table 2 shows the resources pledged for LHCb at the various tier levels for the 2014 period.

2014	CPU (kHS06)	Disk (PB)	Tape (PB)
Tier0	34	4.0	8.5
Tier1	127	12.7	12.3
Tier2	57	1.3	
Total WLCG	**218**	**18.0**	**20.8**

Tab. 2: LHCb 2014 pledges.

The usage of WLCG CPU resources by LHCb is obtained from the different views provided by the EGI Accounting portal. The CPU usage for Tier-0 and Tier-1s is presented in Fig. 3. The same data is presented in tabular form in Tab. 3. In 2014, a new Tier1 site started pledging resources for LHCb at the Kurchatov Institute in Russia. It must be emphasised that CNAF has provided more CPU power than any other centre, including CERN. This has been possible owing to great stability, in particular of the storage system, leading to maximal efficiency in the overall exploitation of the resources.

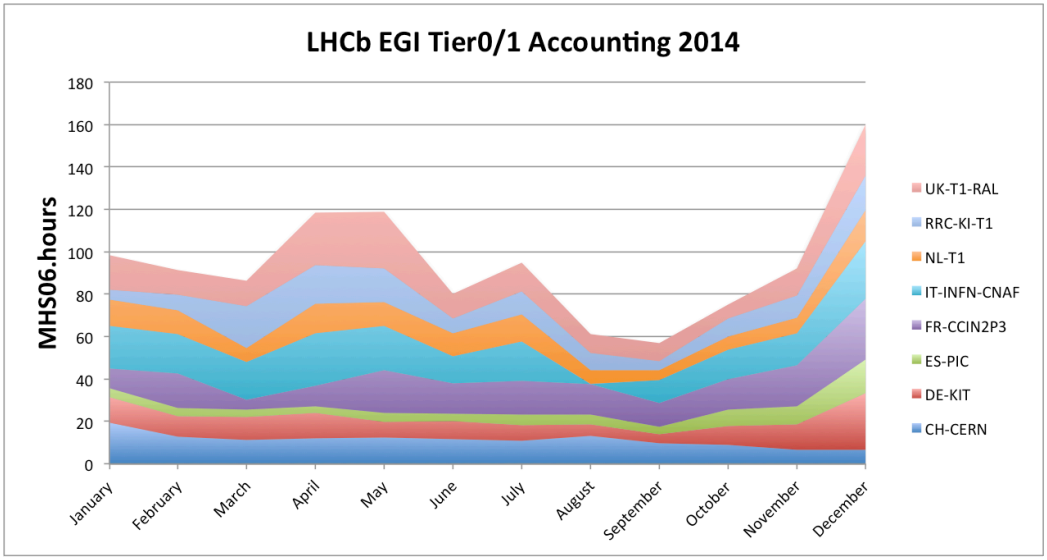


Fig. 3: Monthly CPU work provided by the Tier-0 and Tier-1s to LHCb during 2014.

<Power>	Used (kHS06)	Pledge (kHS06)
CH-CERN	15.6	34.0
DE-KIT	14.6	19.2
ES-PIC	7.8	7.1
FR-CCIN2P3	20.6	21.7
IT-INFN-CNAF	23.2	19.8
NL-T1	13.7	13.8
RRC-KI-T1	15.2	10.8
UK-T1-RAL	20.4	34.7
Total	**131.2**	**161.1**

Tab. 3: Average CPU power provided by the Tier-0 and the Tier-1s to LHCb during 2014.

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

The usage of the Storage is the most complex part of the LHCb computing operations. Not much new tape storage was necessary in 2014, as the new data coming in were due to the archival of MC production and stripping cycles. The total volume added to the tape archive was about 800 TB. The total tape occupancy as of February 6th 2015 is 15.7 PB, 6.1 PB of which are used for RAW data, 4.5 PB for FULL.DST, 5.1 PB for archived. The used tape space at CERN and Tier-1s at the end of March 2015 does not exceed the 2014 pledges. Table 5 shows the situation of disk storage resources at the Tier-0 and Tier-1s at the end of January 2015. Despite the lower disk pledges, CNAF has been the second Tier-1 in terms of disk storage made available to LHCb.

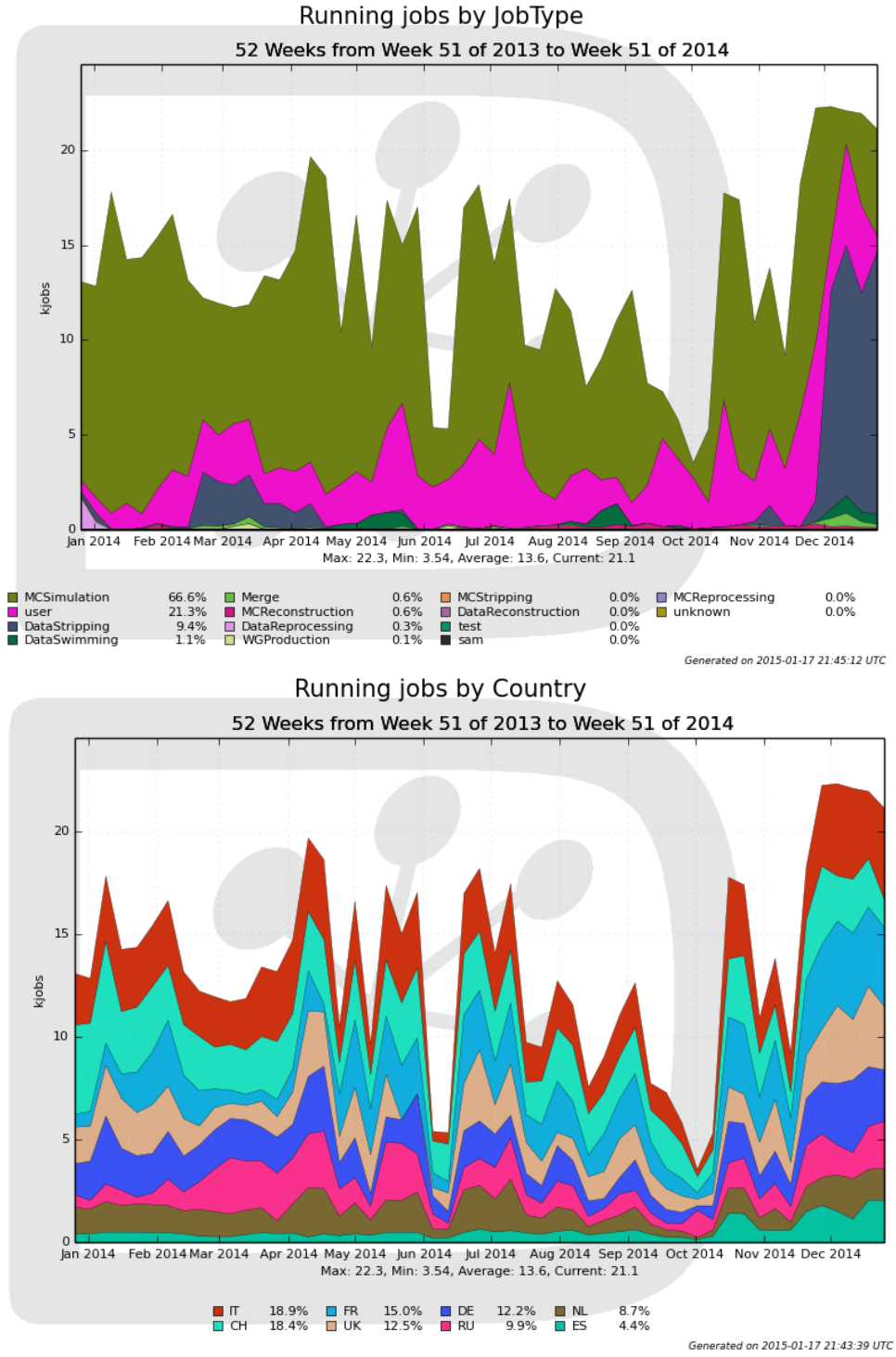


Fig. 4: Usage of LHCb resources at Tier-0 and Tier-1s during 2014. The top plot shows the usage of resources for the various activities, whereas the bottom plot shows the contributions from the different countries.

Disk (PB)	CERN	CNAF	GRIDKA	IN2P3	PIC	RAL	RRCKI	SARA	Tier1s
LHCb accounting	**3.43**	**1.74**	**1.56**	**1.21**	**0.66**	**2.23**	**0.06**	**0.96**	**8.42**
Disk used	3.65	1.76	1.56	1.21	0.65	2.28	0.06	0.96	8.49
Disk free	0.45	0.64	0.57	0.22	0.13	1.30	0.14	0.26	3.27
Stage area (used+free)	0.41	0.77	0.14	0.03	0.01	0.20	0.02	0.03	1.20
Disk total	**4.52**	**3.18**	**2.26**	**1.46**	**0.80**	**3.78**	**0.22**	**1.25**	**12.96**
Pledge 2014	***4.00***	***2.52***	***2.34***	***1.80***	***0.75***	***3.62***	***0.40***	***1.20***	***12.74***

Tab. 5: Situation of disk storage resource usage at the end of January 2015, available and installed capacity, and 2014 pledge.

In summary, the usage of computing resources in 2014 has been quite smooth for LHCb.
An incremental stripping campaign took place during spring, and run for a couple of months. A "swimming" activity was also run during spring. The reprocessing of a small dataset taken in 2010 was performed in summer. A legacy stripping of the Run1 dataset started in December and continued in January 2015.

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

The average CPU power achieved in 2014 is in line with the expectation. The sharing between the various WLCG Tiers was somewhat different with respect to the estimated needs. The usage of fewer resources than anticipated at the Tier0 and Tier1s was compensated by the possibility of utilizing more resources at Tier2s than pledged by the funding agencies, or at other sites.

Tape storage was not really a concern, tape usage was restricted to physics data archive. Staging data from tape for the incremental restripping also went rather smoothly. Due to the availability of disk cache, it was possible to pre-stage a significant amount of data from tape for the Run 1 legacy stripping, thus alleviating the operational burden and the tape-to-disk bandwidth requirements.

Disk storage was also not a major concern, thanks to significant efforts that went into deploying the mDST format, first used in the legacy stripping campaign of Run1 data, and to the analysis of data popularity. LHCb is successfully using disks at Tier2s, which will allow to overcome the anticipated shortfall in disk at Tier1s in the coming years.

Both disk and tape resources pledged by the funding agencies were adequate for the LHCb computing activities until the end of the 2014 WLCG year in March 2015.
The usage of datasets produced for physics analysis is constantly monitored, the subsequent analysis of data popularity having allowed LHCb to free a significant amount of disk space. Progress is being made on the implementation of a dataset classifier that, based on all dataset metadata and popularity history, allows to classify datasets into those that are likely to be used in a given interval and those that are not. This would allow us to tune the number of replicas of a dataset and remove the least popular ones.

4. Expected resource growth

In terms of CPU requirements, Tab. 6 presents for the different activities the CPU work estimates for 2015, 2016 and 2017. Note that in this table there are no efficiency factors applied: these are resource

requirements assuming 100% efficiency in using the available CPU. The last row shows the power averaged over the year required to provide this work, after applying the standard CPU efficiency factors.

CPU Work in WLCG year (kHS06.years)	**2015**	**2016**	**2017**
Prompt Reconstruction	19	31	26
First pass Stripping	8	13	11
Full Restripping	8	20	11
Incremental Restripping	0	4	10
Simulation	134	153	207
VoBoxes and other services		4	4
User Analysis	17	20	24
Total Work (kHS06.years)	**186**	**246**	**293**
Efficiency corrected average power (kHS06)	**220**	**291**	**348**

Tab. 6: Estimated CPU work needed for the various LHCb activities in 2015-2017. Proton physics.

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

Power (kHS06)	**Request 2015**	**Forecast 2016**	**Forecast 2017**
Tier 0	44	51	62
Tier 1	123	156	191
Tier 2	52	88	107
Total WLCG	**219**	**295**	**360**
HLT farm	10	10	10
Yandex	10	10	10
Total non-WLCG	**20**	**20**	**20**

Tab. 7: CPU power requested at the different Tiers in 2015-2017. Minimal additional resources with respect to the ones shown in Tab. 6 are requested for heavy ions physics in 2016 and 2017.

Tables 8 and 9 present, for the different data classes, the forecast total disk and tape space usage at the end of the years 2015-2017. These disk and tape estimates are then broken down into fractions to be provided by the different Tiers. These numbers are shown in Tables 10 and 11. These tables also include small additional resources needed for heavy ions physics in 2016 and 2017. These additional resources are quoted in Table 12. As can be seen the increase in disk storage can be managed to fit inside a reasonable growth envelope by adjustments in the details of the processing strategy. On the other hand, the growth in the tape storage requirement is more challenging but largely incompressible: in Tab. 9 one can see that the major part of the increase is due to raw data that, if not recorded, is lost.

LHCb Disk storage usage forecast (PB)	2015	2016	2017
Stripped Real Data	7.3	13.1	15.3
Simulated Data	8.2	6.9	10.4
User Data	0.9	1.0	1.1
MDST.DST	1.5	1.9	0.0
FULL.DST	3.3		
RAW and other buffers	0.4	1.2	0.9
Other	0.2	0.2	0.2
Total	**21.7**	**24.3**	**27.9**

Tab. 8: Breakdown of estimated disk storage usage for different categories of LHCb data (proton physics)

LHCb Tape storage usage forecast (PB)	2015	2016	2017
Raw Data	12.6	21.7	34.5
FULL.DST	8.7	15.2	20.7
MDST.DST	1.8	5.2	7.9
Archive - Operations	8.6	11.6	15.0
Archive – Data preservation	0.0	6.0	9.2
Total	**31.7**	**59.7**	**87.3**

Tab. 9: Breakdown of estimated tape storage usage for different categories of LHCb data (proton physics)

LHCb Disk (PB)	2015 Request	2016 Forecast	2017 Forecast
Tier0	6.7	7.6	9.1
Tier1	12.5	13.5	15.0
Tier2	2.5	4.0	5.5
Total	**21.7**	**25.2**	**29.6**

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

LHCb Tape (PB)	2015 Request	2016 Forecast	2017 Forecast
Tier0	10.4	20.6	30.9
Tier1	21.3	42.1	62.2
Total	**31.7**	**62.7**	**93.1**

Tab. 11: LHCb tape request for each Tier level (proton + heavy ions physics)

Resources for heavy ion running	2016 Request	2017 Request
CPU (kHS06)	24	32
Disk (PB)	0.9	1.7
Tape (PB)	3.0	5.7

Table 12: LHCb requests for heavy ions physics in 2016 and 2017.

5. Conclusions

A description of the LHCb computing activities has been given, with particular emphasis on the evolutions of the computing model, on the usage of resources and on the forecasts of resource needs until 2017. It has been shown that CNAF has been in 2014 the most important LHCb computing centre in terms of CPU power made available to the collaboration. This is a great achievement, which has been possible due to the hard work of the CNAF Tier-1 staff, to the overall stability of the centre and to the friendly collaboration between CNAF and LHCb people. The importance of CNAF within the LHCb distributed computing infrastructure has been recognised by the LHCb computing management in many occasions.

References

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

NA62 computing at CNAF

P Valente
INFN Roma,
P.le Aldo Moro, 2, I-00185, Rome (Italy)

E-mail: paolo.valente@roma1.infn.it

Abstract. The NA62 experiment at CERN SPS aims at measuring with a 10% precision the branching fraction of the very rare decays $K^+ \rightarrow \pi^+ \nu$ anti-ν. In 2014 the baseline NA62 detector was deployed and commissioned, running with beam from week 41 to week 50 with commissioning data and a sample of K^+ decays, useful to verify the physics sensitivity. A data sample in excess of 100 TB of raw data has been recorded at the Tier-0 and promptly reconstructed, and is now ready for distribution to Tier-1 centers, for further processing, such as calibration, merging, filtering, data quality, etc., in order to enable more refined physics studies.

1. NA62 computing model

The NA62 experiment at CERN SPS aims at measuring with a 10% precision the branching fraction of the very rare decays $K^+ \rightarrow \pi^+ \nu$ anti-ν. This requires collecting 100 signal events (in two years of data taking), among an enormous (10^{13}) sample of Kaon decays with a detector based on several, redundant veto detectors and particle-id. Three different trigger levels have the task of reducing this rate to an acceptable data band-width to be recorded on tape, at the same time keeping the possibility of acquiring other interesting event topologies, in order to perform the other physics measurements, exploiting less rare Kaon decay channels. The redundancy of the detectors should allow the analysis to over-constrain the few candidate events and to discriminate them among the much more abundant background. A detailed description of the NA62 setup, trigger and DAQ, infrastructure, services, etc. can be found in [1].

The NA62 schematic computing model is described in [2] and is essentially based on recording and a first prompt reconstruction of RAW data at the Tier-0, and then distribution and re-processing and user analysis at the remote centres. Up to now three centers have been identified outside CERN: CNAF, RAL and UCL Louvain. The TDAQ has the task of reducing the number of events from the physics rate of about 10 MHz, down to a level acceptable for recording, of a few kHz. It is based on the concept of handling data in a completely digital format and of holding all the information acquired from the detectors in buffers, until the trigger decision is elaborated from the data. After a level-0 trigger decision taken by a dedicated processor working on reduced information from the detectors (trigger primitives), two subsequent trigger levels further reduce the data working on the data stored in the online acquisition farm. Due to the high bandwidth needed for the acquisition of the about 13,000 cells of the liquid Krypton calorimeter, this data is downloaded to the farm only upon a level-1 trigger decision. A further selection can be done at the level-2 on the fully assembled events, and raw data is

stored on the farm storage by dedicated PC (mergers) of the farm. From there, the Central Data Recording system transfers RAW files (one for SPS spill) to the Tier-0 for logging to tape and processing.

The total amount of RAW data expected for a "typical" data taking year (200 days, 60% efficiency, 32% duty-cycle of SPS to the NA62 Kaon beam line) is of the order of 1 PB.

2. 2014 activities

In 2014 the baseline NA62 detector was deployed and commissioned, running with beam from week 41 to week 50 with commissioning data and a sample of K^+ decays, useful to verify the physics sensitivity [3]. A data sample in excess of 100 TB of raw data has been recorded at the Tier-0 and promptly reconstructed, and is now ready for distribution to Tier-1 centers, for further processing, such as calibration, merging, filtering, data quality, etc., in order to enable more refined physics studies.

The dedicated disk area has been setup at CNAF for testing staging and transfer of data from CERN; also the CPU share has been prepared in order to test and run the NA62 Monte Carlo with the production system for the virtual organization setup in the Grid-PP.

[1] F. Hahn et al. 2010, NA62 Technical Design Document, NA62-10-07
[2] P. Valente et al. 2014, A Computing Model for NA62, NA62-14-03
[3] NA62 Collaboration 2015, 2015 NA62 Status Report to the CERN SPSC, CERN-SPSC-2015-012/SPSC-SR-157

OPERA Experiment

S. Dusini, on behalf of the OPERA Collaboration
INFN, Sez. di Padova, via Marzolo 8, 35131 Padova, Italy
E-mail: stefano.dusini@pd.infn.it

1. The OPERA experiment

OPERA [1, 2] is a long-baseline neutrino experiment designed to observe for the first time the $\nu_\mu \to \nu_\tau$ oscillation in direct appearance mode through the detection of the production of the corresponding τ lepton in the CNGS ν_μ beam [3] over a baseline of 730 km. OPERA is an hybrid detector consisting of two instrumented targets, each followed by a muon spectrometer. Each target is a succession of walls of "ECC bricks" interleaved with planes of scintillator strips. The ECC bricks are made of 56 1-mm thick lead plates providing the mass interleaved with emulsion films with a micrometric resolution. The total number of "ECC-bricks" is about 150000 for a total target mass of ~ 1.2 kton. The OPERA detector is located in the Gran Sasso underground laboratory (LNGS) in Italy. For a detailed description of the detector we refer to [4].
The CNGS beam has run for five years, from 2008 till the end of 2012, delivering a total of 17.97×10^{19} protons on target yielding 19505 neutrino interactions recorded in the OPERA targets. The analysis of these events is still in progress.

2. Recent OPERA results

In this section we rrport the most important results obtained by the OPERA collaboration during the 2014.

2.1. ν_τ appearance search

So far the OPERA Collaboration observed four ν_τ candidates: two candidate events are in the $\tau \to h$ channel [5, 6], one in the $\tau \to 3h$ channel [7] and one in the $\tau \to \mu$ channel [8]. In the data sample analysed so far the expected ν_τ signal in all decay channels is 2.11 ± 0.42 events (for $\Delta m^2_{32} = 2.32 \times 10^{-3}$ eV^2 [9] and $sin^2(2\theta_{23}) = 1$), while the total expected background is 0.233 ± 0.041 events. The significance of the observed four ν_τ candidate events has been estimated with two methods. The first method combines the p-values of the single channels (p_i for $i = h, 3h, \mu, e$) according to Fisher's rule into the estimator $p^* = p_h p_{3h} p_\mu p_e$. In order to take into account the systematic uncertainties of the backgrounds, 100 sets of randomised backgrounds are generated. A mean p-value of 1.24×10^{-5} is obtained by a Monte Carlo calculation of the tail probability corresponding to the observed value of p^*. With this method the absence of the of signal can be excluded with a significance of 4.2σ.
The second method is based on the likelihood ratio test [10] with the likelihood function defined as

$$\mathcal{L}(\mu) = \prod_{i=1}^{4} \frac{e^{-(\mu s_i + b_i)}(\mu s_i + b_i)^{n_i}}{n_i!}, \quad i = h, 3h, \mu, e \tag{1}$$

where the parameter μ determines the strength of the signal process ($\mu = 0$ correspond to the background-only hypothesis), s_i and b_i are the numbers of expected signal and background events, n_i the number of observed events. The systematic uncertainties of the backgrounds were taken into account in a similar way as above. The p-value of the observed statistic is 1.03×10^{-5} corresponding to a significance of 4.2σ for the exclusion of the null hypothesis.
In OPERA the rate of ν_τ CC interactions is given by

$$N_\tau = A \int_{E_{th} \simeq 3.5 \text{ GeV}} \phi_{\nu_\mu}(E) P_{\nu_\mu \to \nu_\tau}(E) \sigma_\tau^{CC}(E) \epsilon(E) dE \tag{2}$$

where A is a normalization constant proportional to the detector mass, $\phi_{\nu_\mu}(E)$ in the ν_μ flux at the detector location, $P_{\nu_\mu \to \nu_\tau}(E)$ is the oscillation probability, $\sigma_\tau^{CC}(E)$ is the cross-section for ν_τ^{CC} interactions and $\epsilon(E)$ is the ν_τ detection efficiency. The lower limit of the integral is given by the energy threshold for τ lepton production. On the CNGS beam at the distance of 730 Km the effective two state oscillation probability can be approximated by expanding the oscillation term in power series, and Eq. (2) can be approximated as

$$N_\tau \simeq 1.61 \; A \; \sin^2(2\theta_{23}) \left(\Delta m^2_{32}[\text{eV}^2]\right)^2 L^2[\text{km}] \int_{E_{th} \simeq 3.5 \text{ GeV}} \phi_{\nu_\mu}(E) \sigma_\tau^{CC}(E) \epsilon(E) \frac{dE}{E^2} \; . \tag{3}$$

In this approximation N_τ varies as $\left(\Delta m^2_{32}\right)^2$ and the number of observed ν_τ candidates can be used to measure Δm^2_{32}. Given the 4 observed events and the expected background of (0.233 ± 0.041) events, the confidence interval of Δm^2_{32} has been estimated with the Feldman-Cousins method [11], assuming maximal mixing. The systematic uncertainties of signal and background expectations are taken into account to marginalise the likelihood function used for the ordering principle. The 90% confidence interval for Δm^2_{32} is $[1.8, \; 5.0] \times 10^{-3}$ eV^2. An alternative method using a Bayesian approach [9] with a flat prior on Δm^2_{32} gives for the credible interval of Δm^2_{32} the values $[1.9, 5.0] \times 10^{-3}$ eV^2.

2.2. Search for sterile neutrinos in $\nu_\mu \to \nu_\tau$ oscillations
The study of $\nu_\mu \to \nu_\tau$ oscillation over a long-base line can also be use to set limits on the existence of sterile neutrinos [12]. In the simplest extension of the three-neutrino mixing model with the addition of one massive neutrino ν_4 with mass m_4, the oscillation probability is a function of the 4×4 mixing matrix U and of the three squared mass differences. Defining $C = 2|U_{\mu 3}||U_{\tau 3}|$, $\Delta_{ij} = 1.27 \; \Delta m^2_{ij} \; L/E$ (i,j = 1,2,3,4), $\phi_{\mu\tau} = \; Arg(U_{\mu 3} U^*_{\tau 3} U^*_{\mu 4} U_{\tau 4})$ and $\sin 2\theta_{\mu\tau} = 2|U_{\mu 4}||U_{\tau 4}|$, the $\nu_\mu \to \nu_\tau$ oscillation probability *P(E)* can be parametrised as:

$$\begin{aligned} P(E) \;\; &= \;\; C^2 \; \sin^2 \Delta_{31} + \sin^2 2\theta_{\mu\tau} \sin^2 \Delta_{41} + \frac{1}{2} C \sin 2\theta_{\mu\tau} \cos\phi_{\mu\tau} \sin 2\Delta_{31} \; \sin 2\Delta_{41} \\ &- \;\; C \sin 2\theta_{\mu\tau} \sin\phi_{\mu\tau} \sin^2 \Delta_{31} \sin 2\Delta_{41} + 2C \sin 2\theta_{\mu\tau} \cos\phi_{\mu\tau} \sin^2 \Delta_{31} \sin^2 \Delta_{41} \\ &+ \;\; C \sin 2\theta_{\mu\tau} \sin\phi_{\mu\tau} \sin 2\Delta_{31} \sin^2 \Delta_{41} \end{aligned} \tag{4}$$

where Δm^2_{31} and Δm^2_{41} are expressed in eV^2, L in km and E in GeV. Given the long baseline and the average CNGS neutrino energy, *P(E)* is independent of Δm^2_{21}. The terms proportional to $\sin\phi_{\mu\tau}$ are CP-violating, while those proportional to $\sin 2\Delta_{31}$ are sensitive to the mass hierarchy of the three standard neutrinos, normal ($\Delta m^2_{31} > 0$) or inverted ($\Delta m^2_{31} < 0$). For $\Delta m^2_{41} \gtrsim 1$ ev^2, which is the region indicated by the neutrino oscillation anomalies [13], matter effects are negligible on the CNGS beam. In this domain, taking into account the finite energy

resolution of OPERA detector, $\sin 2\ \Delta_{41}$ and $\sin^2 \Delta_{41}$ average to 0 and $\frac{1}{2}$, respectively. The oscillation probability (4) can thus be approximate to:

$$P(E) = C^2 \sin^2 \Delta_{31} + \frac{1}{2} \sin^2 2\theta_{\mu\tau} + C \sin 2\theta_{\mu\tau} \cos \phi_{\mu\tau} \sin^2 \Delta_{31} + \frac{1}{2} C \sin 2\theta_{\mu\tau} \sin \phi_{\mu\tau} \sin 2\Delta_{31}. \quad (5)$$

The number of observed ν_τ events is compared to the expectation from (5) using the asymptotic χ^2 distribution of the log likelihood ratio test statistics: $q = -2\ \ln(\widetilde{L}(\phi_{\mu\tau}, \sin^2 2\theta_{\mu\tau})/L_0)$, where $L_0 = e^{-n}\ n^n/n!$ and $\widetilde{L}(\phi_{\mu\tau}, \sin^2 2\theta_{\mu\tau})$ is the profile likelihood obtained by maximising $L(\phi_{\mu\tau}$,$\sin^2 2\theta_{\mu\tau}$, C^2) over C^2. In figure 1(a) the 90% CL exclusion limits are presented for both normal and inverted mass hierarchies in the parameter space of $\phi_{\mu\tau}$ vs $\sin^2 2\theta_{\mu\tau}$.

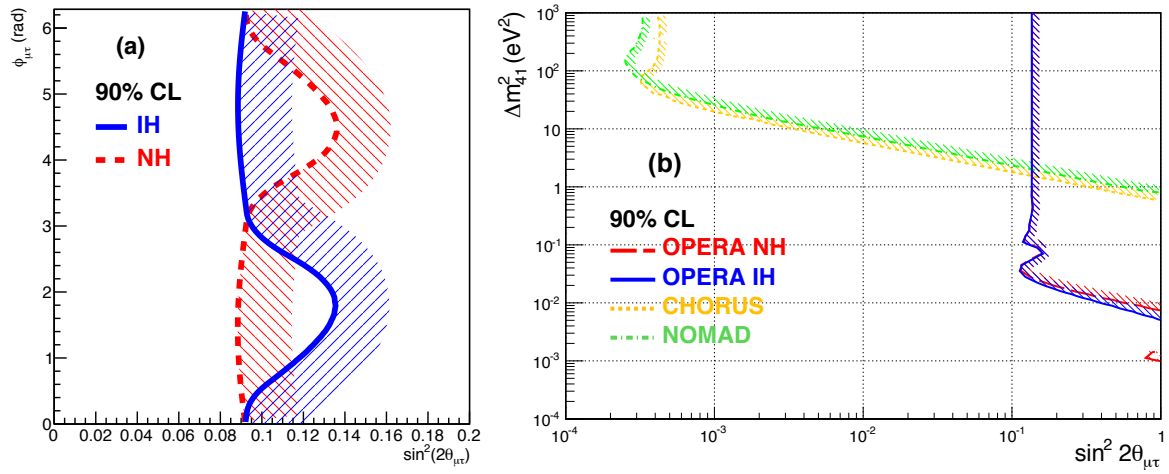


Figure 1. (a) 90% CL exclusion limits in the $\phi_{\mu\tau}$ vs $\sin^2 2\theta_{\mu\tau}$ parameter space for normal (NH, dashed red) and inverted (IH, solid blue) hierarchies assuming $\Delta m^2_{41} > 1$ eV^2. Bands are drawn to indicate the excluded regions. (b) OPERA 90% CL exclusion limits in the Δm^2_{41} vs $\sin^2 2\theta_{\mu\tau}$ parameter space for the normal (NH, dashed red) and inverted (IH, solid blue) hierarchy of the three standard neutrino masses. The exclusion plots by NOMAD [14] and CHORUS [15] are also shown. Bands are drawn to indicate the excluded regions.

To extend the search for a possible fourth sterile neutrino down to small Δm^2_{41} values, the likelihood has been computed using GloBES software [16], which takes into account the non-zero Δm^2_{21} value and also matter effects. The likelihood has been profiled also on the Δm^2_{31} value. More details on the analysis are available in [17]. In Figure 1(b) the 90% CL exclusion plot is reported in the Δm^2_{41} vs $\sin^2 \theta_{\mu\tau}$ parameter space.

3. The Computing Model of the OPERA Simulation Software

In this section we describe the computing model of the OPERA Monte Carlo (MC) simulation. The OPERA simulation framework, called OpRelease, is a set of C++ packages running mainly on Linux platforms. Figure (2) shows a schematic description of the OPERA simulation software.

The OpRelease packages are managed by the CMT [18] system and are stored in the CERN SVN repository [19]. We are planing to migrate the SVN server to CNAF to allow the access to SVN server by none CERN users. CMT is also used to manage external libraries (like ROOT [20], CLHEP [21], Pythia [22], etc.) through the CMT Interface package to hide the system specific properties of those libraries from the OPERA software. The ROOT library is used extensively

for data storage, geometrical description of the detector and as an interface with the Monte Carlo simulation software through the Virtual Monte Carlo (VMC) package [23].
In the spirit of the GAUDI framework of LHCb and ATLAS, the chain of various algorithms used in the simulation and reconstruction programs are steered by a dedicated package called OpAlgo. This allows, together with the package OpIO which manages the I/O, to decouple the algorithms from the particular data storage model (ROOT Trees, ORACLE DB, ASCII files).
The geometrical description of the detector is managed by the OpGeom package using ROOT-TGeoManager classes. The result is a persistent and simulation-independent geometry model which can be accessed by any program sharing the same geometrical information among the various packages of OpRelease.
The detector simulation software, OpSim, is constructed around the VMC package which allows switching between GEANT3 and GEANT4 transportation code while keeping the same external geometry and event steering routines. The simulation of a neutrino interaction is performed externally to OpSim in order to easily switch between different interaction generators. Neutrino interactions are fed to OpSim through a so called beam file.
The reconstruction of events is performed in two steps: in the first step an event is reconstructed using only the information of the electronic detector (ED), in the second step bricks selected by the brick-finding algorithm are reconstructed using hits recorded in the emulsion.
The reconstruction of the electronic detector part of an event is done by a sequence of algorithms managed by the OpRec package. First recorded hits are processed by a pattern recognition algorithm and sub-samples of hits are grouped into three dimensional (3D) tracks. A 3D-track is tagged as a muon if the product of its length and the density along its path is larger than 660 g/cm^2. The momentum of 3D-tracks is calculated from their bending in the spectrometer magnetic field and/or from their range with a Kalman filter-based reconstruction algorithm. A classifier algorithm, called OpCarac [24], is applied to select neutrino interactions inside the OPERA target. These events are further processed by a brick-finding algorithm [25]. The topology and the energy deposition in the OPERA Target Tracker scintillator strips, as well as the muon track information (when available) are used to define a three-dimensional probability density map for the vertex position. This probability is integrated over the volume of the bricks to select and reconstruct the bricks with the highest probability of containing the neutrino interaction.

The reconstruction of the simulated emulsion data is performed by the OpEmuRec package. This software has been designed to simulate the scanning and reconstruction procedure followed in the emulsion scanning laboratories. For this purpose, OpEmuRec implements an interface to SySal.NET [26] and FEDRA [27] *off-line* reconstruction and analysis software used in the European Scanning System (ESS) [28]. SySal.NET is written in *C#* and it runs natively on Windows machines and using Mono [29] on Linux machines. The MC data production is controlled by an ORACLE database through *tcsh* scripts. Configuration files and program settings used to run the various packages are stored in the DB. Each operation is recorded in the DB which allows subsequent tracing of all operations performed for the production of a particular dataset. The MC data files produced are recorded in the DB with their name, location and a link to the operations performed to produce those files.

4. The OPERA activity at CNAF in 2014

The OPERA activity at CNAF started at the end of 2013 with the aim of setting up a second MC production site besides the one at the IN2P3 Computer Center in Lyon. The CNAF site went in production in late summer 2014. The managment of the jobs is done using custom *tcsh* scripts, while the bookkeeping of the produced events is managed with the OPERA Oracle data base.

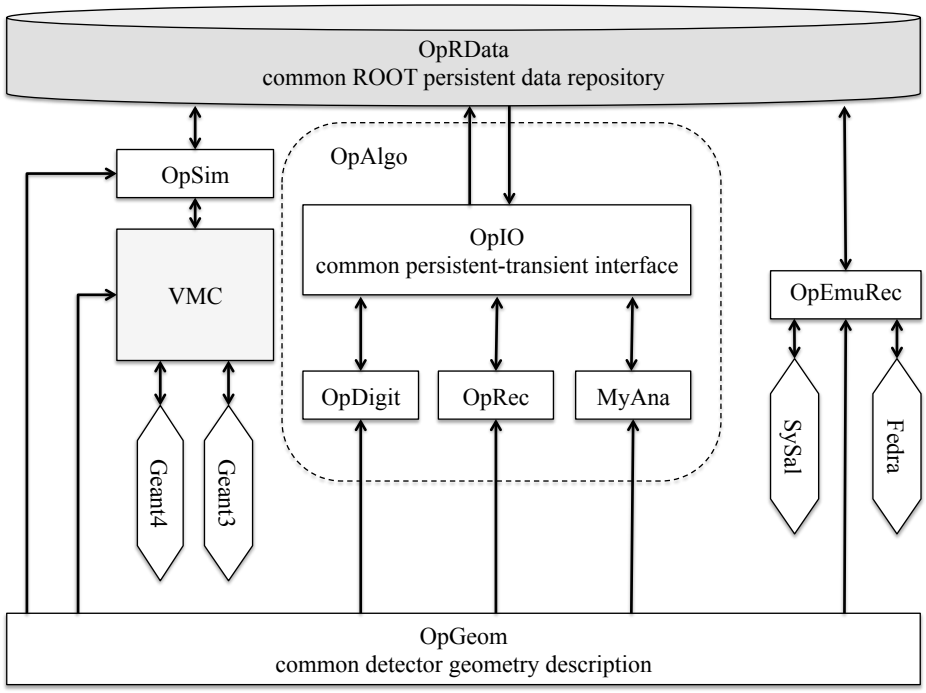


Figure 2. Schema of the `OpRelease` software.

Beside the standard OPERA MC production the CNAF production farm have been used for the statistical analysis of the OPERA results presented in section 2.1. In particular at the end of July, thanks to a temporary share upgrade for OPERA ($\sim$ 100 nodes), some toy Monte Carlo simulations (each lasting 15 minutes) have been run using ROOT/RooStats combined with Proof On Demand resource manager. The aim was the correct evaluation of the significance of the recent OPERA tau neutrino appearance analysis with the full treatment of nuisance parameters which requires simulations and cannot by performed by approximated formulas.
The CNAF farm was also extensively used for the analysis reported in section 2.2 for the search of sterile neutrinos. The Monte Carlo simulation with the General Long Baseline Experiment Simulator (GLoBES) software as well as the statistical analysis have been done entirely at the CNAF computing center.

References

[1] OPERA collaboration, M. Guler et al., *An appearance experiment to search for $\nu_\mu \to \nu_\tau$ oscillations in the CNGS beam: experimental proposal*, CERN-SPSC-2000-028, LNGS P25/2000.
[2] OPERA collaboration, M. Guler et al., *Status Report on the OPERA experiment*, CERN/SPSC 2001-025, LNGS-EXP 30/2001 add. 1/01.
[3] Ed. K. Elsener, *The CERN Neutrino beam to Gran Sasso (Conceptual Technical Design)*, CERN 98-02, INFN/AE-98/05.
R. Bailey et al., *The CERN Neutrino beam to Gran Sasso (NGS) (Addendum to report CERN 98-02, INFN/AE-98/05)*, CERN-SL/99-034(DI), INFN/AE-99/05.
[4] OPERA collaboration, R. Acquafredda et al., *JINST* **4** (2009) P04018.
[5] N. Agafonova *et al.* [OPERA Collaboration], Phys. Lett. B **691** (2010) 138
[6] N. Agafonova *et al.* [OPERA Collaboration], PTEP **2014** (2014) 10, 101C01
[7] N. Agafonova *et al.* [OPERA Collaboration], JHEP **1311** (2013) 036
[8] N. Agafonova *et al.* [OPERA Collaboration], Phys. Rev. D **89** (2014) 5, 051102.
[9] J. Beringer *et al.* [Particle Data Group Collaboration], Phys. Rev. D **86** (2012) 010001.

[10] G. Cowan, K. Cranmer, E. Gross and O. Vitells, Eur. Phys. J. C **71** (2011) 1554 [Eur. Phys. J. C **73** (2013) 2501]
[11] G. J. Feldman and R. D. Cousins, Phys. Rev. D **57** (1998) 3873 [physics/9711021 [physics.data-an]].
[12] B. Pontecorvo, Sov. Phys. JETP **26** (1968) 984 [Zh. Eksp. Teor. Fiz. **53** (1967) 1717].
[13] J. Kopp, P.A.N. Machado, M. Maltoni and T. Schwetz, [arXiv:1303.3011v3].
[14] P. Astier *et al.* [NOMAD Collaboration], Nucl. Phys. B **611** (2001) 3
[15] E. Eskut *et al.* [CHORUS Collaboration], Nucl. Phys. B **793** (2008) 326
[16] P. Huber, M. Lindner and W. Winter, Comput. Phys. Commun. 167 (2005) 195.
P. Huber, J. Kopp, M. Lindner, M. Rolinec and W. Winter, Comput. Phys. Commun. 177 (2007) 432.
[17] S. Dusini et al., "Search for sterile neutrino mixing in the $\nu_\mu \to \nu_\tau$ appearance channel with the OPERA detector", http://operaweb.lngs.infn.it/Opera/notes/notes/175/1/.
[18] See http://www.cmtsite.org.
[19] https://svnweb.cern.ch/cern/wsvn/opera
[20] R. Brun and F. Rademakers, Nucl. Instrum. Meth. **A 389** (1997) 81;
see also http://root.cern.ch/
[21] http://proj-clhep.web.cern.ch/proj-clhep/
[22] T. Sjöstrand, S. Mrenna and P. Skands, Comput. Phys. Comm. **178** (2008) 852 [arXiv:0710.3820]
[23] I. Hrivnacova *et al.* [ALICE Collaboration], eConf C **0303241** (2003) THJT006
see also http://root.cern.ch/drupal/content/vmc
[24] A. Bertolin et al., OPERA public note n.100 (2009), http://operaweb.lngs.infn.it/Opera/publicnotes/note100.pdf
[25] A. Chukanov et al., OPERA public note n. 162 (2013), http://operaweb.lngs.infn.it/Opera/publicnotes/note162.pdf
[26] E. Barbuto, C. Bozza, and C. Sirignano, OPERA note 78 4-06-2006, (2006).
[27] V. Tioukov, I. Kreslo, Y. Petukhov, and G. Sirri. The FEDRAFramework for emulsion data reconstruction and analysis in the OPERA experiment. Nucl. Instrum. and Meth. **A 559** (2006) 103.
[28] L. Arrabito, E. Barbuto, C. Bozza, S. Buontempo, L. Consiglio, D. Coppola, M. Cozzi and J. Damet *et al.*, Nucl. Instrum. Meth. A **568** (2006) 578
[29] http://www.mono-project.com/

Advanced Virgo Computing at INFN CNAF

P Astone[1], A Colla[1], G Debreczeni[2] and G Gemme[3]

[1] INFN, Sezione di Roma, I-00185 Roma, Italy
[2] Wigner RCP, RMKI, H-1121 Budapest, Hungary
[3] INFN, Sezione di Genova, I-16146 Genova, Italy
Virgo experiment `http://www.virgo.infn.it/`

E-mail: `gianluca.gemme@ge.infn.it`

Abstract. A brief description of the Virgo experiment is given, and few highlights of the activity in 2014 are presented. The computing model of the experiment is outlined and particular emphasis is given to the crucial role of the resources provided by CNAF in the framework of the overall strategy.

1. Introduction

Virgo is gravitational wave detector based on a kilometer-scale Michelson interferometer, which took scientific data from May 2007 to September 2011, in four scientific runs (VSR1, VSR2, VSR3, VSR4).

Advanced Virgo is being realized trough the upgrade of the original Virgo interferometer, built by the Centre National de la Recherche Scientifique (CNRS, France) and by the Istituto Nazionale di Fisica Nucleare (INFN, Italy). The realization of Advanced Virgo is also based on the fundamental contribution from the National Institute for Nuclear Physics and High Energy Physics (Nikhef, Netherlands) and on the support coming from the Hungarian Academy of Sciences and from the Polish National Science Centre. A detailed description the main features of the upgrade and of the status of the construction can be found in [1]

Advanced Virgo will be part of the world network of second generation interferometers for gravitational wave detection, which includes the two LIGO detectors in USA [2], the GEO detector in Germany [3] and the KAGRA detector in Japan [4]. Farther in the future the IndIGO detector in India is also expected to join the network [5].

With a sensitivity 10 times better than first generation detectors, Advanced Virgo will be able to monitor a volume of Universe one thousand times greater, observing several tens of binary neutron star coalescences within a range of several hundred millions of light years, and possibly detecting binary black hole events billion of light years away [6]. Continuous signals emitted by distorted rotating neutron stars, as well as the stochastic background of cosmological or astrophysical origin will also be accessible with unprecedented sensitivity.

In parallel with the transition from first to second generation experiments we improved the computing framework, which needed to be upgraded in order to fulfill the new and most ambitious scientific requirements.

The computing model (CM) for Advanced Virgo [7] was designed taking advantage of the experience gained so far with the data taking and analysis from the first engineering runs to the latest Science runs VSR1-VSR4 (the last run ended in September 2011). It also takes into

account the technological progresses of these years, from the original Virgo plan, which is dated back to the year 2002 [8]. The fundamental focus of the CM is to collect the requirements of the science data analysis groups and to find optimal solutions to fulfill them. The CM also reflects needs and constraints raised from the LIGO/Virgo agreement [9], which we have been facing during the last years, and were finally addressed in an organized way in this CM.

2. Activity in 2014

During 2014 the on-site activity was dominated by the works related to the new hardware installation and integration. The most relevant project milestones completed in 2014 were the locking of the Input Mode Cleaner in June, and the start of its commissioning, and the suspension, in December, of the Beam Splitter payload, which represents the first complete system (Suspension+Payload+Mirror) integrated and suspended (see Fig. 1). Moreover the large cryolinks were installed in the end buildings. The West End cryolink was successfully tested in December.

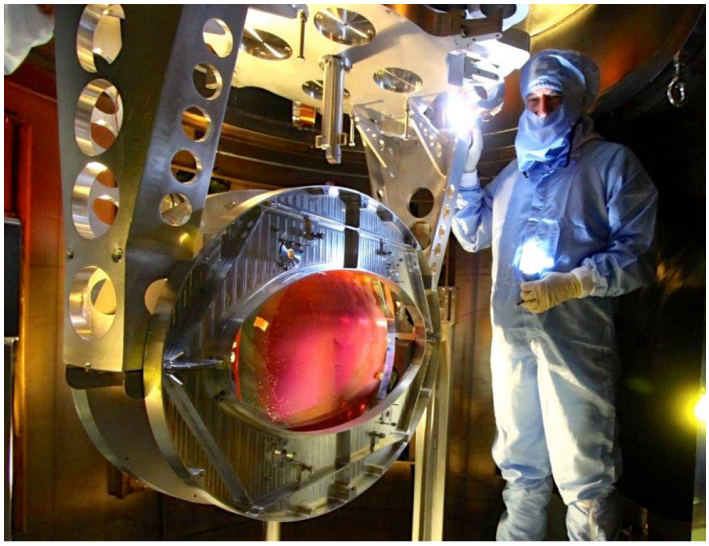


Figure 1. Large Beam Splitter integrated in December 2014.

In the meantime the analysis of the data collected in the science runs that took place in previous years (in particular VSR2, 2009-2010 and VSR4, 2011) has been proceeding. This data analysis activity produced over the last few years a steady flux of ≈ 10 papers/year. The collaboration papers published in 2014 are listed in the References:[10] – [23].

3. The Advanced Virgo Computing Model

The overall computing strategy for Advanced Virgo is described in detail in the Advanced Virgo computing model [7]. Virgo (and Advanced Virgo) has a hierarchical model for data production and distribution: different kinds of data are produced by the detector and firstly stored at the EGO site in Cascina. There is no permanent data storage in Cascina, just a buffer of 6 months of data mainly used for detector characterization.

The external comuputing centers (CCs) receive a copy of the data and provide storage resources for permanent data archiving. They must guarantee fast data access and computing resources for off-line analyses. Finally, they must provide the network links to the other Advanced Virgo computing resources.

For this goal a robust data distribution and access framework (based on file and metadata catalogs) is a crucial element. For data distribution we will use the well-tested framework developed by the EGO IT department, while for file catalogs the LDR DataFindServer has been

installed (or under installation) in the CCs. The collaboration manages also smaller CCs used to run part of some analyses, simulations or for software developments and tests.

During science runs the Cascina facility is dedicated to data production and to different detector characterization and commissioning analysis, which have the need to run "on-line" (with a very short latency, from seconds to minutes, to give rapid information on the quality of the data) or "in-time" (with a higher latency, even hours, but which again produce information on the quality of the data within a well defined time scale). The detector characterization activity gives support to both commissioning and science analysis. Science analyses are carried out offline at the external CCs, with the only exception of the low-latency searches. Some analysis, due to the fact that we analyze data jointly with aLIGO for many searches, are carried on in LSC CCs.

We report here a few considerations on the Advanced Virgo CM, as these have an important impact on the work we will need to do in 2015 and 2016 to be prepared for the first run of the detector, and thus are reflected in the computing requests for the next years. First of all, to face the huge computational demands of gravitational wave searches in the forthcoming years, there will be the need to gather the resources of many CCs into a homogeneous distributed environment (like Grids and/or Clouds) and to adapt the science pipelines to run under such distributed environment. This work has been started and is now in an advanced state. After a short evaluation period we found that the Pegasus workflow scheduler perfectly suits for this purpose i.e. is able to provide a uniform distributed job submission framework and its scheduling and accounting system is compatible with that of LIGO.

Another very important need for the advanced detector era is to provide a Grid-enabled, aLIGO-compatible Condor cluster for Advanced Virgo. While Pegasus can be used for distributed job submission system, due to the strong demand on sharing computing resources and for more efficient workflow development, this request is still very important.

Another important task, which we started to face, is the possibility to run search pipelines in GPU clusters. Due to the very high increase of computing resources which in turn has serious financial consequences, pipeline developers must be enforced to investigate the possibility of re-implementing the most compute intensive part of their algorithms to GPUs. This will either result in a much more consolidated computing budget, or, with similar budget, would open the possibility to examine qualitatively new physics using the same amount of computing resources.

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

Thus the most important issues of the Advanced Virgo computing model may be summarized as follows:

- guarantee adequate storage and computing resources at Cascina, for commissioning, detector characterization and low-latency searches;
- guarantee fast communications between Virgo applications at Cascina and aLIGO CCs/other detectors for low-latency searches;
- guarantee reliable storage and computing resources for off-line analyses in the Advanced Virgo CCs (CNAF and CCIN2P3);
- push towards the use of geographically distributed resources (Grid/Cloud), in external CCs;
- push towards a homogeneous model for data distribution, bookkeeping and access.

Figure 2 gives a big picture of the data workflow for what concerns scientific data analysis and detector characterization activities for Advanced Virgo. Possible additional CCs have also been indicated, as a resource to perform intensive data analysis computation on the most important scientific data channels (which amounts to a really negligible storage need/year).

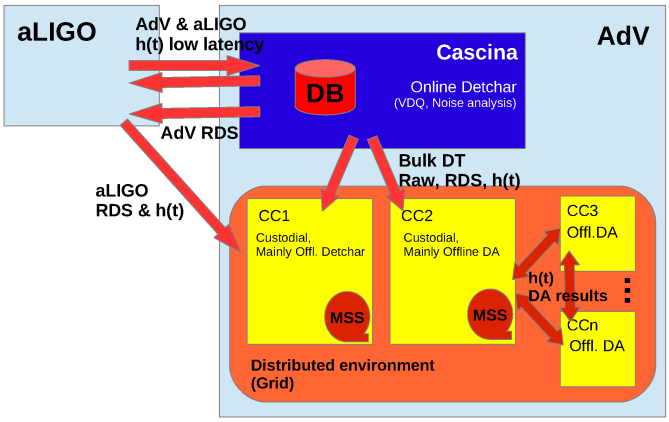


Figure 2. Data workflow for data analysis (DA) and detector characterization (DetChar) activities in Advanced Virgo. CC2 indicates the CNAF, CC1 indicates CCIN2P3.

4. The role of CNAF

The computing usage and needs at CNAF is described in an internal document, annually updated [24].

Over the last year CNAF has mainly been used for all-sky searches for unknown isolated neutron stars by the CW group. Other activities, which required significant amount of computing resources were:

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

Most of these use the Grid.

4.1. Storage

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

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

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

4.2. Computing

At CNAF in 2014, CPU resources have mainly been used for the all-sky CW search and for the CBC Parameter Estimation and General Relativity work. We have also demonstrated successful execution of the MBTA pipeline at CNAF and on other Italian Grid sites. During the year 2014, an important work has been done to do the porting of CBC pipelines from an LSC related submission method to an architecture complaint with our CCs and in particular with Grid. This has removed the limit to run CBC analyses only on LSC clusters. Tests on real data have begun in September 2013 and since then CNAF has granted to Virgo a number of cores $\mathcal{O}(1000)$, which is the minimum needed to prepare the CBC analysis and to run some new CW analysis (on VSR2/VSR4 data) enlarging the parameter space covered so far.

CNAF accounting system[1] provides information on the total and average consumption of computing power in 2014, summarized in Table 2:

Table 2. Accounting at CNAF (date: January, 1st - December, 31 2014)

WCT avg [HS06.day]	CPT avg [HS06.day]	Efficiency avg [CPT/WCT]
4,685.89	5,755.12	1.23

Table 3 shows the evolution since 2007 of the CPU consumptions.

Table 3. Evolution since 2007 of the CPU used at the CNAF

Year	WCT [HS06.day]
2007	60
2008	240
2009	453
2010	162
2011	674
2012	669
2013	850
2014	4686

5. Final remarks

The CNAF support to the VIRGO experiment is not limited to the technical access of the CC facilities. The CNAF expertise was crucial for driving the discussion in the collaboration to define our computing model and focus on right solutions. This kind of support is even more important than the access to their hardware infrastructure. In addition, CNAF support will be fundamental for testing the porting on the GPUs located at CNAF some of most computing demanding searches of gravitational wave signals.

Finally, we note that in the near future of the Advanced detector era, our computing needs will increase: we expect by the year 2018 a need for a continuous power $\mathcal{O}(100)$ kHS06. This implies that the request of computing resources of our experiment on the CNAF infrastructure

[1] `http://tier1.cnaf.infn.it/monitor`

in the forthcoming years, though still below that of the HEP experiments at LHC, will represent a non negligible fraction of the overall resources.

References

[1] Acernese F et al 2015 *Advanced Virgo: a second-generation interferometric gravitational wave detector*, Class Quant Grav **32** 024001

[2] Aasi J et al 2015 *Advanced LIGO*, Class Quant Grav **32** 074001

[3] Affeldt C et al 2015 *Advanced techniques in GEO 600*, Class Quant Grav **31** 224002

[4] Somiya K 2012 *Detector configuration of KAGRAthe Japanese cryogenic gravitational-wave detector*, Class Quant Grav **29** 124007

[5] Unnikishnan C S 2013 *IndIGO and LIGO-India: Scope and Plans for Gravitational Wave Research and Precision Metrology in India*, Int J Mod Phys D **22** 1 1341010

[6] Aasi J et al 2013 *Prospects for Localization of Gravitational Wave Transients by the Advanced LIGO and Advanced Virgo Observatories* `http://arxiv.org/abs/1304.0670`

[7] Virgo Collaboration 2013 *The AdV Computing Model* Virgo Note VIR-0129H-13 `https://tds.ego-gw.it/ql/?c=9474`

[8] Virgo Collaboration 2002 *Virgo Computing Plan* Virgo Note `https://tds.ego-gw.it/ql/?c=1338`

[9] Virgo Collaboration, LSC Collaboration 2013 *Memorandum of Understanding between Virgo and LIGO* Virgo Note VIR-0386A-13 `https://tds.ego-gw.it/ql/?c=9740`

[10] AAsi J et al 2014 *Improved Upper Limits on the Stochastic Gravitational-Wave Background from 2009-2010 LIGO and Virgo Data*, Phys Rev Lett **113** 23 231101

[11] Aartsen M G et al 2014 *Multimessenger search for sources of gravitational waves and high-energy neutrinos: Initial results for LIGO-Virgo and IceCube*, Phys Rev D **90** 10 UNSP 102002

[12] Aasi J et al 2014 *First all-sky search for continuous gravitational waves from unknown sources in binary systems*, Phys Rev D bf 90 6

[13] Aasi J et al 2014 *Implementation of an F-statistic all-sky search for continuous gravitational waves in Virgo VSR1 data*, Class Quant Grav **31** 16 165014

[14] Accadia T et al 2014 *Reconstruction of the gravitational wave signal $h(t)$ during the Virgo science runs and independent validation with a photon calibrator*, Class Quant Grav **31** 16 165013

[15] Aasi J et al 2014 *Search for Gravitational Waves Associated with gamma-ray Bursts Detected by the Interplanetary Network*, Phys Rev Lett **113** 1 011102

[16] Aasi J et al 2014 *Methods and results of a search for gravitational waves associated with gamma-ray bursts using the GEO–600, LIGO, and Virgo detectors*, Phys Rev D **89** 12 122004

[17] Aasi J et al 2014 *Search for gravitational radiation from intermediate mass black hole binaries in data from the second LIGO-Virgo joint science run*, Phys Rev D **89** 12

[18] Aasi J et al 2014 *The NINJA-2 project: detecting and characterizing gravitational waveforms modeled using numerical binary black hole simulations*, Class Quant Grav **31** 11 115004

[19] Aasi J et al 2014 *Search for gravitational wave ringdowns from perturbed intermediate mass black holes in LIGO-Virgo data from 2005-2010*, Phys Rev D **89** 10

[20] Aasi J et al 2014 *Application of a Hough search for continuous gravitational waves on data from the fifth LIGO science run*, Class Quant Grav **31** 8 085014

[21] Aasi J et al 2014 *Gravitational waves from known pulsars: results from the initial detector era*, Astrophys J **785** 2 119

[22] Aasi J et al 2014 *Constraints on Cosmic Strings from the LIGO-Virgo Gravitational-Wave Detectors*, Phys Rev Lett bf 112 13 131101

[23] Aasi J et al 2014 *First searches for optical counterparts to gravitational-wave candidate events*, Astrophys J Supplement Series **211** 1 7

[24] Virgo Collaboration 2013 *Virgo computing status and needs for 2014* Virgo Note VIR-0505C-13 `https://tds.ego-gw.it/ql/?c=9898`

XENON computing activities

G. Sartorelli, F. V. Massoli
INFN e Università di Bologna
E-mail: Gabriella.Sartorelli@bo.infn.it; massoli@bo.infn.it

1. The XENON project

A lot of astrophysical and cosmological observations support the hypothesis that a considerable amount of the energy content of the Universe is made of cold dark matter. Recently, more detailed studies of the Cosmic Microwave Background anisotropies have deduced, with remarkable precision, the abundance of dark matter to be about 25% of the total energy in the Universe. Dark matter candidate particles share some basic properties, mainly: they must be stable or very long lived; they have to be weakly interacting and colorless and they have to be not relativistic. Due to such characteristics, they are identified under the generic name of Weakly Interacting Massive Particles (WIMPs). Among the various experimental strategies to directly detect dark matter, detectors using liquid xenon (LXe), as XENON100 and LUX, have demonstrated the highest sensitivities over the past years. The XENON collaboration is focused on the direct detection of WIMP scattering on a LXe target. Currently, the XENON100 detector is running at LNGS. It set the most stringent limit, for the 2012, on the spin-independent WIMP-nucleon elastic scattering cross section for WIMP masses above 8 GeV/c^2, with a minimum at $2 \cdot 10^{-45}$ cm^2 at 55 GeV/c^2 (90% CL) [1].

In parallel, since 2011, the XENON1T project started. It will be the largest dual phase (LXe/GXe) Xe-based detector ever realized and, after its approval from all funding agencies in 2011, it is now under construction and installation in Hall B at the Gran Sasso Underground Laboratory (LNGS). Both XENON100 and XENON1T detectors are based on the same detection principles. The target volume is hosted in a dual phase (LXe/GXe) Time Projection Chamber (TPC) that contains xenon in liquid phase (LXe) with gaseous phase (GXe) on top. The TPC is enclosed by two meshes: the cathode (at negative voltage) on the bottom and the gate mesh (grounded) on top. This structure contains the LXe active region, called the sensitive volume that represents the volume used to detect the interactions. A particle interacting in LXe produces a prompt scintillation signal (S1) through excitation and recombination of ionization electrons. The electrons that do not recombine are drifted towards the liquid-gas interface where they are extracted into the GXe to produce the secondary scintillation signal (S2). Two PMT arrays, one on top of the TPC inside the GXe and one at its bottom below the cathode, in LXe, are used to detect the scintillation light. The x-y position of the events is determined from the PMTs hit, while from the time 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 of it. With respect to its predecessor, XENON1T will use a larger mount of LXe: about 3.3 tonnes 2 of which will represent the sensitive volume available for the WIMP interactions. Its goal is to lower the current limits on the WIMP interaction cross section of about two orders of magnitude.

To reach such a result, a severe screening campaign 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. Due to the large amount of simulations required to perform that research, the GRID is the most appropriate facility to be used.

2. XENON100

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

3. XENON1T

The XENON1T experiment will use a DAQ machine, hosted in the XENON1T service building underground, to acquire data. It will have a 6TB storage buffer (that could be extended if needed). Such a space is enough to store data for few hours, at the maximum foreseen acquisition rate of 300MB/s during calibration runs, or for few weeks in the case of DM data. The data transfer, to the above-ground XENON computing facility, will take place by means of a 10GB optical fiber connection. For each user will be available a home space of 100 GB on a disk of 10TB. A 64-cores machine is used for analyzing data quickly after they have been acquired. 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 online monitoring web interface, the XENON wiki and GRID UI. Data handled from each service will be hosted in three separate 5 TB disks. For data storage, three separates volumes will be used: a 60TB disk to store the temporary calibration data, another 200TB disk to store the DM raw data during the whole life of the experiment and a 100TB volume to store the processed data. The 200TB for DM data have been evaluated assuming 45TB/year of data taking for more than 4 years. Massive raw data processing (from calibration) will be done remotely with GRID. Moreover, the U-LITE LNGS batch system (360 shared cores) will be used mainly for DM data processing.

CNAF resources have been extensively tested for data simulation (MC with GEANT4 software) using the GRID technology. During the first months of 2014 we already used about 500 HS06 per day and produced about 10 TB of data for the optimization of the detector design and the background evaluation. Due to the success of such a work, it is foreseen to continue to use the GRID to produce simulated data and to store them in the related disk storages. There is also the possibility to store all the processed calibration data on GRID and to move there also the heaviest part of the analysis related to that kind of data. This will certainly increase the amount of CPU and disk space that is foreseen to be used during 2015 and later. Due to the large amount of data for calibrations, the remote data processing would be feasible only if a reasonable bandwidth to connect LNGS to GRID will be guaranteed. The high-speed

telecommunication network provided by the GARR consortium is under test. The final goal is to have a 10Gbps line before the end of 2015.

References

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

The INFN-Tier1 Center and National ICT Services

The INFN Tier-1: a general overview

L dell'Agnello, A Cavalli, L Chiarelli, A Chierici, S Dal Pra, D De Girolamo, M Donatelli, D Gregori, A Mazza, G Misurelli, M Onofri, M Pezzi, A Prosperini, P Ricci, F Rosso, V Sapunenko, A Simonetto, S Virgilio and S Zani

Abstract. This contribution presents a general overview of the situation at the INFN Tier-1, followed by some salient facts that happened during 2014. More detailed reports on specific activities are covered in other contributions to this Annual Report.

1. Current availability of resources

The number of scientific collaborations using the INFN Tier-1 hosted at CNAF has increased during 2014: NA62, Panda, CTA, Darkside and Cuore have joined the existing collaborations using the computing and storage resources available at our data center.

Table 1 summurizes the pledged resources at the INFN Tier-1 by the end of 2014.

Experiment	**CPU** (HS06)	**Disk** (TB-N)	**Tape** (TB)
LHC experiments	98950	11320	16190
HEP experiments	19700	610	4625
Astro-particle experiments	18093	2770	2175
Virgo	10000	428	818
Nuclear physics experiments	0	0	380
Total	**135546**	**15128**	**24188**

Table 1. 2014 pledged resources

1.1. The computing farm

At the beginning of 2014 the general-purpose farm had a total computing power of about 190 kHS06, which was reduced during the course of the year. The quite large "over-pledge" was due to old hardware not yet dismissed.

Figure 1 presents the usage of the computing farm in terms of submitted jobs during the first quarter of 2014 and it shows how the farm was (and is) almost always fully used, despite the large over-pledge. The bumps visible in the graph are typically due to maintenance interventions on the farm, for example to apply patches to the system.

The computing resources of the general-purpose farm were then diminished starting from March as a consequence of a serious incident: in the night between March 8^{th} and 9^{th} a fire

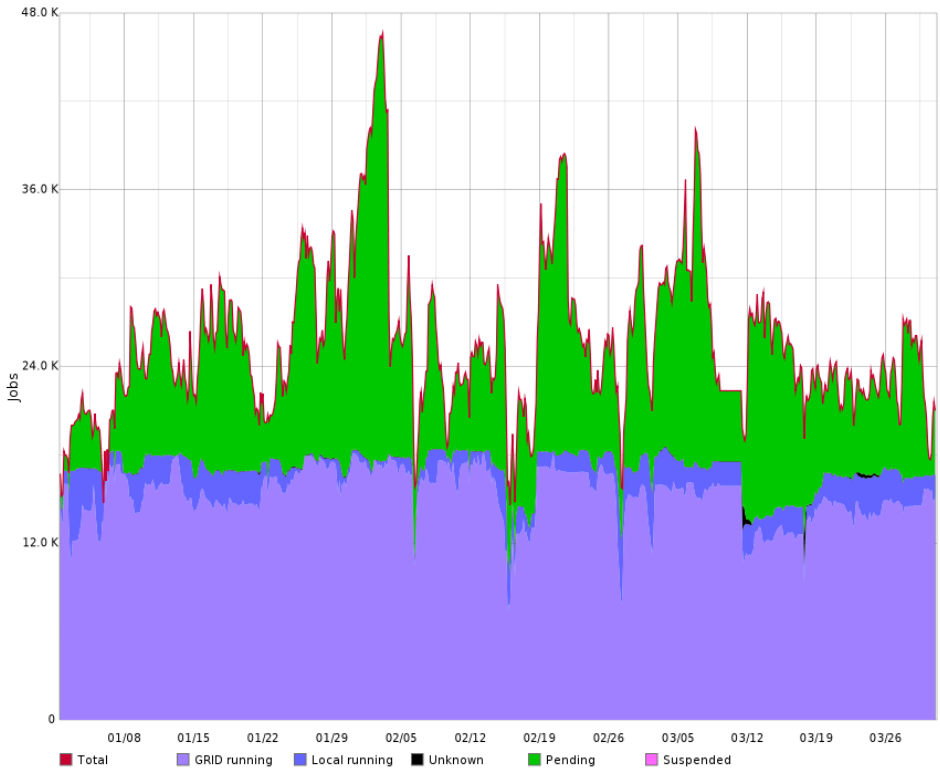


Figure 1. Farm usage in the period January–March 2014

broke out at the control panel of one of the chillers, which caused all of them to go out of power due to the fault of the common power line. Indeed, after the intervention of the fire squad, a quick raise of the temperature in the computing center was registered and an emergency shutdown of all the resources was performed to avoid major damages.

After the restoration of the power line to the chillers (on Sunday March 9^{th}) the services were gradually reopened and on March 10^{th} Tier-1 operations were restored, albeit with a reduced computing capacity.

We then decided to definitively dismiss the oldest worker nodes, hence reducing the computing power of the farm to about 175 kHS06, which was still well above the pledge limit.

1.2. The storage

Storage resources, on the other hand, in 2014 have increased in size of nearly 2 PB for the disk (bringing the total to about 15 PB and more thatn 50 million files), while the real occupancy of the tape storage (about 16 PB) has remained well below the pledge values (about 24 PB). At the end of 2014 a complete repack of the data on tape has started in order to move all the data to the new TK10D drives.

Both disk space and tape space are structured as a Hierarchical Mass Storage system and are managed by GEMSS, the Grid Enable Mass Storage System, a home-made integration of the IBM General Parallel File System (GPFS) with the IBM Tivoli Storage Manager (TSM). An Oracle clustered database infrastructure is deployed for relational data storing and retrieving.

In 2014 also the new activity of the CDF Long Term Data Preservation has been set up and the bulk copy of data from FNAL started [5].

Disk and tape storage services, together with the data transfer services, are operated by the Data storage group within the Tier-1 data center unit.

1.3. The infrastructure
Two improvements in the basic infratructre are worth mentioning:

- Anti-flooding doors were installed for all Tier-1 halls.
- The efficiency of the chillers increased with the installation of a by-pass circuit between the manifold outlet and the collector recovery.

1.4. HPC resources
Besides the main computing and storage resources, a small HPC farm based on GPU cards was installed in January: in total 4 servers with one NVIDIA K20 card and 4 servers with 2 NVIDIA K40 cards were made available to the users.

2. Highlights

The most significant activities performed during 2014 in the context of the data center are presented in other contributions to this Annual Report [1–6].

This section lists some additional activities performed during the period that are worth mentioning.

2.1. Migration to oVirt
To ease the deployment of virtual machines the virtualization infrastructure has been migrated to the oVirt software platform. Thanks to this solution the farming group is now able to delegate machine management to single users and to interface the infrastructure with other software tools (like Foreman, a solution under investigation at CNAF in order to take over the old configuration and installation system).

2.2. Study of alternative batch systems
We have also evaluated alternatives to LSF, like UGE (Univa Grid Engine) and Slurm, both to address economic issues and to prevent potential LSF scalability problems. While Slurm failed to satisfy some basic requirements for our environment, resulting fragile and not scalable enough when applied to our use cases (issues in term of scalability have been reported also by other Tier-1s), the tests with UGE offered us a real alternative. However we were able to negotiate with IBM a reduction of the cost for the LSF license and hence the agreement has been renewed until December 2018.

In the HEP community there is also a renewed and strong interest for HTCondor, thanks to its top-of-the-class features; indeed, it is a serious candidate for substituting LSF at this Tier-1 after 2018.

2.3. Support for multi-core jobs
At the beginning of 2014 we joined the WLCG Multicore Task Force and in August we could enable the multi-core jobs on our farm as required by ATLAS and CMS experiments. This was done through an innovative configuration of the farm allowing its dynamic partition between standard and multi-core jobs, minimizing the waiting time and optimizing the resource usage. We were able to reach an efficiency of over 90%.

2.4. Emergency shut-down procedure
As a follow-up of the recovery from the fire incident mentioned above, we reviewed the emergency shut-down procedures for the computing room. If a cooling problem occurs it is vital to identify the nodes that can be easily switched off in order to keep the room temperature at an acceptable level: the largest fraction of the machines in the computing room is constituted by worker nodes,

which can be easily switched off in case of emergencies, leaving time to other, more complicated systems (for example the storage), to be taken care of (and eventually be switched off too). The farming group implemented a procedure to smoothly switch off all the computing nodes, using both a software approach (a "poweroff" command issued via ssh) and a hardware one (IPMI command to cut the power of a node). In the future this procedure, still in the testing phase, could be made available to the on-call operator. allowing us to increase security and safety of the computing room.

References

[1] Low-power CPU investigation, this report.
[2] Adapting a custom accounting system to APEL, this report.
[3] Dynamic partitioning for multi-core and high-memory provisioning with LSF, this report.
[4] Towards a common monitoring dashboard for the Tier-1, this report.
[5] Projecting the CDF computing model to the long-term future, this report.
[6] The INFN Tier-1: networking, this report.

Low-power CPU investigation

Andrea Chierici

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

Stefano Dal Pra

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

Giuseppe Misurelli

E-mail: giuseppe.misurelli@cnaf.infn.it

Saverio Virgilio

E-mail: saverio.virgilio@cnaf.infn.it

1. Introduction

INFN-T1 is quite expensive to maintain, due the cost of electricity. We started to investigate some promising low power solutions developed by chip makers, trying to understand if they could fit our requirements of both computing power and cost per watt.

2. Intel Avoton

Our computing room PUE[1] is 1,6 and the average monthly power required hits 1100kW. Our farm hosts several old nodes with a poor hs06/w ratio, hence we felt the need to better investigate solutions to improve power efficiency. Recently Intel introduced the Avoton[2] processor (C2000 family) a system-on-chip built on 22 nanometer process technology. This chip comes with up to 8 64-bit cores based on Silvermont micro architecture, originally conceived to address the needs of microserver, entry level communication infrastructure and cloud storage markets. It's remarkable power consumption of 20W pushed us to investigate the possibility to use this chip in our farm as opposed to our standard server chip solutions (to give an idea, a common processor in our farm requires 115W).

3. Tests

We were able to test 2 different solutions, one based on a stand-alone appliance made by Supermicro, and the other based on the *Moonshot* system by HP[3]. In both cases the chip was the same and we obtained comparable results: the main difference is the form factor of the 2 solutions, one more targeted at stand-alone appliances, while the other to high density low storage solutions. Figure 1 shows a good increase in HS06 according to the number of concurrent runs, meaning that the CPU is capable of scaling well till we reach the number of physical cores (8 in this CPU). More important is to compare this result with other CPUs in our computing farm, as depicted in figure 2. The Avoton CPU performs significantly slower than the others,

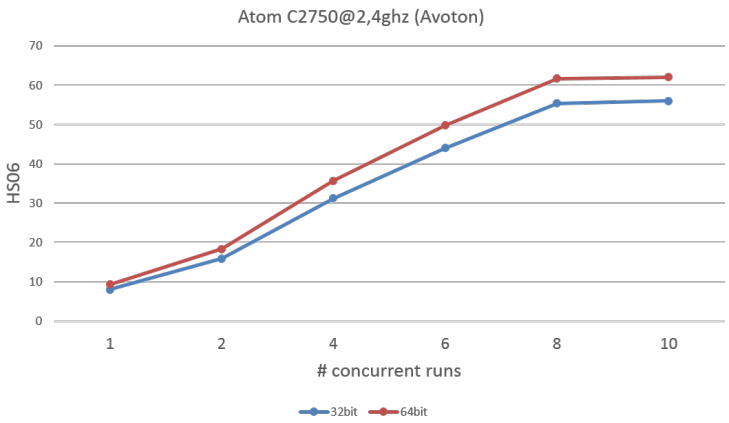


Figure 1. C2750 HS06 Calculation

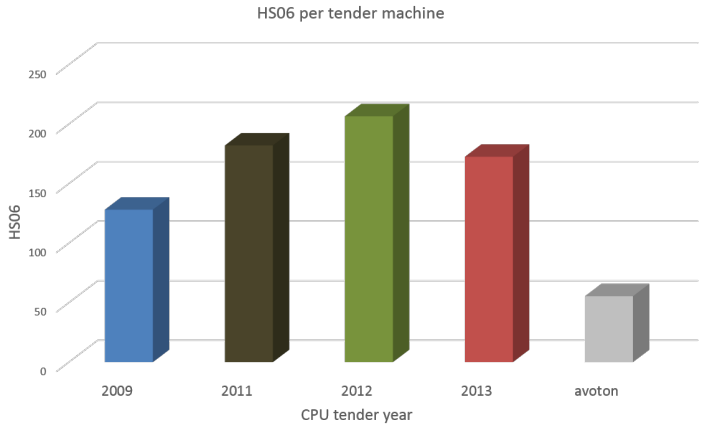


Figure 2. HS06 per tender

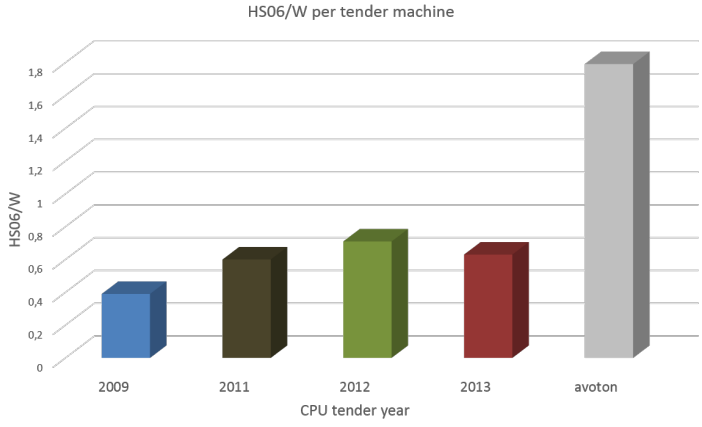


Figure 3. HS06 per watt

60k HS06	*last tender*	*HP Moonshot*
n. Enclosures	87	25
n. Racks	6	4
n.Motherboards	348	1125
kW required	88	35
w/HS06	1,47	0,56
Purchase Cost	x	1,68*x
Electricity Cost/Y	x	0,4*x

Table 1. TCO Simulation

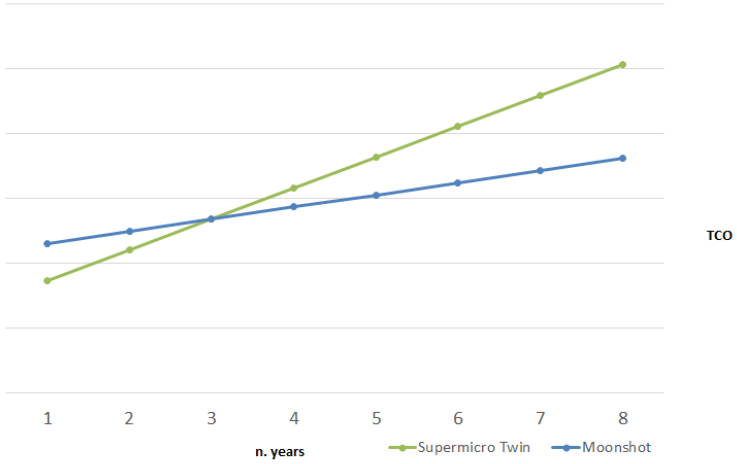


Figure 4. TCO simulation: Avoton CPU vs last tender

less than half. This result is overturned if we evaluate the number of HS06 per Watt consumed by the CPUs: in this case, as depicted in figure 3, Avoton CPU greatly outperforms all the other solutions currently installed in our computing farm.

4. TCO simulation

To better understand the possibility to switch to this type of solution, we calculated a TCO simulation comparing the declared street price of a HP Moonshot system with the last CPU tender we acquired: the results are shown in table 1.

From the table we can understand that the initial cost of the Avoton solution is significantly higher compared to the last tender, but we have to stress the fact that it was computed on street price, not on a tender price (where prices generally are significantly lower). Anyway starting from the fourth year (as depicted in figure 4) the solution becomes cost-effective.

5. Conclusions

Since the Avoton CPU uses the same microcode architecture of server chips, switching to this kind of solution is rather easy and simple, indeed no code recompilation is required (as compared to other low power solutions, like arm chips). The CPU TDP of only 20W means less cooling power is required to cool the computing room and this implies the possibility to save quite a lot of money. This solution apparently has got some drawbacks that need further investigation: even if the results are promising, the Avoton CPU is not so powerful compared to standard server

class CPU, meaning that more computers have to be purchased in order to reach the pledged resources experiments require. More computers mean bigger human effort to keep all the nodes up and running and possibly more hardware failures. It's not easy to understand and decide if this solution is the best for our computing requirements, we will continue the investigation, possibly with the collaboration of hardware vendors, asking the possibility to deploy a certain number of such computers in our farm in order to test them in a real production environment.

6. References

[1] PUE on Wikipedia: http://en.wikipedia.org/wiki/Power_usage_effectiveness
[2] Avoton on Intel Website: http://ark.intel.com/products/codename/54859/Avoton
[3] HP Moonshot system Website: http://www8.hp.com/us/en/products/servers/moonshot/

Adapting a custom accounting system to APEL

Andrea Chierici

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

Stefano Dal Pra

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

Giuseppe Misurelli

E-mail: giuseppe.misurelli@cnaf.infn.it

Saverio Virgilio

E-mail: saverio.virgilio@cnaf.infn.it

1. Introduction

Starting from October 2013, we upgraded the lower layer of the accounting system of the Tier–1 with a solution designed and implemented from scratch, due to an incident occurred with the previous system, the Distributed Grid Accounting System (DGAS). Now, accounting data for Grid and local jobs are stored in a `acct` PostgreSQL database whose content and schema is managed by the farming staff[1].

Data propagation to the `accounting.egi.eu` portal is performed as before, through the DGAS–HLR hierarchy: the MySQL database in the first–level HLR is updated with the newest accounting records. This step, formerly performed by specific DGAS components, has been replaced by a custom component whose main task is to fetch new records from the `acct` database and to insert them into the `hlr` database hosted in the site HLR. These data were then transmitted to the second–level HLR, hosted and managed by staff at INFN–Torino. This system acted as a central collector for all the accounting data from all the sites in the italian Regional Operation Center (ROC) and was also responsible of their delivering to the `accounting.egi.eu` portal of WLCG accounting, managed by APEL[2].

On September 2014 the second–level HLR of the Italian ROC had to be dismissed and after a short discussion, the IGI consortium decided that Italian sites had to replace their DGAS infrastructure with the one provided by APEL: the main difference is that each APEL instance directly delivers usage records straight to the `accounting.egi.eu` portal.

2. Incompatibility issues

A few issues prevented us from directly adopting the APEL accounting model and forced us to study a different solution:

multi–site batch system Three Grid sites are supported by the "Tier–1" computing centre, and they all rely on a single instance of the Platform/LSF batch system deployed on the

farm: APEL however cannot fairly handle this case, as it assumes that each Grid site relies on its own batch system.

LSF log parsing Even in the most recent APEL release available, the component responsible of parsing the raw accounting records logged by LSF, proved to be bugged, failing to recognize a number of valid accounting records.

Wrong WallClockTime LSF logs the time each job spends in `RUN` status in the `runtime` field. APEL however does not collect that value and sets `WallDuration` as "`endtime - starttime`". This is correct if the job is not suspended during its lifetime and if it is dispatched only once, which is not always guaranteed. Furthermore, if a job is killed before being started, the `starttime` field of its log record remains at its default value, which is zero. This fools the APEL parser to account for a job with a `WallDuration` equal to *epoch*, i.e. as if it started in 01-01-1970.

multi–core The number of used cores is not considered in the computation of `WallDuration` for multi–core jobs, which then turns out to have an efficiency $E = \frac{CPUtime}{WCtime} > 1$. This is not consistent with the case of single–core jobs and should instead be computed as $E = \frac{CPUtime}{cores \times WCtime}$.

HPC INFN–T1 hosts a small HPC cluster, managed by LSF 9.1, running many–core, multi–core or GPU based applications. The accounting of the activity of this cluster is done using the same custom accounting system described in [1]. This enables us to quickly adapt our accounting requirements even on new and different use-cases, such as that of GPU computing.

Custom needs A generic accounting system should aim at collecting and storing only those accounting information that make sense on the widest set of batch systems: still, many useful specific data are logged by the LSF batch system, and these are valuable to the farm staff (such as the queue name, the submission or execution host, the job exit status, the submission request parameters, etc.). Thus, the overall information gathered by our local accounting is a superset of those currently required and requested to provide accounting information. A single accounting system (with all the data of interest stored in it) is simpler to maintain and eases the task of extracting a specific subset of info to fulfil a particular request.

3. Switching the data propagation model

After an incident happened at the 2^{nd} level HLR and its subsequent dismissal, the way to propagate the accounting data from the Tier–1 to the EGI portal had to be changed. The data delivered from our local `acct` database had to be adapted to directly transmit accounting records to the EGI–APEL database using the `ssmsend` utility coming from the APEL distribution.

3.1. The APEL patching attempt

At first, an attempt to adapt APEL software to solve the aforementioned issues was made. Three APEL python programs were patched to provide multi–site awareness and multi–core compliance. Furthermore, the low level parsing of the raw LSF log files was adapted to use a reliable open source library, from the `python-lsf-tools` collection. The former log parsing approach used the `lsf.py` low level module from the APEL distribution, that is based on a regular expression which simply is too fragile and fails to correctly parse the records on a certain number of corner cases. The `accounting.py` library from `python-lsf-tools` proved indeed to be more reliable.

After verifying the updates were correct, a ticket (#109485) had been submitted to `https://ggus.eu` to propose the APEL team the adoption of the patches and to get support during the process of the adaptation of the accounting data delivery.

3.2. The working approach
Even if this approach proved to produce correct data, while waiting for the feedback about the proposed patches, we followed a different strategy: the data to be delivered using the `ssmsend` APEL utility from the `apel-ssm` package are produced by a custom script which extracts them from the `acct` database and then dumps them into a set of files equivalent to those produced by the original APEL system. Each file contains one–thousand usage records and is deleted just after its delivery. The final outcome of the implemented solution has a significant difference with respect to what would be done using current APEL approach: the `Site` field doesn't come any more from a configuration file, it hasn't static values hard coded in it but is defined instead as a map of the submission host. This enables one single host to deliver all the accounting records produced by a Batch System instance, despite the number of CEs used or the logical Grid Sites relying on it.

3.3. Faust
Short after dismissing the old 2^{nd} level HLR, the former DGAS staff introduced *Faust*, a new accounting repository for the Italian ROC, offering a web–service based API interface intended for extracting graphical reports for the client. This repository has to be fed with the same data delivered to APEL, so the transmission process is repeated twice, simply providing a different configuration file to the `ssmsend` utility.

3.4. Validation
The validation process for the new system was followed through the aforementioned ggus ticket, which provides a pretty detailed report on the progress of this activity.

4. Conclusions
Our implementation proved to be reliable enough to be used in every day production: we have now a more robust accounting system that is used every day by a great number of people flawlessly.

References
[1] S. Dal Pra, "Accounting Data Recovery. A Case Report from INFN-T1" Nota interna, Commissione Calcolo e Reti dell'INFN, `CCR-48/2014/P`
[2] APEL, `https://wiki.egi.eu/wiki/APEL`

Dynamic partitioning for multi-core and high-memory provisioning with LSF

Andrea Chierici

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

Stefano Dal Pra

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

Giuseppe Misurelli

E-mail: giuseppe.misurelli@cnaf.infn.it

Saverio Virgilio

E-mail: saverio.virgilio@cnaf.infn.it

1. Introduction

During January 2014, The INFN–T1 was requested to enable the execution of multi–core applications, using 8 cores simultaneously in the same Worker Node. Even thought current batch systems are commonly able to accept and manage a great variety of jobs, this kind of activity requires special configurations in a typical WLCG farm: this is because the common workload is given by single–core jobs. A 8–core job can only start when eight free slots are available at the same time in the same node, which is an extremely unlikely event.

2. Traditional solutions and their drawbacks

Host partitioning: this is the most simple and straightforward solution. A fixed subset of the Worker Nodes of the site are exclusively dedicated to multi–core only activity. This is unacceptable at the INFN–T1, because too many resources would remain unused when not enough multi–core processing is in progress or because subset may bee undersized during a boosted multi–core demand.

Advanced reservation: this is a standard feature provided by most batch systems, LSF included. Assuming that the batch system knows the maximum expected end–time for each running job, it can select the node where enough slots are early going to be free. The node can be exclusively reserved for a specific multi–core job during a suitable time window, when currently running jobs are expected to be finished.

The duration of most of the running jobs is totally uncertain, varying from a few seconds (extremely frequent with *empty pilot* jobs) to a large upper run limit defined at queue level. This means that each time a node gets reserved, a *drain time* begins, during which no single–core jobs can be scheduled on it, causing free slots to be unusable until enough are

available to host a multi–core job. Furthermore, the batch system would systematically trigger an advanced reservation for each pending multi–core job, adding more and more CPU power loss due to draining times.

The available known methods to run a mixture of single and multi core jobs are not general enough to fulfil our case, so a different model had to be designed.

3. The dynamic partitioning model

A desired feature would consist in having a varying number of nodes dedicated to multi–core according on needs. This would be equivalent to an auto resizing host partition, and can be defined according to the following principles:

Drain one, run many: once selected for multi–core and put on draining, a node should run multi–core jobs as long as possible, in order to minimize the impact on the drain time.

Follow the demand: depending on the number of pending multi–core jobs, more nodes should be dedicated, up to a defined threshold.

Avoid emptiness: if no multi–core jobs are being scheduled to a dedicated node, it should be reverted back to work with ordinary single core activity.

4. Implementation with LSF

The dynamic partitioning has been implemented by a few python scripts, a couple of simple C programs and a configuration file. The main elements are:

elim: to obtain a dynamic partition, worker nodes are tagged with a `mcore` flag. WNs whose flags equal one are dedicated to multi–core. This means that:

- single–core jobs cannot be dispatched to nodes having `mcore==1`
- multi–core jobs must be dispatched to nodes having `mcore==1`

The flag is, in LSF terminology, an *external load index* and its value is computed and reported by the WN itself, by running an *elim* script written by the administrator. This computes and prints to its `stdout` the value of the flag every 60 seconds.

esub: for each job, at submission time, another custom script, the external submitter *esub* is executed in the submission host (the CREAM–CE, in our case). It distinguishes multi–core jobs from single–core ones by inspecting the submission parameters, then modifying one of them, the *resource request* to ask `mcore==1` for multi–core jobs and `mcore!=1` for single–core jobs.

director: a third script, the *director*, implements the logic of the partitioning model. It runs every six minutes and decides which nodes are to be added or removed from the mcore partition. The result is written in a Json file on a shared filesystem, accessible to any node in the LSF cluster. The *elim* running on each WN decides the value of its own mcore flag status by inspecting this file.

status: the director makes use of status information, which are provided by a couple of simple `C` programs based on the `lsf/lsbatch.h` api. These information are:

- the resource request of each pending or running job.
- the current set of unavailable or closed nodes.

configuration: it is possible to tune the behaviour of the dynamic partitioning by configuring the host groups available for multi–core, specified as a list of racks in our case, the maximum number of nodes that can be in drain status, the maximum acceptable number of unused slots, the higher acceptable ratio of unused slots, and a number of other minor details.

hosts: we selected for mcore queue, WNs with 16 and with 24 slots, thus being able to host two or three mcore jobs.

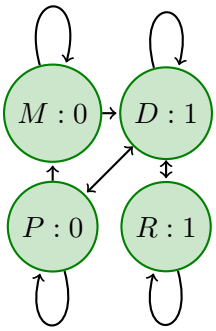


Figure 1. The status transition map

4.1. Transitions
The director considers each node as a member of one of four disjoint sets:

- M: available for mcore. This is initially the set of hosts defined in the configuration file. Only single–core jobs can be dispatched to nodes in this set, and they have their mcore flag set to 0.
- D: Assigned to mcore. A node in this set is in drain time. It may have single–core running jobs but no single–core can be dispatched there. It might have up to eight free slots. Nodes in this sets have their mcore flag set to 1.
- R: Running only multi–core jobs, drain time finished, no free slots. Nodes in this sets have their mcore flag set to 1.
- P: Purged from mcore. A node in this set comes from the D set, where it was found to have free room for multi–core jobs after many (default: three) consecutive checks. There are multi–core jobs still running, however single–core only can be dispatched to Nodes in this set. The mcore flag is set to 0.

4.2. Dynamic partitioning
The dynamic partitioning works like a finite state machine, as represented in Fig. 1

(i) At $T = 0$, all WNs are w_i in the set $M = \{w_1, \ldots, w_N\}$

(ii) When $Q_m > 0$ multi–core jobs are queued, k WN are moved from M to $D = \{w_1, \ldots, w_k\}$ by the director.

(iii) When a node is full of multi–core, it is moved from P to R.

(iv) When a node $w_i \in D$ has free room for a multi–core and no jobs have started there after a timeout, it is moved from D to P.

(v) When more multi–core nodes are required, they are moved from $P \cup M$ to D, beginning with P.

(vi) The elim script on each node w_i updates its mcore status:

$$mcore(w_i) = \begin{cases} 1 & if\ w_i \in D \cup R \\ 0 & if\ w_i \in M \cup P \end{cases}$$

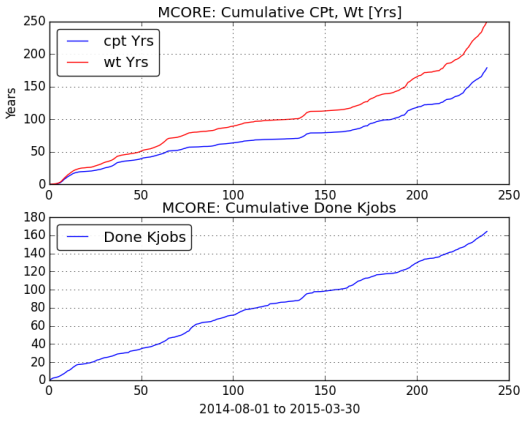


Figure 2. MCORE: Done jobs

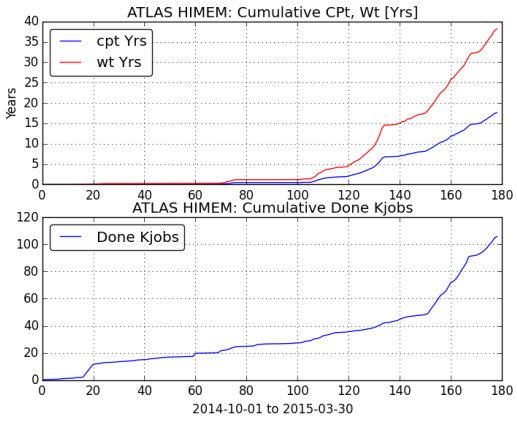


Figure 3. ATLAS HIMEM: Done Jobs

5. Deployment and improvements: high memory jobs

The dynamic partitioning system started working on the production batch system on the 1st of August 2014, and proved to work successfully (Fig. 4) without the need for manual interventions.

During October, a new requirement came from the ATLAS LHC experiment to provide resources for the so called *himem* jobs. These are special single–core jobs requiring twice the RAM of an ordinary one.

To fulfil this, himem jobs are treated like multi–core ones, except that they require two slots in the mcore partition. This way one core remains idle, however the working one has the requested amount of RAM. Furthermore, LSF is configured to run no more than four such jobs per node. Doing so, resources are always granted for at least two 8–core jobs on nodes with 24 slots.

This solution has the benefit of reducing the number of unused slots, at the cost of a little delay in the start of 8–core jobs on the node.

Month	*cpu*	*VO*	*n*	*CPT_days*	*WCT_days*	*%Eff*	$E[CPT]\,[h]$	$E[WCT]\,[h]$
2015-03	1	atlas	338737	57508.103	60272.685	67.217	4.075	34.163
2015-03	1	cms	207049	57664.471	74034.616	39.111	6.684	68.654
2015-03	2	atlas	52900	3529.259	7683.582	43.272	1.601	13.944
2015-03	8	atlas	23848	13799.668	18293.281	70.298	13.888	18.410
2015-03	8	cms	2798	8511.604	11948.709	28.167	73.009	102.491

Table 1. Activity by core and VO, March 2015. Noticeably, the average efficiency of multi–core jobs ($cpu = 8$) is higher than that of single–core. This partially compensates the cpu power loss due to draining.

6. Results

The dynamic partitioning started in August 2014, enabling access to compute resources for 8–core jobs. Two months later it was adapted to enable high–memory jobs too. Fig. 2 and 3 report cumulative CPUTime and WallclockTime for multi–core and high–memory activity. Tab. 1 reports *per cpu* accounting data for March 2015.

In order to evaluate performances, a *Fill Factor* was considered, defined as $FF = \frac{used\ slots}{dedicated\ slots}$, with optimal value $FF = 1$. Fig. 4 shows an insufficient behaviour, with an average fill factor

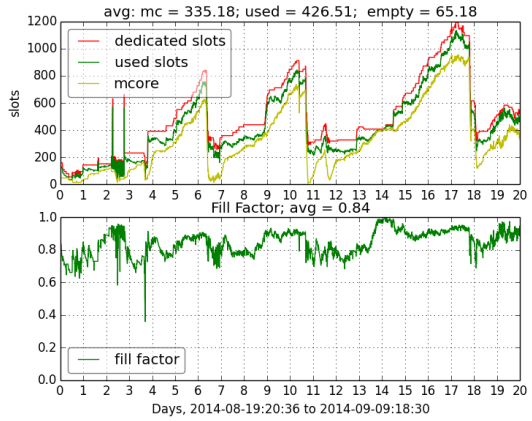


Figure 4. Dynamic partition early days, August 2014, multi–core. The space between red and green line represents unused slots. Sudden submission dropdown have negative impact on average efficiency (see Fig. 6). Different configurations have impact on the reactivity of the system, i.e. how quickly the partition grows or shrinks.

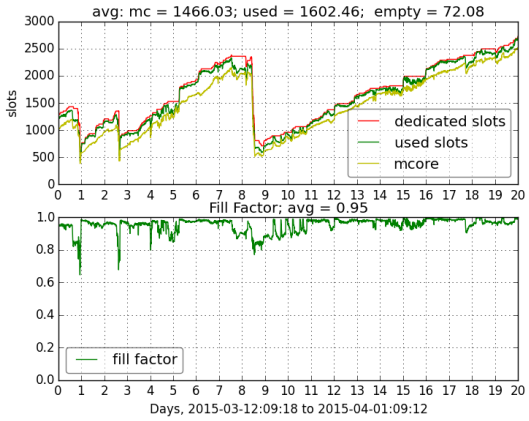


Figure 5. Dynamic partition, March 2015, multi–core and high–memory. The number of unused slots over time is much reduced respect to the case of Fig. 4. By inspecting the *used slots* line it can be noted how the fill factor decreases when there is lack of himem jobs. The partition reaches greater size and dropdown are less frequent.

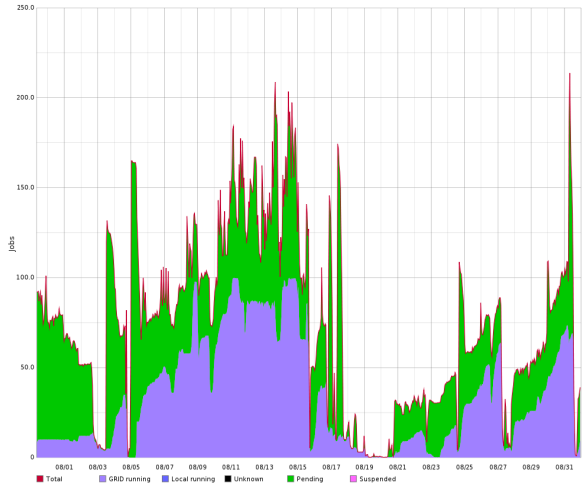


Figure 6. Atlas, multi–core, 2014 Aug. The submission flow suddenly interrupts several times a month. As a consequence nodes in the mcore partition are put back at work with single–core jobs. Short after, submission flow restart, triggering the need for a new draining session.

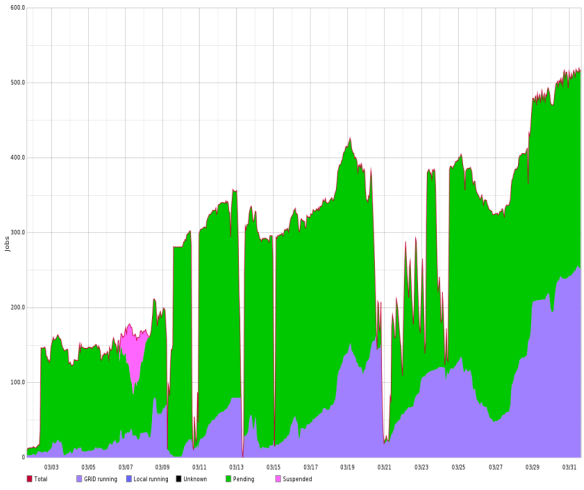


Figure 7. Atlas, multi–core and high–memory, March 2015. The number of unused slots over time is greatly reduced respect to the case of Fig. 6. By inspecting the *used slots* line it can be noted how the fill factor reduces when the himem job submission flow stops.

$FF = 0.84$ and a partition size of 335 slots, 65 of which unused. On the other hand, the partition grows and shrinks itself according to the multi–core jobs submission flow (Fig. 6). As expected, during time periods of steady submission flow, the fill factor increases near to 1.

Fig. 7 shows much better performances, with $FF = 0.95$ and an average partition size of

1466 slots, 71 of which unused. The improvement is due to a more regular submission rate and to high–memory jobs, running in the mcore partition with 2 assigned slots each.

7. Conclusions

A dynamic partitioning system for the LSF batch system has been designed and implemented at INFN–T1. It works smoothly, provides resources for multi–core and high–memory jobs. The system is more efficient with smooth job submission flows and suffer with sudden interruptions and restarts, because of the need for draining filled resources. (Fig. 6). High–memory jobs are helpful to reduce the number of unused slots on the draining nodes. Having more independent multicore submitters also helps in preventing partition collapsing.

Towards a common monitoring dashboard for the Tier-1

Andrea Chierici

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

Stefano Dal Pra

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

Giuseppe Misurelli

E-mail: giuseppe.misurelli@cnaf.infn.it

Saverio Virgilio

E-mail: saverio.virgilio@cnaf.infn.it

1. Introduction

INFN-T1 monitoring and alarming agents produce tons of data describing state, performance and usage of our resources. Before we began this project, there was a lack in usage statistics reporting, a simple page where management board could look at how many jobs run during a certain period, how many HS06 they consumed as well as other kind of statistics showing how the data center is used by experiments. Hence, at the beginning of 2014, we started a cross-groups project aimed to collect, store, process and publish metric data centrally on a web portal (that we called Monviso): now we can provide both our resource administrators and user community with a central monitoring dashboard.

Every group inside INFN-T1 (like farming, storage and network) collect data and metric in a different way, due to the different devices used and solutions implemented during the years. It was not requested to drop every tool and start from scratch, but instead we focused on implementing a common set of rules to present the different data in a similar way. Every collector (tipically an INFN-T1 internal group) must follow these rules:

- sends their own data based on a common metric domain
- metrics are stored in a time-series database
- metric data should be consumed in a programmatic way

We chose to avoid the development of in house solutions in favour of open source ones, already adopted by large communities: we opted for the Graphite[1] project which provides a suite of components able to deal with the aforementioned requirements. In fact, it consists of three modules:

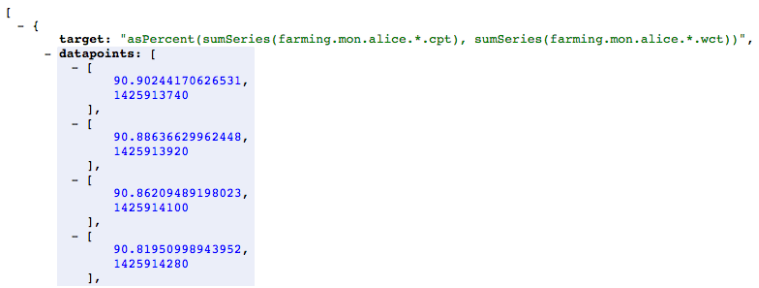


Figure 1. Snippet of the json data returned by the Graphite url API

Carbon component
for dealing with incoming data saving them into the database

Whisper time-series
database library to implement the required round-robin-database like functionality[2]

Graphite-Web
application that renders graphs and dashboards

In order to collect data in the time-series database, before deploying Graphite and data producers we agreed on a metric domain with the rationale to distinguish each group by its prefix followed by the metric name of the type of data it carries on. For instance, *farming.accounting* gives access to the farming group accounting data about executed jobs. In this scope, metrics are organized within the sub-domain *experiment_name.type_of_job.unit* (possible query about HS06 for all the experiment and grid jobs).

Consequently, we are now able to consume Graphite data-points based on the json format it provides by simply setting the proper url request parameters.

For example, to compare the amount of cpu time consumed by the Alice experiments over the wall clock time we can just make the following http request:

```
http://graphite_fqdn/render?target= \\
asPercent(sumSeries(farming.mon.alice.*.cpt), \\
sumSeries(farming.mon.alice.*.wct))&from=-1h&format=json
```

which results in the job efficiency (percentage of CPT/WCT) for all the Alice jobs (figure 1).

The result of this work is the Monviso web portal, which handles `json` data-points and depicts them in charts. Its web structure is based on the Bootstrap[3] framework to deal with all the `html`, `css` and `js` stuff. Charts are generated by exploiting the Jqplot[4] library plus a set of jQuery[5] and server side logic to get `json` data from Graphite and depict results in charts.

With the Monviso portal (figure 2) we have now a central point to gather resource usage reports on computing, storage, network and facility.

References

[1] Graphite on Github repo, `https://github.com/graphite-project`
[2] Time-serie database on Wikipedia, `http://en.wikipedia.org/wiki/Time_series_database`
[3] Bootstrap website, `http://getbootstrap.com`
[4] JQplot website, `http://www.jqplot.com`
[5] JQuery website, `http://jquery.com`

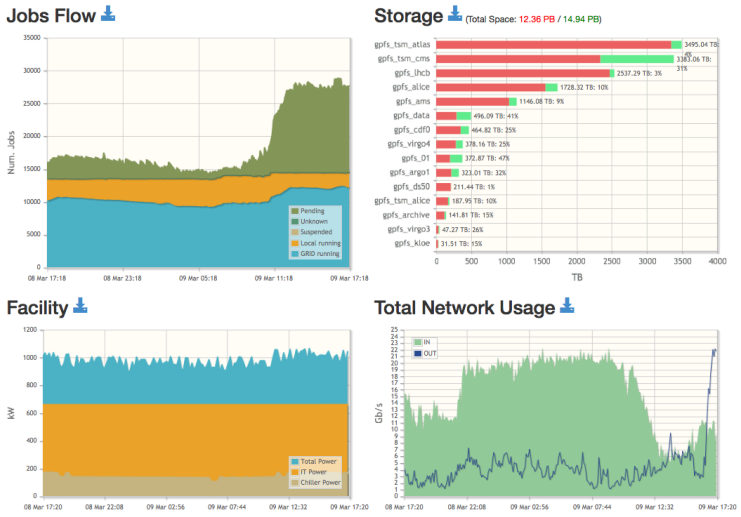


Figure 2. Monviso homepage

Projecting the CDF computing model to the long-term future

Silvia Amerio[a], **Michele Pezzi**[b]

[a]Dept. of Physics, University of Padova, via Marzolo 8, Padova (Italy)
[b]INFN-CNAF, viale Pichat 6/2, Bologna (Italy)

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

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

1. The CDF experiment and recent results

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

CDF ended its data taking in 2011. More than 70 papers have been published since then, and many analysis are still ongoing on its $10 fb^{-1}$ data sample. In the top sector, a measurement of the single top quark production cross section on the full data sample was published in 2014 [4]. CDF and D0 measurements were combined and resulted in the first observation of single-top-quark production in the s-channel [5]. In the B sector, recent results include measurements of direct and indirect CP violating asymmetries in D_0 decays [6] [7]. Higgs sector is also under investigation, with limits on Higgs spin and parity, searches for fermiophobic Higgs and limits on charged Higgs boson production [8] [9] [10].

2. CDF Computing Model for the long term future

CDF computing architecture is evolving towards a model which will allow data access and analysis in the long term future. At Tevatron a data preservation project has been implemented in 2014. The project is divided into the following areas:

- data preservation: all CDF data have been migrated to the most recent tape technology (T10Kd);
- data access: CDF moved to a new data handling system, *SAMWeb* [11], based on http protocol for communication between the database and the experiment framework. This new system is easier to maintain and will be used by Intesity Frontier experiments at FNAL, so long term future support is guaranteed;

- analysis software: a version of CDF software based on SL6 operating system has been released in 2014 and will be the legacy software release for the long term future;
- data analysis: a new submission system, *jobsub* [12] has been developed to allow Intensity Frontier experiments to submit their jobs to the Open Science Grid. CDF moved to this new system to submit data analysis jobs to FNAL computing resources. As for SAMWeb, this system will be supported for many years to come.
- documentation: all CDF internal notes have been migrated to Inspire and CDF webpages updated.

3. CDF computing at CNAF: the long term future data preservation project

CNAF has been one of the major contributors to CDF computing outside Fermilab in the past and it now maintains a leading role in the data preservation effort. CDF has dedicated resources at CNAF: 8000 HS06 of computing power, 400 TB of disk and a set of machines for data access and analysis services. In 2014, CNAF contributed to CDF computing with the implementation of a long term future preservation project for CDF data. The project is being implemented in collaboration with CDF experiment and within the DPHEP collaboration. This is the first project funded by INFN on long term data preservation and will serve as a prototype for other experiments hosting their data at CNAF and other INFN sites. The project is divided into two main areas: bit (data) preservation and analysis framework development.

3.1. Bit Preservation

During 2014, 4 PB of CDF data (raw data and ntuples) were copied from FNAL using a dedicated link (see fig. 1). A mechanism able to copy the data at 5 Gb/s rate and store it in CNAF tape library has been setup in the second half of 2013 and used throughout 2014 to copy the data from FNAL. A dedicated 10Gb/s link and a reserved network allowed to manage and monitor CDF data movement independently from CNAF Tier 1 network resources and to have a secure high speed channel always available for data transfers, with no sharing of any resource. The storage layout consists of a pool of disks managed by GPFS, a tape library infrastructure for the archive back-end and an integration system to transfer data from disk to tape and vice versa. The CNAF Tier 1 storage solution is GEMSS [13], an integration of GPFS, TSM and StoRM. A single 10 Gb/s GridFtp Server, connected directly through the Storage Area Network (SAN) to the CDF GPFS file system disks and to the CNAF Tier 1 network switch 10 Gb backbone is used for the data copy. This allowed a plain method for transferring data from Fermilab to CNAF through a single point. In fig. 2 shows the data transfer rate during 2014.

3.2. Analysis framework

In order to access and process the data in the long term future it is essential to have an infrastructure that allows to use the experiment software. The goal is to make the software and data available and functional for many years in the future, beyond CDF collaboration.

In the current CDF analysis framework at CNAF, the user submits a job from his/her user area. The job submission system contacts the data access machine (SAM station) to access the data. Data is sent to worker nodes together with a tarball containing the analysis code. If needed, e.g. for MC production, detector and run conditions are retrieved from a dedicated database. All these services are already installed at CNAF and for the long-term future we plan to replicate this system as much as possible, upgrading the services to the latest versions. We expect in the long-term future data will be accessed and processed very rarely. For this it is necessary to study a solution which is robust but at the same time allows to minimize the needed resources (number of physical machines within the framework). In the analysis framework at CNAF the only real machine, except the storage system, is the database, which is

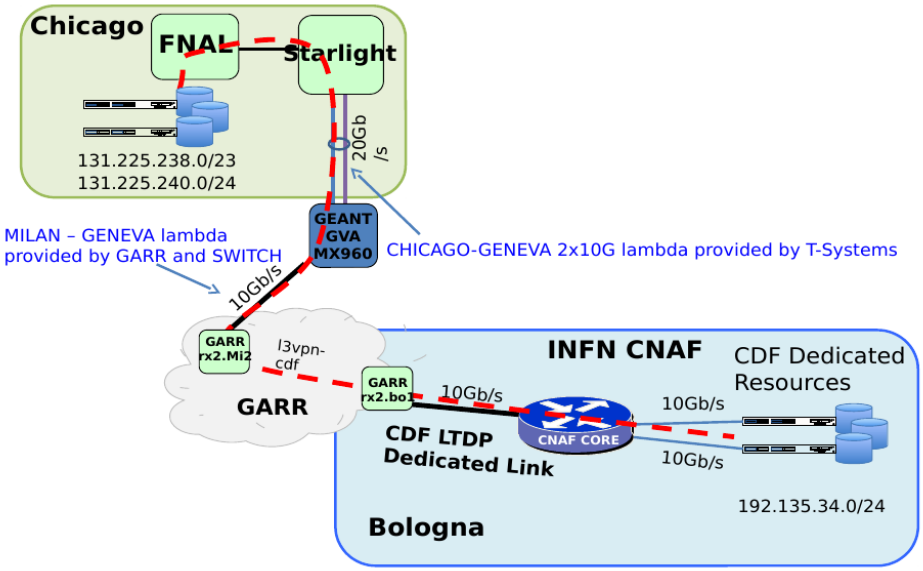


Figure 1. Layout of the FNAL-CNAF copy network.

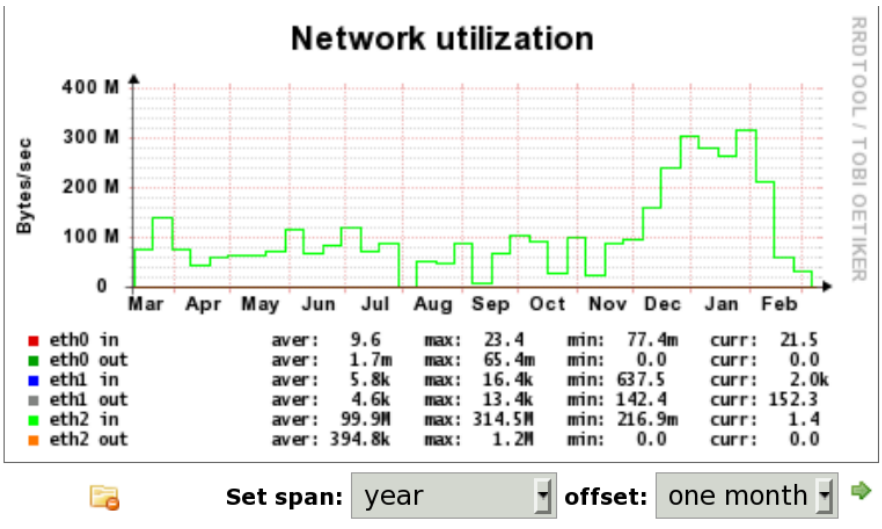


Figure 2. Data transfer rate from Fermilab to CNAF during 2014.

currently located at FNAL. All other services are instantiated on virtual machines. The virtual machines will be handled by WNoDeS [14], an INFN-developed framework that makes possible to dynamically allocate virtual resources out of a common resource pool, in order to instantiate the virtual machines when a job is started.

For the data access, the data access machine has been upgraded to SAMWeb. CNAF tape system is transparent for SAMWeb: upon a request to access a file, the file is recalled from tape on a disk cache before being sent to the final destination (worker node or user area). As already stated in the previous section, in the long term future we foresee intermittent access to CDF data, so the necessary disk cache will be allocated on-demand using CNAF resources in opportunistic mode.

Future analysis on CDF data will use a software legacy release based on SL6 distributed

through CVMFS. At CNAF as a first step squid proxy servers have been setup to access the CVMFS server located at FNAL. In a second phase the FNAL CVMFS server will be replicated at CNAF.

As far as job submission is concerned, the system currently installed for CDF at CNAF is based on glideinWMS [15] to exploit computing resources at CNAF and additional LCG resources at different Tier-2 sites in Italy and other European countries. In 2014 the current system has been maintained, upgrading it to use the legacy release. In 2015 we plan to move to the new job submission system developed at FNAL, jobsub.

3.3. Conclusions

During CDF RunII (2001-2011) operations CNAF has been one of the major computing centers for the experiment. A portal to access CNAF Tier-1 and other LCG resources is hosted at CNAF, together with dedicated data processing and storage resources. Now, three years after the end of data taking, CDF has entered the data preservation phase and CNAF is contributing with the implementation of a long term future data preservation project: complete copy of all CDF data is now available at CNAF, and the setup of a long term future analysis framework started at the end of 2014 and will be completed in 2015.

References

[1] Abe F *et al.* (CDF Collaboration) 1995 *Phys.Rev.Lett.* **74** 2626–2631 (*Preprint* `hep-ex/9503002`)
[2] Abulencia A *et al.* (CDF Collaboration) 2006 *Phys.Rev.Lett.* **97** 242003 (*Preprint* `hep-ex/0609040`)
[3] Aaltonen T *et al.* (CDF Collaboration) 2009 *Phys.Rev.Lett.* **103** 092002 (*Preprint* `0903.0885`)
[4] Aaltonen T A *et al.* (CDF) 2014 (*Preprint* `1410.4909`)
[5] Aaltonen T A *et al.* (CDF, D0) 2014 *Phys.Rev.Lett.* **112** 231803 (*Preprint* `1402.5126`)
[6] Aaltonen T A *et al.* (CDF) 2014 *Phys.Rev.Lett.* **113** 242001 (*Preprint* `1403.5586`)
[7] Aaltonen T A *et al.* (CDF) 2014 *Phys.Rev.* **D90** 111103 (*Preprint* `1410.5435`)
[8] Aaltonen T *et al.* (CDF, D0) 2015 (*Preprint* `1502.00967`)
[9] Aaltonen T A *et al.* (CDF) 2014 *Phys.Rev.* **D89** 091101 (*Preprint* `1402.6728`)
[10] Aaltonen T A *et al.* (CDF) 2015 *Phys.Rev.Lett.* **114** 141802 (*Preprint* `1501.04875`)
[11] Lyon A L, Illingworth R A, Mengel M and Norman A J 2012 *J.Phys.Conf.Ser.* **396** 032069
[12] Box D 2014 *J.Phys.Conf.Ser.* **513** 032010
[13] Ricci P P, Bonacorsi D, Cavalli A, Dell'Agnello L, Gregori D *et al.* 2012 *J.Phys.Conf.Ser.* **396** 042051
[14] WNoDeS URL `http://web.infn.it/wnodes/index.php/wnodes`
[15] Amerio S, Benjamin D, Dost J, Compostella G, Lucchesi D *et al.* 2012 *J.Phys.Conf.Ser.* **396** 032001

The INFN Tier-1: networking

S Zani, D De Girolamo and L Chiarelli

1. Introduction

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

2. Wide Area Network

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

As shown in Figure 1, CNAF is connected to the WAN via GARR/GEANT essentially with two physical links:

- A general IP with a 10 Gb/s connection via GARR and GEANT
- A link to WLCG destinations, which has been upgraded to 40 Gb/s, shared between the LHC-OPN Network for Tier0-Tier1 and Tier1-Tier1 traffic and LHC-ONE network mainly for Tier2 and Tier3 traffic.

2.1. Routing changes in 2014

During 2014 the traffic between the Tier1 centres hosted at CNAF, KIT and IN2P3 has been moved from LHC-OPN to LHC-ONE in order to optimize the network resources available within the GEANT network and between the GEANT network and the national NRENs. In Italy the peering between the GARR network and the GEANT network is made of $2x100$ Gb/s links; moving part of some of the inter-Tier1 traffic on that link frees bandwidth on the LHC-OPN uplink for Tier0-Tier1 transfers and the rest of Tier1-Tier1 traffic.

2.2. Possible Capacity Evolution

The WLCG link can be upgraded any time to 60 Gb/s ($6x10$ Gb/s) and GARR has in its roadmap a 100 Gb/s link between the GARR POP at CNAF and GEANT (the fibers are 100 Gb-ready but an upgrade of the GARR optical devices is needed).

An upgrade of the general IP link to 20 Gb/s is planned for 2015.

3. Local Area Network

The Tier1 LAN is essentially a star topology network based on a fully redundant switch router (Cisco Nexus 7018), used both as a core switch and an access router for LHC-OPN and LHC-ONE networks. In addition, more than 100 aggregation ("Top Of the Rack") switches are installed, with Gigabit Ethernet interfaces for the Worker Nodes of the farm and 10 Gb Ethernet

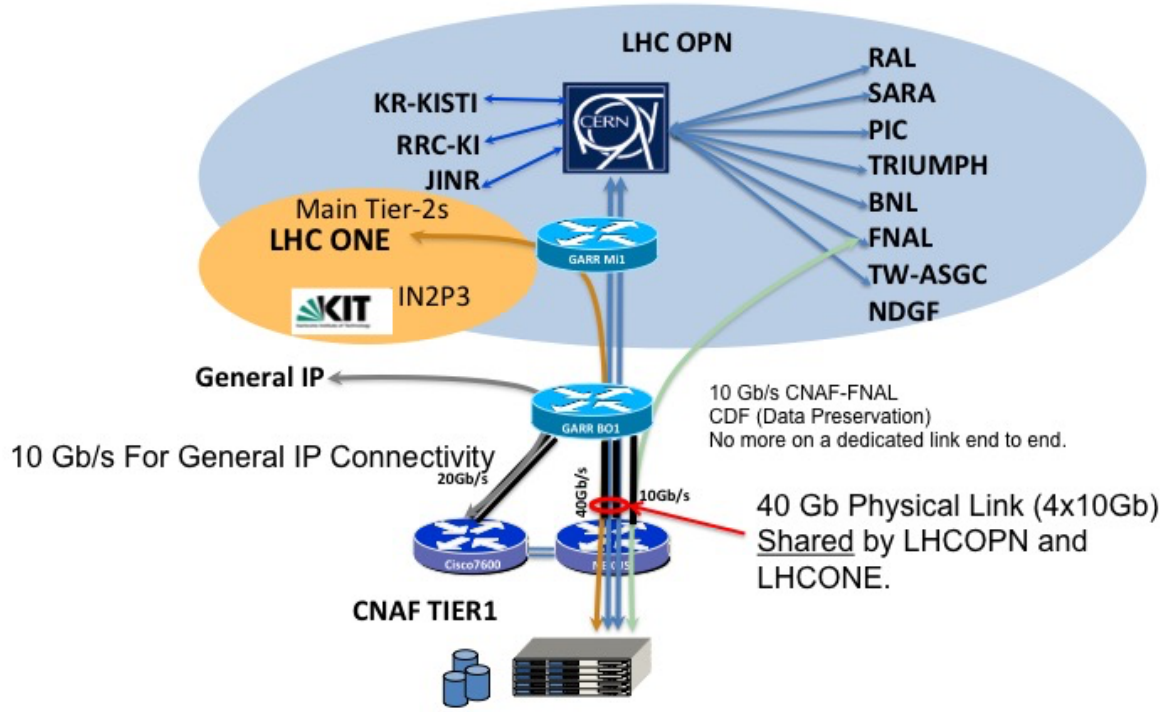


Figure 1. WAN connection schema

interfaces used as uplinks to the core switch. Disk servers and GridFTP servers are directly connected to the core switch at 10 Gb/s.

General Internet access, local connections to the offices and INFN national services provided by CNAF are managed by another network infrastructure based on a Cisco 7606 Router, a Cisco Catalyst 6509 and an Extreme Networks Black Diamond 8810. CNAF owns an IPv4 B class (131.154.0.0/16) and a couple of C classes for specific purposes: half of the B class is used for Tier1 resources and the other half is used for all the other services, thus providing sufficient IP addresses. The private address classes are used for IPMI and other internal services.

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

4. Network monitoring and security

In addition to the perfSONAR-PS and the perfSONAR-MDM infrastructures [1] required by WLCG, the monitoring system is based on several tools organized in the "Net-board", a monitoring dashboard realized at CNAF (see Figure 2). The Net-board integrates MRTG [2], NetFlow Analyser [3] and Nagios [4], with some scripts and web applications to give a complete view of the network usage, which allows to promptly identify possible problems. The alarm system is based on Nagios.

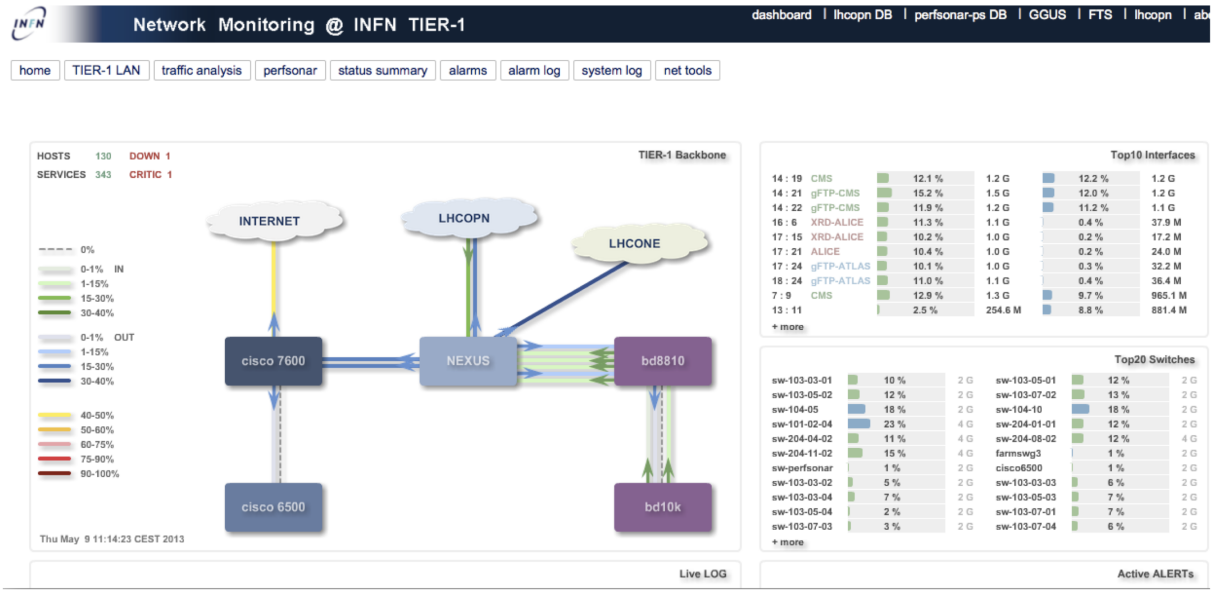


Figure 2. A screenshot of the monitoring Net-board

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

The network group, in coordination with GARR-CERT and EGI-CSIRT, also takes care of security incidents at CNAF, both for compromised systems or credentials and discovered vulnerabilities of software and Grid middleware, cooperating with all involved parties to identify and apply the appropriate solutions.

5. Software Defined Networks

Software Defined Networks (SDN) provide a level of abstraction on top of the physical network, allowing to delegate some decisions to application software.

The CNAF Network department has been investigating SDN technology since 2013 [5], with the purpose to evaluate its use in a number of present and foreseable scenarios typical of the centre, including: physical network abstraction, integration with virtualization stacks like OpenStack, centralization and automation of network configuration, network configuration on-demand, seamless integration of geographically-separated sites.

The experimentation is performed on a dedicated infrastructure, NetLab@CNAF, shown in Figure 3.

During 2014 the investigation has concerned mainly two products:

- Programmable Flow Control by NEC [7]
- OpenDaylight [8]

Both are based on the OpenFlow standard protocol [6] and provide convenient interfaces for the definition of virtual routers and switches for the establishment of Virtual Tenant Networks, especially useful when assigning cloud resources to different users, which should appear separated network-wise.

References

[1] perfSONAR, `http://psps.perfsonar.net/`

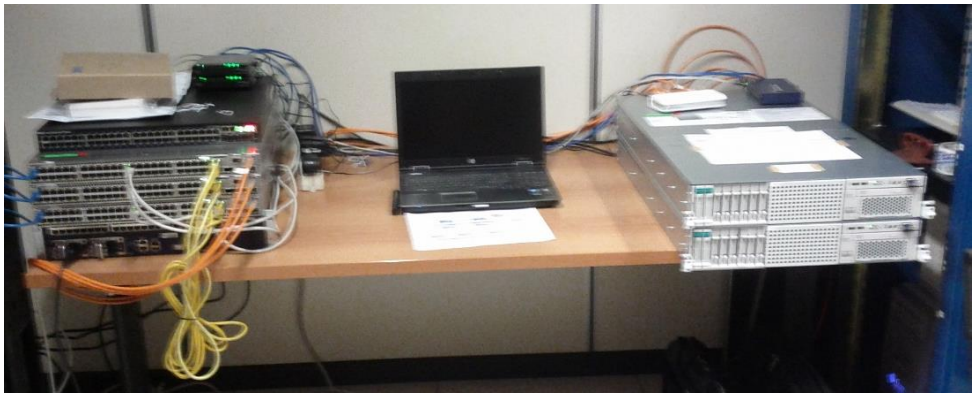


Figure 3. The NetLab@CNAF

[2] MRTG – Multi Router Traffic Grapher, http://it.wikipedia.org/wiki/Multi_Router_Traffic_Grapher
[3] NetFlow, http://en.wikipedia.org/wiki/NetFlow
[4] NAGIOS, http://www.nagios.org
[5] L. Chiarelli, *Software Defined Networks: Studio del modello e implementazione di una rete basata sullo standard OpenFlow, Master Thesis*
[6] OpenFlow, https://www.opennetworking.org/
[7] NEC Programmable Flow Control, http://www.necam.com/SDN/
[8] OpenDaylight, https://www.opendaylight.org/

National ICT infrastructures and services

S Antonelli, S Longo, R Veraldi, and S Zani

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

1. Strategic and critical high-priority services

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

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

2. Medium-priority non-critical services

- National Mailing lists: we manage over 1000 lists for the INFN domains with a centralized system based on Sympa.
- Centralized Web Site management: this is a service which allows people to develop a web site for their particular project or experiment, in which INFN may be involved. We chose a common CMS (Content Management System) based on JOOMLA for everyone. At the moment almost 90 sites are managed.
- NTP service: we manage one of the three Network Time Protocol servers for INFN computers.
- Backup service for the INFN Certification Authority (CA): a daily encrypted backup mirrors data of the INFN CA hosted at the Florence INFN site.

- Eduroam and TRIP management: we coordinate the eduroam and TRIP 802.1x-based WiFi national infrastructure. Actually TRIP is the precursor of eduroam inside INFN, uses the same technology and is widely adopted by many internal INFN users.
- Centralized License distribution: we deployed an infrastructure for license distribution related to several software packages: Ansys, Comsol, Autodesk, NX, Esacomp, Mathematica, Cliosoft, Mathlab. It is based on Linux Flex servers. The management of the infrastructure is done in collaboration with colleagues from other INFN sites: Padova, LNF, Napoli and Pisa.
- Activation of Microsoft Operating Systems and Office: we manage the KMS server for Windows Vista, Windows 7 and Windows 8 Operating Systems and Microsoft Office activation for all the INFN computers.
- National Multimedia Services: the following services are managed:
  - Asterisk server for INFN phone conferences
  - MCU H.323 for INFN video conferences
  - Real Networks and Flash video servers for streaming INFN events
  - SeeVogh reflector
  - vidyo router, part of the CERN vidyo infrastructure
- Centralized INFN Authentication and Authorization (AAI): we host three servers belonging to the INFN-AAI Service.
- AFS and Kerberos: we manage three servers for the INFN national AFS service.

3. Expanding services

- A new DNS architecture is available that meets the strict requirements concerning the DNS resolution for high-availability services. The new DNS architecture is based on a pool of DNS servers distributed across INFN sites and implementing a multi-master ISC bind solution with underlying Dynamic Loadable Zones and GaleraDB technologies. At the moment the service is used by the so-called "INFN Corporate Cloud" and shortly it will be further extended to other services.
- An official INFN Document Management System, based on Alfresco, is available. A section (a *site* in Alfresco parlance) is available for each INFN site, where documentation and other material related to local services and experiments can be stored. Various scientific collaborations are already using this service for their own documents: BELLE-II, BESIII, !CHAOS, GINGER, LHCb Italia, PRISMA, ReCaS, ReCaS-PRISMA, SPES e SR2S. Additionally it is used for the INFN official collection of documents.
- Disaster recovery: CNAF coordinates a working group for disaster recovery of important services. The involved sites are CNAF and LNF. At the moment the project is focused and limited to the INFN Information System.
- Synchronization and desktop backup Pandora: the service implements a Dropbox-like solution built at CNAF with standard open-source tools. It includes a synchronization client, multi-platform access via a web interface, data sharing through web URLs. It also offers a WebDAV interface and the possibility to create mini-sites for the distribution of images of operating systems and licensed software that is made available to the entire community. The service has been put into production and is available to all users belonging to the INFN-wide AAI.
- Collaboration tools aimed at supporting proper software engineering practices during software development.

4. HW and SW architecture

The National ICT Infrastructures and Services are mainly implemented using virtualization technology inside cluster architectures. We phased out the CentOS Cluster architecture in favour of VMWare and oVirt architectures.

Inside oVirt we are managing 3 different data centers abstractions:

- ServNaz: dedicated to production services INFN-wide and local to CNAF
- ServNaz-testing: holds several virtual machines for testing, development purpose and semi-production services
- ISSS: dedicated to software development infrastructure with a continuous integration environment for different Linux distributions
- VDI: this is a new infrastructure dedicated to Virtual Desktop needs belonging to INFN Administration Divisions.

Software Services and Distributed Systems

CNAF activities in the !CHAOS project

E. Fattibene[1], D. Salomoni[1], P. Veronesi[1]

in collaboration with:

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

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

E-mail: enrico.fattibene@cnaf.infn.it

Abstract. The "!CHAOS: a cloud of controls" project has been financed by MIUR (Italian Ministry of Research and Education) and aims to develop a new concept of control system and data acquisition framework (DAQ) by providing, with a high level of abstraction, all the services needed for controlling and managing a DAQ through a large, distributed infrastructure. The final product of the project is expected to be a prototype of a dynamic, on-demand cloud-based infrastructure, devoted to accelerator control systems. The !CHAOS application will be deployed on OpenStack, an open source cloud framework that can be executed on open source platforms and that has strong support from the industry. This paper presents the activity carried on by CNAF in the framework of the !CHAOS project.

1. The "!CHAOS: a cloud of controls" project

The !CHAOS activity was originally born within the context of High Energy Physics (HEP) as a candidate of Distributed Control Systems (DCS) and Data Acquisition (DAQ) for the SuperB experiment. In 2014 it evolved to the project "!CHAOS: a cloud of controls" [1] supported by MIUR and developed by INFN through its four sites Laboratori Nazionali di Frascati (LNF), Laboratori Nazionali del Sud (LNS), Sezione di Padova and CNAF. National Instruments and ESCO also support the project for the implementation of several case studies. The aim of !CHAOS is to develop by the end of 2015 a prototype of a dynamic application based on a cloud infrastructure offering "controls as a service" also to society and industries.

The project is divided in 4 major research and development activities, besides a coordination, communication and documentation activity. Each of these activities is carried out in a specific Work Package:

- WP2 - Framework development: development of routines of the common architecture aimed to ensure the communication of data among the five nodes of the system: data acquisition (CU – Control Unit), presentation (UI – User Interface), proxies/indexing/storage (CDS - !CHAOS Data Service), data handling (EU – Execution Unit) and system state information (MDS - Metadata Service).
- WP3 - Drivers and CU development and integration: development of routines devoted to the drivers implementation and CU development; integration and deployment of these components for the use cases; framework tests and debugs.
- WP4 - Use cases implementation: software and hardware implementation of the specific CU, EU and UI for the three main use cases: LNF Beam Test Facility (BTF) DAQ, accelerator devices and diagnostic controls; LNS beam source control; LNF Touschek auditorium environmental control.
- WP5 - IT infrastructure development and implementation: analysis and implementation of the cloud infrastructure and services to offer the !CHAOS framework as a service. This is the WP in which CNAF people are involved.

2. The !CHAOS cloud infrastructure design

In the scope of this project and in particular in the WP5, the architectural design of the !CHAOS software and the use cases have led the !CHAOS team to choose a cloud Infrastructure as a Service (IaaS) as the way to deploy both backend and frontend services.

The design of the application deployment environment took into account the following requirements:

- high availability and reliability
- scalability
- disaster recovery
- on-demand deployment according to user requests
- automatic setting of the number of instances and the size of components according to the parameters chosen by the user
- auto-scaling on the basis of monitoring information

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

The !CHAOS [3] main services (CDS and MDS) rely on backend services, such as a file system, a database and a shared cache object. The !CHAOS team studied a solution for each of these services (both the common and the !CHAOS specific ones), to integrate them in a generic IaaS based on OpenStack. The final goal of this activity is to produce a Software as a Service (SaaS) implementation of !CHAOS, in order to give users the possibility to automatically deploy a complete !CHAOS instance on a private or public cloud environment.

Figure 1 represents a simplified design schema of the !CHAOS deployment on OpenStack. On top of OpenStack, the backend services should be automatically provided by the cloud environment, as Platform as a Service (PaaS) components; the !CHAOS frontend services (CDS and MDS) exploit the

virtual instances of PaaS components and can run as unique or multiple instances. The frontend services have to communicate in a bidirectional way with the remote !CHAOS clients, such as the CU and the UI. Since the remote clients could not be identified by a public IP address, a VPN service will be deployed within the OpenStack cloud to allow network traffic to flow to and from the frontend services.

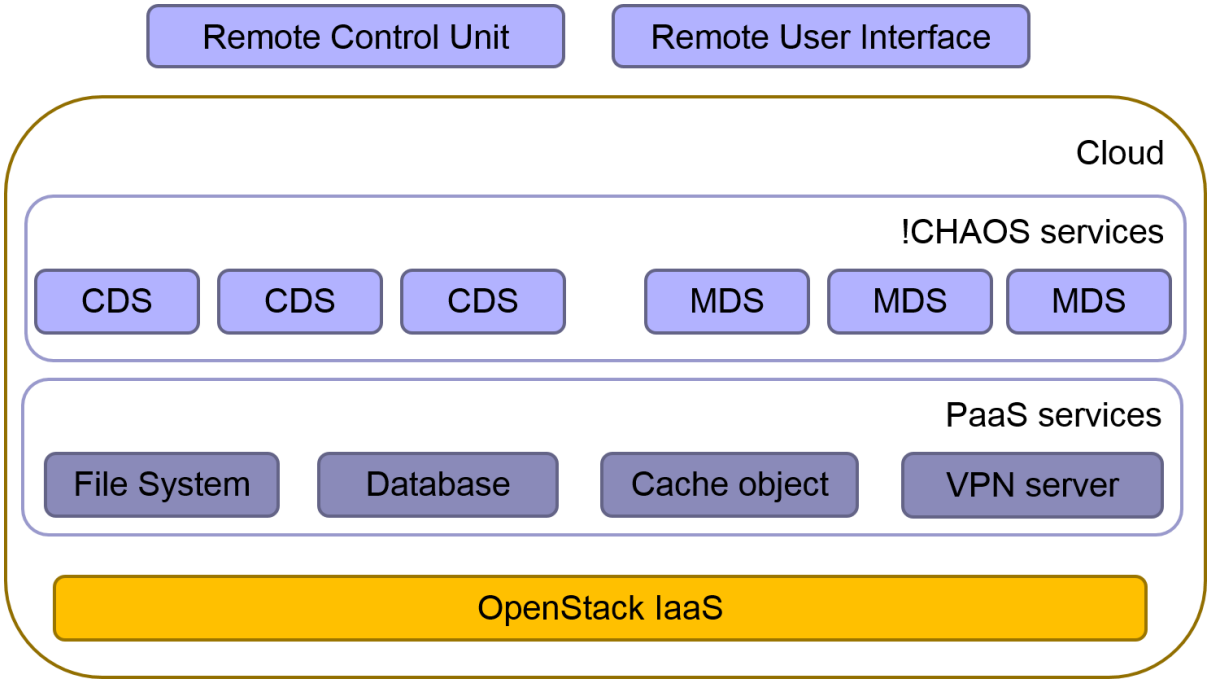


Figure 1: Design schema of the !CHAOS cloud deployment

In the !CHAOS software version available at the end of 2014, the frontend services depended on a POSIX file system, on a MongoDB [4] NoSQL database in a clustered configuration and on a Couchbase [5] cluster as shared cache object. To supply these PaaS components the !CHAOS team investigated the capability of Heat [6], the OpenStack component that implements the function of infrastructure orchestration, i.e. a programmatic approach to create and deploy full stack configurations. The idea is to create a set of Heat templates to automatically deploy the different PaaS elements. In order to deploy a MongoDB cluster, Trove [7] (the Database as a Service OpenStack component) was evaluated but the current version (the last version of 2014 is part of the OpenStack Juno release [8]) is not able to deploy a MongoDB cluster composed by a configurable number of shards.

A first Heat template has been developed to provide a file system based on a Ceph [9] virtual infrastructure composed by 3 Monitor, 3 Object Storage Device (OSD), with a configurable backing store size, and 1 Metadata Service (MDS).

To deploy the frontend services, ad-hoc virtual images have been produced, based on an Ubuntu 14.04 operating system, and tested on the CNAF OpenStack infrastructure.

3. Future work

During 2015 the CNAF people involved in the !CHAOS project will work on deploying the !CHAOS frontend and backend components on an OpenStack-based infrastructure, following the design architecture. Once !CHAOS is available on OpenStack, the !CHAOS team will work toward implementing a highly available solution, by deploying the instances of backend and frontend services on different availability zones (i.e. logical groupings of computing resources). The deployment of services on different regions (i.e. IaaS running on different data centres) will also be tested in order to provide geographical redundancy for services. These tests will be performed on the OpenStack cloud infrastructures running at CNAF and LNF.

Another planned development is the exploitation of the Heat APIs in order to dynamically size the !CHAOS infrastructure, on the basis of parameters chosen by the user (e.g. through a web dashboard). A solution for auto-scaling of the infrastructure will therefore be developed, in order to add (or remove) instances of !CHAOS services in case of need. This feature will use the Heat capability to auto-scale deployed infrastructures on the basis of monitoring metrics calculated by Ceilometer [10], the OpenStack monitoring service.

References

[1] F. Antonucci et al, "!CHAOS: a cloud of controls -MIUR project proposal", INFN -14-15/LNF; https://www.lnf.infn.it/sis/preprint/
[2] OpenStack, www.openstack.org/
[3] L. Catani et al, "Introducing a new paradigm for accelerators and large experimental apparatus control systems", Phys. Rev. ST Accel. Beams 15, 112804 –Published 29 November 2012
[4] MongoDB, docs.mongodb.org/manual/core/introduction/
[5] Couchbase, www.couchbase.com/nosql-databases/couchbase-server
[6] Heat, wiki.openstack.org/wiki/Heat
[7] Trove, wiki.openstack.org/wiki/Trove
[8] Juno, www.openstack.org/software/juno
[9] Ceph, www.ceph.com
[10] Ceilometer, wiki.openstack.org/wiki/Ceilometer

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

M. Manzali [2 3], T. Chiarusi [1], M. Favaro [1 2 3], F. Giacomini [2], C. Pellegrino [1 4] on behalf of the KM3NeT-Italy Collaboration

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

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

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

1. Introduction

KM3NeT-Italy is an INFN project supported with Italian PON[1] fundings for building the first part of the Italian node of the KM3NeT neutrino telescope [1] [2]. The detector will be placed in the Ionian Sea (Italy) at 3500 m of depth. The KM3NeT-Italy detector will be made of 700 optical modules (OMs), each one containing a 10" PMT and the readout electronics. The OMs are organized in vertical structures called towers. Each tower is composed of 14 horizontal bars with 6 OMs each. The detection principle is based on the measurement of the Cherenkov light from high-energy neutrino induced charged particles produced within a fiducial volume around the telescope [3]. The "all data to shore" approach is assumed to reduce the complexity of the submarine detector, demanding for an on-line trigger integrated in the data acquisition system running in the shore station, called TriDAS [4]. Due to the large optical background in the sea from ^{40}K decays and bioluminescence, the throughput from the sea can range up to 30 Gbps. This puts strong constraints on the performances of the TriDAS processes and the related network infrastructure.

[1] The research leading to these results has received funding from the Programma Operativo Nazionale "Ricerca & Competitività" 2007-2013.

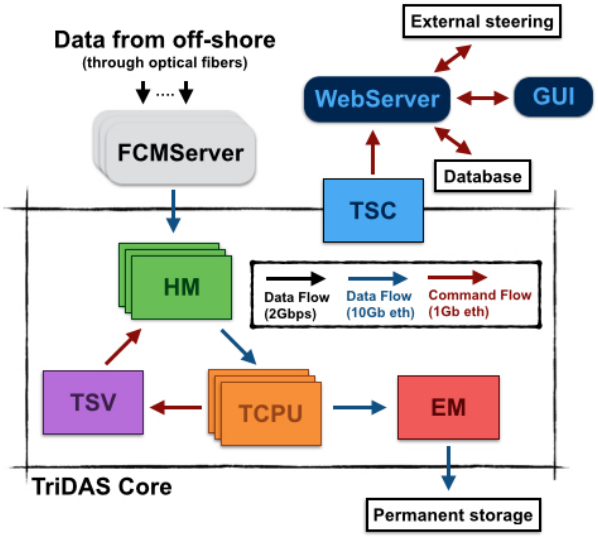


Figure 1. Scheme of TriDAS components and their interactions with external services.

2. The evolution of the TriDAS software

TriDAS, designed for the prototype tower deployed by KM3NeT-Italy and running at the offshore station of Portopalo, has been checked with a long period of data acquisition, started more than two years ago. After an important upgrade of the dedicated computing infrastructure and a revision of the existing software (Fig. 1), TriDAS will also be used for the data acquisition of the first block of the KM3NeT-Italy telescope, that is composed of 8 towers. In order to meet the new requirements, the revision of TriDAS has involved the adoption of modern software design solutions and high-level libraries and the update to the new raw data format.

2.1. Floor Control Module Server (FCMServer)

The FCMServer represents the interface of TriDAS with the data from the off-shore detector. It performs the read-out of data coming from a number (up to 4) of floors through a dedicated ASIC [5]. After having performed a consistency check, it sends the data to the connected HitManager.

2.2. Hit Manager (HM)

The HMs are the first step of aggregation of the detector. Each HM receives data from a set of floors called sector: the sector's dimension can be dynamically decided before each run, but it is seen that the best fit is 1/2 or 1 tower. The main task of HMs is to split data into slices of a fixed time duration (about 200 ms) called SectorTimeSlices and to send them to a specific TriggerCPU, a dedicated object for the last step of aggregation and online analysis.

2.3. TriggerCPU (TCPU)

The TCPUs are responsabile for the last step of data aggregation and online analysis. Each TCPU receives from Hit Managers all the SectorTimeSlices related to a specific time interval and creates the TelescopeTimeSlice, rappresenting the state-of-the-art of the whole detector in a specific time interval. Once that a TelescopeTimeSlice is ready, the TCPU applies first and second level triggers in order to find interesting events and to remove backgroud noise. Finally the TelescopeTimeSlice with related events is sent to the Event Manager and stored.

2.4. Event Manager (EM)
The EM is the software component of the TriDAS deputed to the storage of triggered data. A single EM process collects triggered data from the whole TCPU set and performs data writing on local storage. Data are written in strict temporal order in so-called PostTrigger files (PT files), which maximum size is fixed by means of configuration parameter. Each PT file also contains all the run setup parameters and the detector geometry.

2.5. TriDAS SuperVisor (TSV)
The TSV is the process responsible to decide wich SectorTimeSlices have to be sent to wich TCPU. When a TCPU can handle a TelescopeTimeSlice, it sends a token to the TSV: this last one chooses a SectorTimeSlice ID and communicates to all the HMs that ID and which TCPU has sent the token. HMs are constantly waiting for comunications from the TSV and when a SectorTimeSlice ID is received they sent the requested SectorTimeSlice to the related TPCU. Finally the TSV has a fault tollerant mechanism that permit to resend a SectorTimeSlice ID to a different TPCU if the previous send has failed.

2.6. Tridas Controller (TSC)
The TSC is the software interface that permits to control the entire TriDAS enviroment. Its purpose is organize and control the launch of each sotware in order to allow a correct acquisition and real time analysis of the data coming from the sub-marine detector. The TSC is modeled on top of a state machine that represent a stable situation of the system. Each state indicates a well know situation of the Tridas. With this model the TSC is able to understand and keep managed the state of the indipendent software that form the TriDAS. Moreover it can recover some faulty situations with some retry mechanisms.

2.7. GUI
The TriDAS enviroment, as described so far, is a closed system with a unique control access point located on the TSC. The TriDAS System Controller, in fact, permits to a only one actor at once to comunicate and query the TriDAS. But we want a collective and collaborative system in order to allow multiple reasearchers to view how the acquisition is performing and the health of the TriDAS. In order to achieve this functionality it has been made a GUI that sets itself as broker between the TSC and the rest of the world. It has the capability to allow only one TSC controller at once (human or not). It mirrors the TSC information so it can serve to multiple client in a faster way. On top of this the GUI is developed as single-page application (SPA) that is one of the most modern implementation of online applications.

2.8. Tools and Libraries
New tools for software development and modern design solutions have been adopted in order to con- tinuously improve software quality at decreasing cost [6].

3. Tests at the KM3NeT Bologna Computing Infrastructure

The TriDAS implementation is currently being tested with a setup realized in the KM3NeT DAQ laboratory at the INFN-Sezione di Bologna. The various TriDAS processes run on a dedicated computing farm interconnected according to the same network layout that will be used in production. The data stream incoming to the TriDAS is obtained with both an experimental setup, with the real electronics for two complete floors attached to two FCMServers (Fig. 2), and a set of FCMServer simulators which can run on the farm, allowing to simulate the data flow from all the towers.

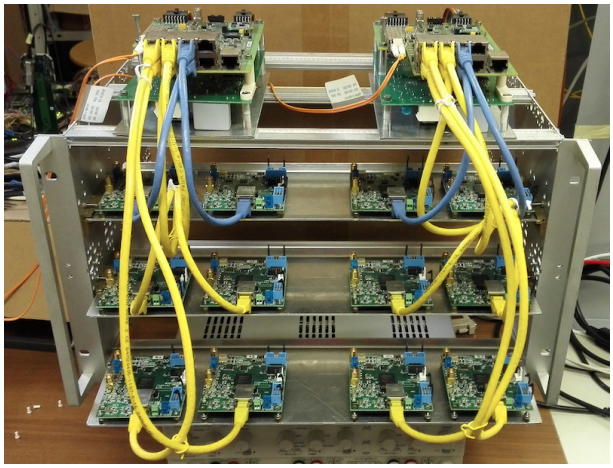


Figure 2. The photo shows a crate containing the electronics for two complete floors at the Bologna test bench.

4. Conclusion

The improved TriDAS for the 8 towers is currently under an intense test session. The aim is validating the system scalability and stability with long duration runs and with varying the the incoming throughput via the FCMServer simulators. At the same time, the trigger framework is optimized for enhancing the performances of the TCPU processes. Finally, new trigger algorithms for different kind of physics searches (e.g. multi-messangers alerts, high energy neutrino induced showers) are under development and tests.

References

[1] T. Chiarusi, M. Spurio, "High-Energy Astrophysics with Neutrino Telescopes", Eur. Phys. J. C65, 649 (2010)
[2] KM3NeT web site: http://www.km3net.org/home.php
[3] S. Aiello et al., "Measurement of the atmospheric muon depth intensity relation with the NEMO Phase-2 tower", Astroparticle Physics, Vol. 66, Pag. 1-7 (2015)
[4] C. Pellegrino, F. Simeone, T. Chiarusi, "The Trigger and Data Acquisition for the NEMO-Phase 2 tower", AIP Conference Proceedings 1630, 158 (2014)
[5] A. Lonardo et al., "NaNet: a configurable NIC bridging the gap between HPC and real-time HEP GPU computing" JINST 10 C04011 (2015)
[6] F. Giacomini et al., "An integrated infrastructure in support of software development", Journal of Physics: Conference Series 513 (2014) 062018.

The 40 MHz trigger-less DAQ for the LHCb upgrade

D H Campora Perez[1], A Falabella[2], D Galli[3,4], F Giacomini[2], V Gligorov[1], M Manzali[2], U Marconi[4], N Neufeld[1], A Otto[1], F Pisani[1,6] and V M Vagnoni[4]

E-mail: antonio.falabella@cnaf.infn.it
[1] CERN, Geneva, Switzerland
[2] INFN CNAF, Bologna, Italy
[3]Università Bologna, Bologna, Italy
[4] INFN Sezione di Bologna, Bologna, Italy
[5] Università Ferrara, Ferrara, Italy
[6] Università la Sapienza, Roma, Italy

Abstract. The LHCb experiment will undergo a major upgrade during the second long shutdown (2018 - 2019), 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 read-out at the LHC bunch crossing frequency of 40 MHz, using an entirely software based trigger. A new high-throughput PCIe Generation 3 based read-out board, named PCIe40, has been designed on this purpose. The read-out 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. Possibile technologies candidates for the foreseen event-builder under study are InfiniBand and Gigabit Ethernet. In order to define the best implementation of the event-builder we are performing tests of the event-builder on different platforms with different technologies. For testing we are using an event-builder evaluator, which consists of a flexible software implementation, to be used on small size test beds as well as on HPC scale facilities. The architecture of DAQ system and up to date performance results will be presented.

1. Introduction

LHCb is one of the four main experiments in operation at the Large Hadron Collider at CERN [1]. The LHCb main goals are to make precise measurements of CP violation parameters and studying rare decays of b and c-quark hadrons. It has already performed successful physics measurements exploiting the Run 1 data and it has just started the Run 2 data taking. The LHCb detector will be upgraded during the Long Shutdown 2, foreseen by 2018-2019, in order to record data at the maximum design LHC energy of 14 TeV, running at an instantaneous luminosity of $2 \times 10^{33}\mathrm{cm}^{-2}\mathrm{s}^{-1}$. The key features of the upgraded detector are the trigger-less readout system and the full software trigger [2].

2. Trigger evolution

The trigger system of the present LHCb detector consists of a two stage-trigger: the fixed latency, near-detector, Level-0 (L0) hardware trigger and a software trigger, the so called High-Level Trigger (HLT). Aim of the L0 trigger is to reduce the bunch crossing rate to $\sim$ 1 MHz,

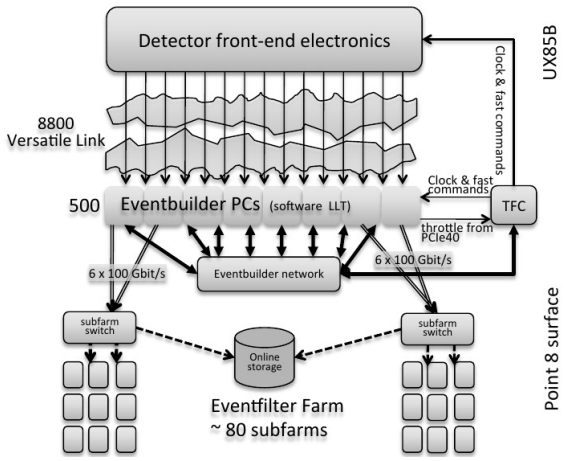


Figure 1. The architecture of the upgraded LHCb readout-system.

at which the detector is read-out. The HLT applies then more advanced selections, in order to further reduce the rate to ~ 5 KHz, for storage and offline data processing.

The upgraded LHCb detector will operate at a luminosity of $2 \times 10^{33}\text{cm}^{-2}\text{s}^{-1}$, five times higher with respect to Run 1[3]. The 1 MHz readout is a bottleneck because of the limited bandwidth and the limited hadronic trigger efficiency.

3. Upgrade DAQ implementation

The main feature of the LHCb upgraded detector is the trigger-less readout system, without any hardware trigger. The logic scheme of the readout system is shown in Figure 1, with the main components: The event builder, the Timing and Fast Control(TFC) system, The Experiment Control System (ECS) for monitoring and configuration, and the trigger Event Filter Farm (EFF).

Table 1. Constraints for the online system.

Event rate	40 MHz
Mean nominal event size	100 KBytes
Readout board bandwidth	up to 100 Gbit/s
CPU nodes	up to 4000
Aggregated throughput	32 Tbit/s

The design requirements for the implementation of the trigger-less readout system are summarised in Table 1. The aggregated bandwidth of the event builder network can be estimated given the foreseen nominal event size of 100 KBytes, assuming the maximum event rate of 40 MHz, to be of the order of 32 Tbit/s. The size of the event-builder farm can be estimated of the of the order 500 PCs, this by assuming the input rate of a server, through the PCIe40, is about 100 Gbit/s.

The event-builder network can be effectively implemented by using commercial local area network technologies such as Ethernet or InfiniBand. In the following we present the results of the first scalability tests of the event builder InfiniBand-based implementation. The InfiniBand standard is widely used in HPC clusters, since it provides cost-effective high-speed performance.

4. Event Building Performance Evaluator

In order to test the InfiniBand event-builder implementation we developed a performance evaluator software. The software evaluator emulates the behaviour of the Readout Unit (RU) and of the Builder Unit (BU). The RU receives event fragments from the detector and send them to the BU. The BU collects event fragments from all the RUs and assembles them into a complete event to be afterwards sent to the EFF. The orchestration of the event building is managed by an Event Manager (EM) unit. The EM sets a server as the BU, balancing the load among the servers. The focus of our test is on the RDMA (*Remote Direct Memory Access*) capabilities of the InfiniBand network. RDMA allows to write or read data by a user-space application without the intervention of the CPU and the OS, resulting in low latency communications. Among the possible different protocols that can be implemented, we tested the *push* mode. Running in push mode the data transfer is started by the RU (in pull mode it is started by the BU). The EM addresses events from the RUs to the current specific BU at a tuneable readout frequency (the expected visible event bunch-crossing rate at the foreseen luminosity being around 30 MHz).

5. Event builder nodes tuning

The InfiniBand adapters reach the best performance following the recommendations as explained at [4]. The PCIe motherboard configuration should be of generation 3.0, since it provides a bandwidth of $\sim$ 126 Gbit/s on 16 lanes slot. A second key aspect to be considered is the CPU power management. The CPU can be commanded by the Linux OS to enter into a low-power mode, called the "c-states", when in idle state. Another important aspect to be considered concerns the case of NUMA (*Non Uniform Memory Access*) architecture of the multi-processors systems. In this case, the best performance can be achieved running the application on the CPU-cores directly connected to the PCIe adapter.

In order to estimate the role of these effects we measured the data transfer rate on a basic system, consisting of one BU and one RUs in a point-to-point connection. The CPU and HCA (Host Channel Adapter) used for these tests are reported in Table 2.

Table 2. Test bed characteristic for the point-to-point rate measurements.

CPU	Intel Xeon CPU E5-2620 2.00GHz
HCA	Mellanox MCB194A-FCAT 56 Gbit/s (FDR)

To perform the tests we used the OpenMPI implementation of RDMA over InfiniBand [5]. The result of the tests can be seen in Figure 2. The Mellanox FDR HCA allows for 54.3 Gbit/s maximum transfer rate (taking into account data encoding). The measured transfer rate when the CPU power management is active, without binding the application to a specific CPU, are shown as blue dots. The measured data transfer rate turns out to be $\sim$ 10% of the maximum value. Binding the application to the CPU directly connected to the PCIe adapter, we get a bandwidth average value of $\sim$ 11.8 Gbit/s (green dots). Disabling the CPU power management we get an average value of $\sim$ 52.5 Gbit/s (red dots), corresponding to the $\sim$ 98% of the maximum rate. The conclusion we reach is that the event builder software is able to saturate the network capability when the proper settings of the system are applied. We also observed that the transmission rate is stable with time.

6. Scalability tests

We performed scalability tests of the event-builder on a large scale system, running the event-builder software evaluator at the 516 nodes Galileo cluster of the Cineca Consortium [6]. The size of the cluster is similar to the expected size of the event builder, however the Galileo cluster uses InfiniBand HCAs of the QDR type, which provide a bandwidth of about 27 Gbit/s. We first

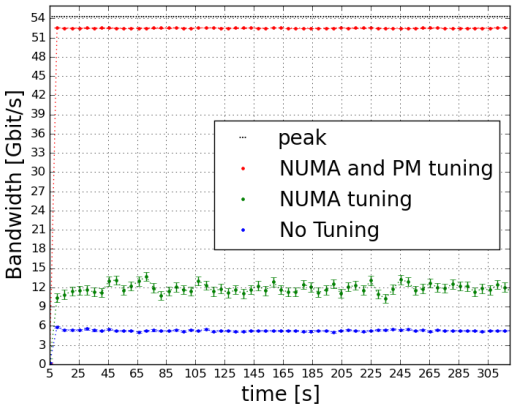


Figure 2. Event building point-to-point transmission rate measurements. The blue dots represent the transmission rate seen by the BU without any optimisation applied. The green dots represent the results we get when binding the application to the CPU directly connected to the PCIe interface. The red dots represents the results obtained when disabling also the power management.

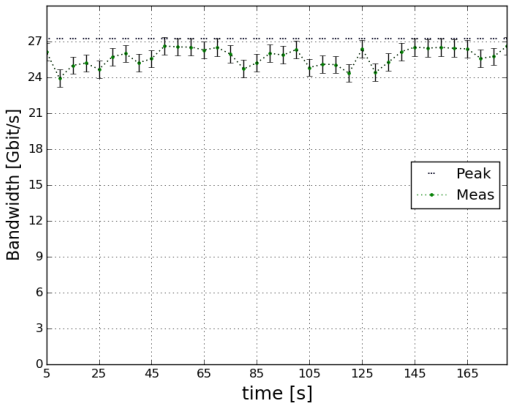


Figure 3. Characterization of the Intel QDR HCA of the Galileo cluster. The transmission rate has been measured using standard test tools.

performed reference measurements of the transmission rate achievable between any two nodes of the cluster selected randomly, using standard performance test tools (e.g. ib_write_bw). The results of the reference measurements are reported in Figure 3. The average transfer rate we get is 25.7 Gbit/s, corresponding to 94% of the expected maximum bandwidth (single duplex).

We run then the event builder software evaluator with different farm configurations, progressively increasing the number of nodes. Figure 4 shows the average transmission rate recorded by the BU versus the number of RUs (nodes used for testing). The results show the good scalability of the event-builder based on the OpenMPI implementation of RDMA. Figure 5 shows the results for a test performed using 128 nodes. As can be seen the data transfer is stable over time at $\sim$ 56% of maximum full-duplex bandwidth allowed for these HCA. The event builder the software showed good performance in term of scalability and stability in a cluster of high complexity.

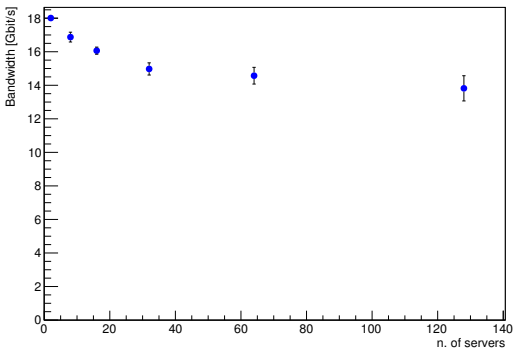


Figure 4. Scalability test results. The measured transfer rate as a function of the number of RU nodes.

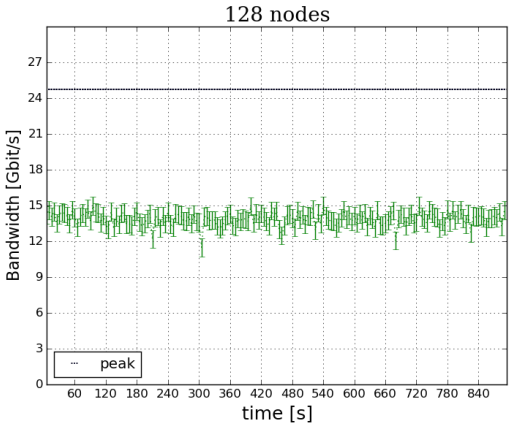


Figure 5. Data transfer rate for the BU for 128 nodes test of the event builder software.

7. Conclusion

The DAQ system for LHCb Upgrade has been redesigned in order to cope with the higher luminosity foreseen for the Run 3 of LHCb. Its implementation requires an high-throughput event builder network that must handle an aggregated traffic of the order of 32 Tbit/s. Such a network can be built using commercial hardware components such as InfiniBand. We developed a prototype software for testing purpose. The results of the measurements performed show that the event builder is stable and scalable up to 128 nodes. In order to reach good performance a proper tuning of the system is mandatory.

Acknowledgments

The authors thank the HPC User Support team at Cineca for their prompt support during the tests.

[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-2014-016. LHCB-TDR-016 (2014).
[3] LHCb Collaboration, The Upgrade of the LHCb Trigger System, JINST 9 (2014) C10026.
[4] Mellanox Technologies 2014, Performance Tuning Guidelines for Mellanox Network Adapters
[5] http://www.open-mpi.org
[6] http://www.hpc.cineca.it/content/galileo

WNoDeS: The error-correcting virtualization framework

V Ciaschini and S Dal Pra
INFN CNAF, Viale Berti Pichat 6/2, 40126, Bologna, Italy

Abstract. The CNAF Tier1 uses WNoDeS as a virtualization system for some of its worker nodes to allow jobs coming from experiments with unusual requirements like non-standard OSes, traffic shaping or special configurations, to be able to run and still have full integration into its batch system.

This report describes the main activities performed on WNoDeS in 2014, which were focused on increasing robustness and ease of administration.

1. Current State

In 2014 development activity on WNoDeS was split into two main lines:

- ease of administration
- fault tolerance and resilience

These activities concluded in December 2014 with the release of WNoDeS 3.0, which was put into production early January 2015, with one bug fix release since, and a total of one reported issue at the time of writing (October 2015), fixed by the bug fix.

1.1. Ease of Administration

This activity focused on improving ease of management for system administrators who run WNoDeS on their system. In practice this required an extensive rewrite of wnodes-manager, both to allow a more comfortable syntax, and produce output that could be parsed an analyzed with common unix tools like `grep` and `cut`. Sub-commands were also modified to accept wildcards when specifying machines, so the caller could specify something like `vwn-08*` to mean all virtual nodes whose hostname starts with `vwn-08` without having to call the same command `n` times, one for each VM.

An entirely new facility was also developed, the `history`. This facility registers status changes of all Virtual Nodes and all Hypervisor on a sliding window lasting 30 days by default. This feature was strongly requested by the system administrator because it allows, for example, to see on which hypervisors a virtual machine has run, and therefore simplifies diagnosing problems caused by interactions between hypervisors and virtual machines.

The final ease of administration change was a complete rewrite of the configuration system of WNoDeS. Before this rewrite, configuration was a mess, requiring for example to maintain exactly the same information, with slightly different syntax, and without documentation, present on multiple configuration files that had to remain consistent at all times. All parameters were required, even when there were reasonable defaults or their value could be inferred from the

value of other configuration parameters. After rewriting the system, only one configuration file exist, any option whose value could be derived from others is no longer needed; wherever a default value makes sense, the option assumes that value by default; finally all options have a multiline help text explaining exactly what it is for, and what the default value is.

1.2. Fault Tolerance and Resilience

This has been the main focus of the development effort for WNoDeS during 2014. The main problem is that WNoDeS relies on and tries to manage, a large number elements that it does not own nor fully control: jobs belong to, and are controlled by, the batch system, while virtual machines belong to the hypervisor. This could cause a large array of issues. For example LSF may decide to reschedule a job or kill it when it is being handled by WNoDeS, or the hypervisor could decide to stop a virtual machine, or to die itself but leave the VM in a semi-running state. A node could become closed due to the action taken by LSF or due to admin action, possibly by mistake. Various timeouts could be reached when a system is busy, and in addition to that, *normal* issues like hardware and network failures happen.

To have a more resilient system, all of these problems had to be accounted for. Unfortunately, they are by they happen asynchronously to WNoDeS internals, and trying to account for all of them in practically every line of code of the services is both impractical and fragile.

These problems were solved with the development of watchdog threads in both the global WNoDeS daemon (the nameserver) and the local daemon (the hypervisor) along with making the whole structure resilient to the death of any of its components.

The watchdogs are threads running once every 5 minutes. The hypervisor watchdog takes a snapshot of the real situation of the machine and compares it with what it expects to see, reconciling any differences. For example, if it finds a VM with no job running or scheduled to run on it, it destroys the VM, releasing its resources, and adjusts the hypervisor's internal state accordingly, assuming that the job was killed or rescheduled by LSF itself. The nameserver watchdog periodically does the same thing but with a global view of the infrastructure, querying different subsections of it (to avoid taking too much time, since the operation is synchronized) and comparing the answers with what it expects to see, taking also into account that any part may be down or temporarily isolated.

Additionally, any part of the infrastructure, including the nameserver, can crash or be turned off at any time without any lasting global damage. A hypervisor that restarts from being down automatically recreates its internal state on the bases of what virtual machines and jobs it sees running, and if it were the whole host that went down, it will correctly consider itself empty, and the nameserver would adjust the global vision of it next time it asks for a report.

The nameserver itself may go down at any time, and all state changes that would have been taken place during the downtime would be rediscovered and reapplied even though not necessarily immediately.

The net result of these changes is a much stabler and resilient system, which is running in production with no reported issues.

References

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

[2] V. Ciaschini and S. Dal Pra and G. Dalla Torre and E. Ronchieri and D. Salomoni, WNoDeS: a virtualization framework in continuing evolution, Annual Report 2104

Cloud@CNAF

D Michelotto, F Cappanini, E Fattibene, D Salomoni and P Veronesi
INFN CNAF, Viale Berti Pichat 6/2, 40126 Bologna, Italy
E-mail: diego.michelotto@cnaf.infn.it

Abstract.
Cloud@CNAF is a project aimed at deploying a Cloud Infrastructure, based on open source solutions to serve the different CNAF use cases. The project is the result of the collaboration of a transverse group of people from all CNAF departments: network, storage, farming, national services, distributed systems. The Cloud@CNAF IaaS (Infrastructure as a Service) is based on OpenStack, Havana version, a free and open-source cloud-computing software platform. The present infrastructure is used by many projects such as EEE, !CHAOS, Middleware Developers group. This paper presents the activity carried out at CNAF to set up the infrastructure, its deployment and the related training events devoted to the users. A perspective on the evolution of the infrastructure is also presented.

1. Introduction

The main goal of Cloud@CNAF project is to provide a Cloud Infrastructure for CNAF users and for projects taking place at CNAF:

- Provisioning VM for CNAF departments
  - Backend for provisioning of VMs to other services/processes like Jenkins, LBaaS
  - Provisioning of VM for the User Support service (VirtualBox-like)
- Provisioning of VM for CNAF staff members
- VM for experiments hosted at CNAF
  - DIRAC-OpenStack integration
  - access to VM for pilot jobs - CMS
  - Long Term Data Preservation
- Tutorials and Hands-on

The infrastructure made available is based on OpenStack [1], an open source product that can be deployed on open source platforms and has strong support from the industry. The version deployed in production is Havana [2].

2. Infrastructure

The present infrastructure is composed of:

- one Controller Node providing the core services - Keystone, Glance (LVM), Heat, Horizon, Ceilometer, MySQL, QPID, hosted on a physical machine with:
  - 2x8 HT (32) Intel(R) Xeon(R) CPU E5-2450 @ 2.10GHz, 64 GB RAM

- one Network Node providing Neutron configured to use 100 VLANs with Open vSwitch[3] virtual switch, hosted on a physical machine with:
  - 2x8 HT (32) Intel(R) Xeon(R) CPU E5-2450 @ 2.10GHz, 64 GB RAM
- the infrastructure is completed by four Compute nodes (for a total of 64 CPU and 256 GB of RAM) that provide Nova services based on KVM/QEMU hypervisors.

In this environment GPFS[4] is used as distributed file system for the backend infrastructure storage. The GPFS cluster is composed of three GPFS servers that expose four LUNs exported from a Dell PowerVault StorageSystem, providing 16TB of storage space for Nova backend.

A cloud User Interface node is also present to guarantee the access to the infrastructure using the powerful OpenStack APIs for an advanced use of OpenStack.

In early 2015 more than 50 users have been registered in the infrastructure, providing for them 50 tenants with 48 instances that use about 90 VCPUs and 180GB of RAM.

3. Use cases

The active use cases working at CNAF can be separated in two sets, the first represents the internal CNAF use cases, and the second collects the external projects where CNAF is involved in.

3.1. CNAF Internal use case

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

- The **Middleware Developers** team is composed by people active on the development of new software like EMI Middleware [5] for the Grid distributed computing resources, VOMS [6], StoRM [7] and Argus [8]. Middleware developers are involved in two kinds of projects:
  - *Middleware-project*: the instances of this project are coordinated by an external Jenkins [9] master. The instances are used for building the software and deploying it (Installation and Upgrade). Moreover, some instances based on CoreOS [10] are implementing Docker [11] nodes to speed up test deployment. (See Fig. 1).

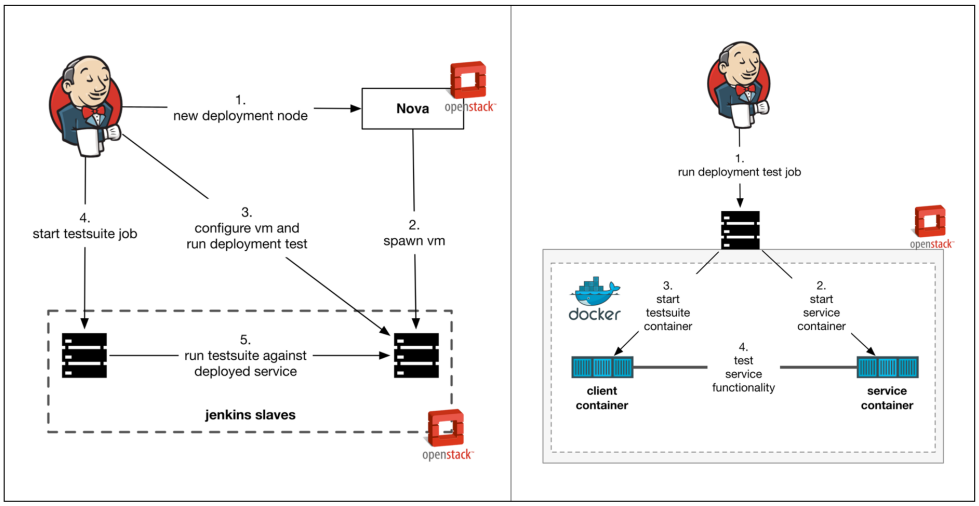


Figure 1. The Jenkins build node architecture (Left hand side) and the Jenkins docker deployment test architecture (Right hand side) are shown.

– *Storm Distributed Testbed*: this project is used for study and test the deployment in high availability for StoRM infrastructure. (See Fig. 2).

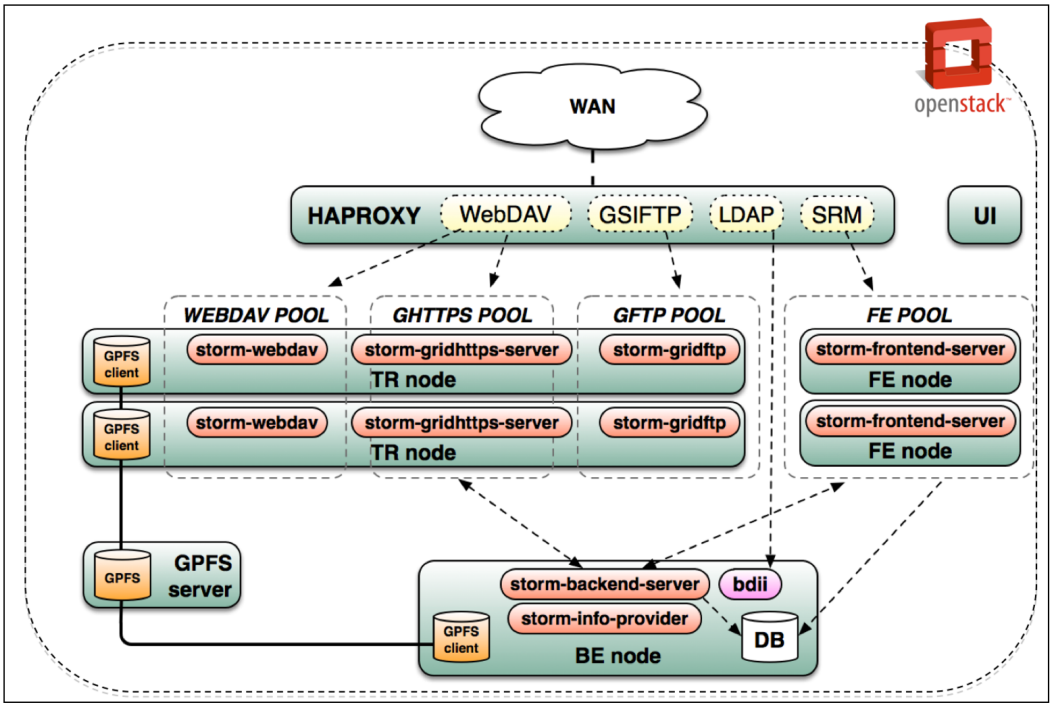


Figure 2. The StoRM distributed architecture.

- **CNAF staff.** Every employee at CNAF can rely on a personal project that can be used for internal needs. Each project has limited resources: 10 instances, 10 VCPU, 25GB RAM, 5 Floating IPs and 10 security groups. At the time of writing there are about 30 active users having personal projects.

3.2. External Projects

CNAF participates to several external project, in particular two of them make a massively use of the Cloud@CNAF infrastructure.

- The **Extreme Energy Event** (EEE) [12] experiment is devoted to the search of high energy cosmic rays through a network of telescopes installed in about fifty high-grade schools distributed throughout the Italian territory.
  The EEE activity at CNAF started in 2014 with the data collection during the EEE pilot run. The activity involved 21 schools (plus two INFN telescopes) in a coordinated data acquisition. During the pilot run all the schools were connected/authenticated at CNAF in order to transfer the amount of data acquired from the school's telescopes. To this scope, in our cloud infrastructure a node has been dedicated to play the role of frontend in order to receive all the data (with a total required bandwidth of 300 kB/s) to collect the expected 10 TB per year using BTSynk [13]. All the information collected by the experiment are considered in custody and for this reason the migration of the data to a tape system has been implemented.
- The **!CHAOS** [14] activity was originally developed within the context of High Energy Physics (HEP) as a candidate of Distributed Control Systems (DCS) and Data Acquisition

(DAQ) for the SuperB experiment. In 2014 it evolved into the project *!CHAOS: a cloud of controls* supported by MIUR and developed by INFN through its four sites: Laboratori Nazionali di Frascati (LNF), Laboratori Nazionali del Sud (LNS), Padova Section and CNAF Centre.
On top of OpenStack, the backend services are automatically provided as Platform as a Service (PaaS) components. The !CHAOS frontend services (CDS and MDS), instead, exploit the virtual instances of PaaS components and can run on single or multiple nodes. The frontend services communicate in a bidirectional way with the remote !CHAOS clients, such as the Control Unit (CU) and the User Interface (UI). Since the remote clients cannot be identified by a public IP address, a VPN service has been deployed within the Cloud@CNAF infrastructure to allow network traffic flow to and from the frontend services.

4. Training

In 2014 at CNAF four OpenStack training sessions have been organized by the Software Developer and Distributed System group. The first training course covered basic information about OpenStack, such as its architecture and components.

The other three were dedicated to hands on OpenStack from a user point of view. During these training, many arguments were treated, in particular was explained how to use the OpenStack dashboard and how to use its command line interface. More than 30 users participated to these training courses. In addition, a seminar on Heat component has been organized.

5. Future Work

In 2015 the Cloud@CNAF project will be improved in order to obtain a highly available infrastructure. Next steps foresee the deployment of a new infrastructure with redundant components (Controller and Network) and an increasing number of hypervisor resources.

For stability reason the Neutron services will be based on Linux Bridge instead of Open vSwitch. At the same time, the compute nodes will be increased from 4 to 13 nodes aiming to make available a total computing power of about 200 CPUs and 800 GB of RAM.

The GPFS filesystem will be used as backend storage for all OpenStacke services, Nova, Cinder and Glance. Data persistence and the messaging service, AMQP [15], will be guaranteed by a three nodes cluster hosting MySQL Percona [16] and RabbitMQ [17] software tools.

References

[1] OpenStack, *http://www.OpenStack.org/*
[2] OpenStack Havana, *https://www.OpenStack.org/software/havana/*
[3] Open vSwitch, *http://openvswitch.org/*
[4] GPFS aka Spectrum Scale, *http://www-03.ibm.com/systems/uk/storage/spectrum/scale/*
[5] EMI, *http://www.eu-emi.eu*
[6] VOMS, *http://italiangrid.github.io/voms/*
[7] StoRM, *http://italiangrid.github.io/storm/*
[8] Argus, *http://argus-authz.github.io/*
[9] Jenkins, *https://jenkins-ci.org/*
[10] CoreOS, *https://coreos.com/*
[11] Docker, *https://www.docker.com/*
[12] E Fattibene et al., CNAF activities in the !CHAOS project, this report
[13] BitTorrent Sync, *https://www.getsync.com/intl/it/*
[14] E Fattibene et al., The EEE Project activity at CNAF, this report
[15] AMQP, *https://www.amqp.org/*
[16] Percona, *https://www.percona.com/*
[17] RabbitMQ, *https://www.rabbitmq.com/*

Middleware support, maintenance and development

A Ceccanti, D Andreotti, E Vianello, G Dalla Torre and F Giacomini
INFN-CNAF, Bologna, Italy
E-mail: andrea.ceccanti@cnaf.infn.it

Abstract.
INFN-CNAF plays a major role in the support, maintenance and development activities of key middleware components (VOMS, StoRM, Argus PAP) widely used in the WLCG and EGI computing infrastructures. In this report, we discuss the main activities performed in 2014 by the CNAF middleware development team.

1. Introduction

The CNAF middleware development team has focused, in 2014, 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 [6]: the lightweight storage element in production at the CNAF Tier-1 and in several other WLCG sites;
- Argus Policy Administration Point (PAP) [5]: the Argus administrative interface and policy repository.

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 software development and maintenance process

The software development, maintenance and evolution activities for VOMS, StoRM and Argus follow the same process, and are driven by user requirements and support requests that identify problems or shortcomings in the code.

All Requests for changes (RFCs) are tracked in the INFN JIRA tracker [10, 11, 12, 13], prioritised, and linked to development sprints that are usually three weeks long. A development sprint leads typically (but not always) to a software release which is announced via the product website [4] and announcement mailing lists.

2.1. Continuous Integration and testing

All the code for VOMS, StoRM and the Argus PAP is hosted on Github [8]. The software is continuously built and tested by our Jenkins [7] server, which is configured with build nodes for the main supported platforms (Scientific Linux 5 and 6).

The Github Webhooks [18] mechanism provides efficient integration between the code repository and our CI server, so that whenever a change is pushed to one of the managed repositories a new build job is started on our CI infrastructure.

Continuous deployment tests for VOMS and StoRM run nightly, in order to check that the latest versions of our products install and run correctly on all supported platforms. The clean environment required for the deployment tests is provided by the integration with the Cloud@CNAF local private cloud infrastructure based on OpenStack [9] Havana.

3. VOMS

During 2014, 68 new issues were opened in the VOMS issue tracker [12] to track maintenance, development and release activities. In the same period, 74 issues were resolved.

The main highlights for VOMS are:

- The release of VOMS Admin server version 3.3.0 and 3.3.1 [20, 21], providing several improvements and bug fixes mainly targeted at the VOMS deployment at CERN which serves the main LHC experiments
- The release of VOMS server version 2.0.12 [22], to fix several issues found in production
- Minor fixes on the VOMS Java APIs [24] and clients [23]
- Evolution of the VOMS functional and regression testsuite [15]

4. StoRM

During 2014, 85 new issues were opened in the StoRM issue tracker [11] to track maintenance, development and release activities. In the same period, 86 issues were resolved.

The main highlights for StoRM are:

- The development of the StoRM WebDAV service, officially released after an initial testing phase in february 2015 [27], to replace the StoRM GridHTTPs providing a more scalable webdav solution for StoRM
- The StoRM 1.11.4 and 1.11.5 releases [25, 26], providing fixes for several issues found in production and during development
- Evolution of the StoRM functional and regression testsuite [16]
- Evolution of the load testsuite [17]

5. Argus

In 2014, work on the Argus PAP focused on the support activities. No new releases of the Argus PAP were issued.

6. Future work

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

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

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 VOMS website news sectionhttp://italiangrid.github.io/voms/news
[5] Argus authorization service website http://argus-authz.github.io
[6] StoRM website http://italiangrid.github.io/storm
[7] Jenkins https://jenkins-ci.org/
[8] GitHub https://github.com/
[9] Openstack http://www.openstack.org
[10] INFN issue tracker https://issues.infn.it
[11] StoRM on INFN JIRA https://issues.infn.it/jira/browse/STOR
[12] VOMS on INFN JIRA https://issues.infn.it/jira/browse/VOMS
[13] Argus on INFN JIRA https://issues.infn.it/jira/browse/ARGUS
[14] The European Middleware Initiative http://www.eu-emi.eu
[15] The VOMS clients testsuite https://github.com/italiangrid/voms-testsuite
[16] The StoRM testsuite https://github.com/italiangrid/storm-testsuite
[17] The StoRM load testsuite https://github.com/italiangrid/grinder-load-testsuite
[18] Github webhooks https://help.github.com/articles/about-webhooks
[19] Github pages https://pages.github.com
[20] VOMS Admin 3.3.0 http://italiangrid.github.io/voms/release-notes/voms-admin-server/3.3.0/
[21] VOMS Admin 3.3.1 http://italiangrid.github.io/voms/release-notes/voms-admin-server/3.3.1/
[22] VOMS 2.0.12 http://italiangrid.github.io/voms/release-notes/voms-server/2.0.12/
[23] VOMS clients 3.0.6 http://italiangrid.github.io/voms/release-notes/voms-clients/3.0.6/
[24] VOMS Java APIs 3.0.5 http://italiangrid.github.io/voms/release-notes/voms-api-java/3.0.5/
[25] StoRM v. 1.11.4 http://italiangrid.github.io/storm/2014/04/23/storm-v.1.11.4-released.html
[26] StoRM v. 1.11.5 http://italiangrid.github.io/storm/2015/01/07/storm-v.1.11.5-released.html
[27] StoRM WebDAV v. 1.0.2 http://italiangrid.github.io/storm/release-notes/storm-webdav/1.0.2/

A novel software quality model

M Canaparo, E Ronchieri and D Salomoni
INFN CNAF, Viale Berti Pichat 6/2, 40126, Bologna, Italy
E-mail: marco.canaparo@cnaf.infn.it, elisabetta.ronchieri@cnaf.infn.it, davide.salomoni@cnaf.infn.it

Abstract. Existing software quality models identify a set of software characteristics that are defined by standards, such as ISO-9126, and measured by metric tools. The characteristics require a set of metrics, typically shared among them, which are measured periodically during the software development life cycle. Software quality models and metrics determine an upside-down approach to quantify high reliability balancing effort and results. However, quality models often provide a partial view of the analyzed software, resulting unhelpful for developers during their daily activities.

In this report, we are going to describe our solution to fulfil the aforementioned issue. We designed mathematically and tested our own quality model prototype where we connected software best practices with code metrics. Furthermore, we supplied the model with statistical techniques, such as discriminant analysis and linear regression, to predict the quality at any stage of development. For the validation process, we used some EMI software products whose defects were already known. Our solution has proved to reasonably reproduce reality.

1. Current State

In the context of the software development in several European projects, such as DataGrid [1], EGEE [2], EGI [3] and EMI [4], researchers including those from INFN CNAF have analysed a set of quality models (such as McCall [1], Bohem [2], Dromey [3] and ISO [4]) according to the projects' requirements. These models have revealed to be specific for the domain they were designed (e.g., telecommunication and aerospace). Furthermore, they address a subset of software characteristics [4] resulting too complex to extrapolate the needed information. The proposed software quality model connects software best practices [5] with a set of metrics [6] to improve the prediction of the quality of software at any stage of development. While for the best practices we considered those to improve the success of software development process, for the metrics we derived them from best practices, static and dynamic analysis. The best practices [7] we selected refer to software structure, the construction of the code, deployment, testing and configuration management. The metrics we decided to include are specific of both best practices (such as those related to file and code conventions and software portability) and analysis [8] (such as Lines Of Code and Number of Defects). However, we planned the validation of our model with a progressive increase in the data set in terms of product metrics and software packages to properly speculate on the variables included in the model.

[1] The DataGrid project, http://eu-datagrid.web.cern.ch/eu-datagrid/
[2] EGEE, Enabling Grid for e-Science egee, http://eu-egee-org.web.cern.ch/eu-egee-org/index.html
[3] European Grid Infrastructure (EGEE), http://www.egi.eu/
[4] European Middleware Initiative (EMI), http://www.eu-emi.eu/

We fully described the model [9] by using the mathematical description formalism to express various levels of abstraction from the fundamental concepts of software engineering up to metrics. It leverages predictive techniques, called risk-threshold Discriminant Analysis (DA) [10] and linear regression [11], whose starting point is the measurement of the foregoing metrics, while its outcomes determine risky software products that may contain defects. The calculated risk-threshold [12] contributes to detecting fault-prone and non fault-prone components, whilst the predictive technique determines the metrics that influence the behaviour of the component the most. As result, DA confirmed the correctness given in [13], while the regression method determined an inaccurate number of defects. However, the outcomes can be improved by increasing the data set size and better contextualizing the predictive methods.

We used some EMI products under INFN responsibilities in the EMI distributions [14]. We selected source code mainly written in `Python`, `sh`, `Java`, `C` and `C++`. The analysis exploited up to 5,489 files in 52 software components amounting to a 1,570,323 total lines of code by using a Matlab-based prototype tool that codes the presented solution. The prototype classified all the components in faulty and non-faulty groups with a correctness of about 88%. We validated our model enlarging the data set by increasing the number of metrics and software products [9]. For the former we also considered complexity metrics, while for the latter we used CREAM (Computing Resource Execution And Management) [15], VOMS (Virtual Organization Management System) [16], WMS (Workload Management System) [17] and YAIM (Yet Another Installation Manager) [18], WNoDeS (Worker Nodes on Demand Services) [12] and StoRM (STOrage Resource Manager) [19] products released in the EMI 3 Monte Bianco distribution, to highlight similarities and differences among development scenarios.

2. Future Work

Starting from the current state, there are several improvements that we can conduct in the following periods. In the short-term, we are going to study the correlation among metrics and to express defects as function of various metrics: the former provides us with details about how they influence one another; the latter determines which metric has a greater weight than others. The statistical computing tool, called R, represents the best tool to fulfil these achievements. Our aim is to identify and improve the predictive technique that reproduces reality as much as possible strengthened by our knowledge of the problem. In the medium-term, we would like to increase the data set by adding new software products and metrics both static and dynamic ones. Our purpose is to improve the % of correctness in the prediction of our model. In the long-term, we will consider and adopt further predictive techniques, in addition to DA and regression, such as k-fold cross-validation [20] and support vector method [21], to strengthen the validation of our model. Furthermore, we will dedicate effort to evaluate the applicability of existing software metrics thresholds to be included in our model.

References

[1] McCall J, Richards P K and Walters G F 1977 Factors in software quality Tech. rep. Griffiths Air Force Base
[2] Bohem B 1978 Characteristics of software quality Tech. rep.
[3] Dromey G 1995 *IEEE Transactions on Software Engineering* **21** 146–162
[4] Iso/iec 25010:2011 - systems and software engineering – systems and software quality requirements and evaluation (square) – system and software quality models
[5] CMMI P T 2010 CMMI for Development, Version 1.3 Technical Report CMU/SEI-2010-TR-033 Software Engineering Institute URL `http://resources.sei.cmu.edu/library/asset-view.cfm?assetid=9661`
[6] Kan S H 2002 *Metrics and Models in Software Quality Engineering* (Addison-Wesley Professional)
[7] Perks M 2006 Best practices for software development projects Tech. rep. IBM
[8] Chidamber S R and Kemerer C F 1994 *IEEE Transactions on Software Engineering* **20** 476–493
[9] Ronchieri E, Canaparo M and Salomoni D 2014 *Journal of Integrated Design and Process Science* **18** IOS Press Amsterdam, The Netherlands, The Netherlands
[10] Guo G and Guo P 2008 *International Conference on Computational Intelligence and Security*

[11] Fenton N 1990 *Journal of Software Engineering* **5** 65–78
[12] Salomoni D, Italiano A and Ronchieri E 2011 *Journal of Physics: Conference Series (JPCS)* **331**
[13] Ronchieri E and Canaparo M 2013 *The 8th International Conference on Software Engineering and Applications (ICSOFT-EA 2013)*
[14] Aiftimiei C, Ceccanti A, Dongiovanni D, Meglio A D and Giacomini F 2012 *Journal of Physics: Conference Series (JPCS)* **396**
[15] Andreetto P, Bertocco S, Capannini F, Cecchi M, Dorigo A, Frizziero E, Gianelle A, Giacomini F, Mezzadri M, Monforte S, Prelz F, Molinari E, Rebatto D, Sgaravatto M and Zangrando L 2011 *Journal of Physics: Conference Series* **331**
[16] Ceccanti A, Ciaschini V, Dimou M, Garzoglio G, Levshina T, Traylen S and Venturi V 2009 *Journal of Physics: Conference Series* **219**
[17] Cecchi M, Capannini F, Dorigo A, Ghiselli A, Giacomini F, Maraschini A, Marzolla M, Monforte S, Pacini F, Petronzio L and Prelz F 2009 *Advanced in Grid and Pervasive Computing* (*Lecture Notes in Computer Science* vol 5529) (Springer Berlin Heidelberg) pp 256–268
[18] Jayalal M L, Rajeswari S and Murty S A V S 2009 Application of yaim tool in grid computing Tech. rep. Superintendents Advisory Committee on Enrollment and Transfers (SACET)
[19] Zappi R, Ronchieri E, Forti A and Ghiselli A 2011 *An Efficient Grid Data Access with StoRM* (Springer New York) chap VI Grid Middleware and Interoperability, pp 239–250 Data Driven e-Schience. Use Cases and Successful Applications of Distributed Computing Infrastructures (ISGC 2010)
[20] Rodriguez J, Perez A and Lozano J 2010 *IEEE Transactions on Pattern Analysis and Machine Intelligence* **32** IEEE Computer Society
[21] Lo J H 2010 *The 2nd International Conference on Computer Research and Development* (Kuala Lumpur, Malesia) pp 765–769

An assessment of software metrics tools

E Ronchieri and F Giacomini

E-mail: elisabetta.ronchieri@cnaf.infn.it

Abstract.

Software metrics are a special kind of analysis, namely one in which the correspondent measures express source code characteristics and establish improvement priorities. Such metrics allow judging the quality of software and highlighting the lower-accuracy portions of code. Existing tools are able to collect meaningful measurements for trend quality over time. However, they interpret and implement the definitions of software metrics differently, causing a considerable unequal interpretation of the assessment of a software system from tool to tool.

In this report, we provide details of a number of software metric tools – both commercial and free – that we considered while assessing the software quality of two Geant4 sub-packages, called `geometry` and `processes`, over various releases. Measurements showed that the results are tool-dependent.

1. Introduction

Geant4 [1, 2] is a simulation system that is used in a wide variety of scientific context, such as shielding and radiation protection, and developed and maintained by an international widespread collaboration. It is a mature system (20 years old) that can be used as a playground to study metrics and metrics tools: in particular, can metrics help addressing its maintenability in the next 20 years? To answer this question, we have started to perform the quantitative assessment of a subset of Geant4 packages, which play a key role in scientific applications and are representative of different software development processes.

Geometry makes it possible to describe a geometrical structure and propagate particles through it. In turn, it includes a set of sub-packages such as biasing, divisions, magnetic field, management, navigation, solids and volumes. Any simulation application involves some geometrical modelling of the experimental configuration.

Processes handles particle interactions: electromagnetic interactions of leptons, photons, hadrons and ions, and hadronic interactions and transportation. Like the geometry package, it comprehends a set of sub-packages such as biasing, cuts, decay, electromagnetic, hadronic, management, optical, parameterization, scoring and transportation. Electromagnetic physics represents the core of particle transport, as almost any simulation scenario involves electromagnetic interactions either of primary or secondary particles.

In this report, we present some software metrics tools – both commercial and free – that mainly support the C++ programming language, the one in which Geant4 is written. Furthermore, they calculate software product metrics, that provide the characteristics of software over time in terms of various categories, such as size, complexity, object-orientation and quality.

2. Initial Assessment

An initial assessment of these tools is reported in [3], where we included the following tools CCCC (C and C++ Code Counter) [4], CLOC (Count Lines of Code) [5], Metrics [6], Pmccabe [7], SLOCCount (Source Lines of Code Count) [8] and Understand [9].

CCCC (C and C++ Code Counter) v. 3.1.4 analyses and reports measurements of source code primarily in C++. It processes the files listed on its command line as described below, generating a report in HTML format on various measurement of the code processed:

```
$ cccc --outdir=dname --lang=c++ --lang=c file1 ...
```

The main file is `cccc.htm` with detailed reports on each module. The report contains a number of tables that cover: a rejected extents presenting a list of code regions which the analyser was unable to parse; a structural summary including relationships to each module identified; a procedural summary presenting values of procedural measures summed for each module identified in the code submitted; a project summary report.

Cloc (Count Lines Of Code) v. 1.60 counts blank lines, comment lines and physical lines of source code. It processes the files listed on its command line as described below, generating a report on size measurement of the code processed:

```
$ cloc file1 ... --by-file-by-lang report-file=fname.txt
$ cloc file1 ... --skip-uniqueness --by-file-by-lang report-file=fname.txt
```

The metrics software package was born in 2010. It counts lines of code and McCabe metrics [10]. Metrics v. 0.2.6 processes the files listed on its command line as described below, generating a report in a csv file:

```
$ metrics format=csv file1
```

Pmccabe v. 2.6 counts non-commented lines and McCabe complexity for C and C++. It processes the files listed on its command line as described below, showing data on the stdandard output:

```
$ pmccabe -t -b file1 ...
$ pmccabe -t -F -C file1 ...
$ pmccabe -t -f -C file1 ...
```

SLOCCount v. 2.26 identifies and measures several languages. It counts the comment lines and counts lines automatically. SLOCCount processes the files listed on its command line as described below, showing data in the stdoutput:

```
$ sloccount datadir dname file1 ...
```

Understand v. 3.1.728 is a static analysis tool for maintaining, measuring and analysing code, supporting the main programming language, which is under an evaluation license. It processes the files listed on its command line `und`, generating a report in a csv file.

We performed measures on a standard PC with the Ubuntu operating system version 13.10. We considered various Geant4 release distributions in the range [6.1, 10.0] downloaded in a designated directory. Furthermore, we installed all the metrics tools, such as CCCC, CLOC, Metrics, Pmccabe, SLOCCount and Understand, and used their command-line tool. Referring to the measures, we also implemented a set of scripts: to automatize measurements on the same source tree of the geometry and processes packages in all the Geant4 distributions; to generate intermediate files containing raw measured data; to remove unnecessary information and filter them. We used the Microsoft Excel 2010 and RStudio version 0.98.1091 tools to perform some graphics.

The software metrics tools used in this study return the same value for the Number of Files (NFiles), whereas they provide similar values for Lines Of Code (LOC), Comment LOC (CLOC), Blank LOC (BLOC). The Cloc tool supports all the aforementioned metrics, while the SLOCCount tool only provides information for the LOC metric. Furthermore, the Cloc and Pmccabe tools provide the same Total LOC due to the use of the same code to measure the LOC, CLOC and BLOC metrics. These tools provide metric values that are often unequal, mainly due to the way the lines with comments are counted, but also to the lines removed from the tool parser, and the way curly brackets are counted. CCCC and Understand are able to provide measurements about object-orientation category in addition to the others.

3. Future Work

The next steps of this activity consist of: doing statistical analysis for inference; evaluating the applicability of existing quality thresholds; adding further object-oriented metrics that contribute to evaluate maintainability software factor; determining which metrics are most effective at identifying risks; correlating interactions between software quality and functional quality. At the end of this study, we intend to create a baseline for evaluating correlations between software quality embedded in the Geant4 development process and simulation observables produced by Geant4-based applications.

References

[1] Agostinelli S, Allison J, Amako K and et al 2003 *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **506** 250 – 303
[2] Allison J, Amako K and et al 2006 *IEEE Transactions on Nuclear Science* **53** 270 – 278
[3] Ronchieri E, Pia M G and Giacomini F 2014 *ANS RPSD*
[4] CCCC URL `http://cccc.sourceforge.net/`
[5] CLOC URL `http://cloc.sourceforge.net/`
[6] Metrics URL `https://pypi.python.org/pypi/metrics`
[7] Pmccabe URL `http://people.debian.org/`
[8] SLOCCount URL `www.dwheller.com/sloccount/`
[9] Understand source code analysis and metrics URL `www.scitools.com`
[10] McCabe T J 1976 *IEEE Transactions on Software Engineering* **SE-2**

Provisioning IaaS for the Open City Platform project

C Aiftimiei[1,2], M Antonacci[3], G Donvito[3], E Fattibene[1], A Italiano[3], D Michelotto[1], D Salomoni[1], S Traldi[4], C Vistoli[1] and G Zizzi[1]

[1]INFN CNAF, Viale Berti Pichat 6/2, 40126 Bologna, Italy
[2]IFIN - "Horia Hulubei", Bucharest - Magurele, Romania
[3]INFN Bari, Via E. Orabona n. 4, 70125 Bari, Italy
[4]INFN Padova, Via Marzolo 8, 35131 Padova, Italy

E-mail: cristina.aiftimiei@cnaf.infn.it

Abstract.
The Open City Platform (OCP) project [1] aims to research, develop and test new technological solutions open, interoperable and that can be used on-demand in the context of cloud computing, along with new organizational models, sustainable in time, for the public administration, to innovate, with scientific results, with new standards and technological solutions, the provision of services by local public administrations (PAL) and regional citizens, companies and other government departments. OCP inherits the experience of other projects of Cloud Computing applied to the Public Administration and Research like PRISMA[2], Marche Cloud[3], INFN Cloud. The IaaS layer of OCP is based on OpenStack [4], the open source cloud solution most widespread in the world to manage physical resources, virtual machines and containers. In this paper we will present the reaserch activity done at CNAF in order to indentify the best tools for an automatic provisioning as the IaaS layer of the OCP platform.

1. Introduction

As is well-known, Cloud computing is a model for procurement and use of distributed storage and computing resources with a strong development and adoption by the industry, academia and government. For these players, however, the adoption of the models offered by cloud, although it represents a strong opportunity, is often hampered by different problems: technical and scientific issues, rules and regulations that impose specific behavior and protection for themselves and for the recipient of public services (citizen or company).

Therefore the main areas of research conducted by the OCP can be traced to:

- scientific and technological challenges like:
  - Federated management of heterogeneous cloud platforms
  - Integrated monitoring and support to billing systems
  - Design and reengineering Cloud applications
  - Disaster recovery as a Service
  - Integration of PaaS components in particular Paas for eGov
  - Open data and the Open Service and integration into business models
  - Federated identity management and its trust relationship
- legal, organizational, functional challenges and new business models that are necessary to ensure a concrete feasibility within government scope of the results achieved by the project, like:

- Define new organizational models and public governance where regions have the role of infrastructure intermediaries
- Decouple the exercise of administrative functions (public role) from the ICT instrument (private role) on which the function is performed
- Adherence to new models regulations
- Accountability and attribution of precise responsibilities ensuring mutual protection among those involved in the chain of service

The OCP architecture has been designed as a scalable, multilayer and interoperable platform that inherits also components from previous projects, like PRISMA. The main components are shown in Figure 1:

- orchestrator module
- monitoring/billing module
- PaaS layer/services
- IaaS layer/services
- Open Data e Open Service Engine
- Service/citizen market place

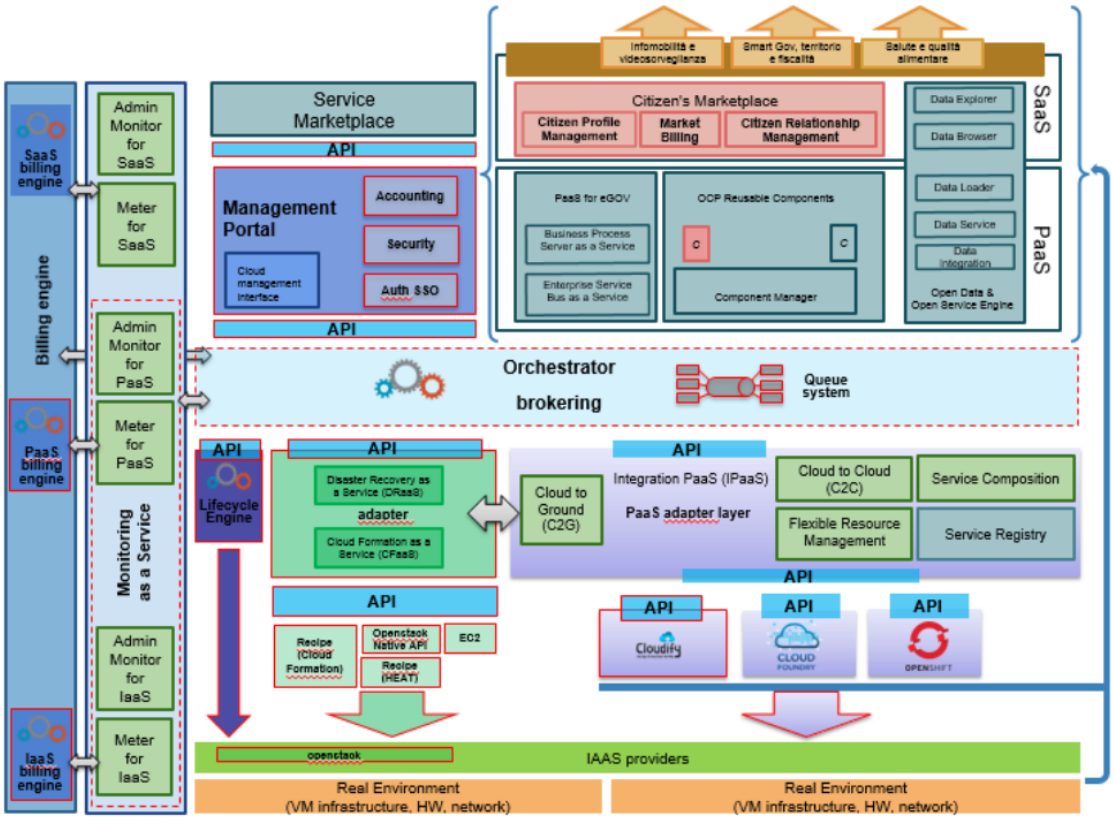


Figure 1. OCP platform architecture

The IaaS layer of OCP is based on OpenStack, the most widely used open-source cloud solution that allows to manage physical resources, virtual machines and containers.

2. IaaS and installation scenarios

Open City Platform makes available a Cloud Computing platform open source, flexible, scalable and compliant to international standards that consists of components organized in different layers fully integrated with each other that can already be installed and used for activation of on demand services and access data in the Data Centers of the PA. The platform will be continuously updated with additional OCP components developed by the project in order to progressively meet the various specific needs of PA. The layers that constitute the cloud platform OCP consists of:

- LAYER 1: A Infrastructure as a Service (IaaS) platform based on OpenStack, suitably configured by capitalizing on the pioneering experiences made in the INFN Cloud-infrastructure, Marche Cloud project and especially in the PRISMA project.
- LAYER 2: A Platform as a Service (PaaS) platform that allows to use heterogeneous IaaS sites and easily manage the activation and execution of new applications
- LAYER 3: A Software as a Service (SaaS) platform that consists of an Application Store and a set of new services that allow PA to choose the application of interest, to configure it to suit their needs and activate it digitally native on cloud-infrastructure, also different from that of OCP.

In the first phase OCP will make available the components related to the LAYER 1 and LAYER 2. OCP is evaluating three different installation methods for the IaaS layer. The selection of the installation method that will be adopted by the regions involved in the pilot study will depend on factors such as:

- degree of familiarity with OpenStack
- interest in deepening configuration knowledge
- requirement of special configurations, etc.

The proposed methods are:

- manual installation and configuration ('hardcore').
- installation with data center automation tools like Puppet [5] and/or Foreman [6].
- semi-automatic installation (GUI) through the Fuel [7] tool.

2.1. Manual Installation and Configuration

This method consists of:

- The installation of the Openstack middleware based on the use of the package manager (apt-get, yum, etc.) and the repositories available for the chosen distribution of the Operating System
- Changing the configuration of the individual services by editing the corresponding configuration files.
- Verifying the proper installation and configuration by checking the state of the processes involved, the status of the connections between related services, the messages collected in the logs, etc.
- In the case of deploying multi-node HA, services are first installed in standalone mode and then clustered.

Some of the advantages of this method are:

- better understanding of OpenStack dependencies between components, of the possible choices in the deployment of services

- more control over configurations, the possibility to enable advanced functionality, support multi-region, SSL support of the identity service, etc.

Yet, the following disadvantages have been identified:

- this method requires requires basic knowledge of Linux OS - bash, network configuration
- error-prone - for ex. typo in the configuration change
- time-consuming - many of the operations are repetitive
- it requires additional effort to keep aligned configurations on more servers (eg in the case of multi-node deployment HA).

This method is widely used in INFN cloud projects both for production and for test and pre-production environments. Based on our experience we can say that the manual installation method is to be considered preferable as a first approach to the installation of OpenStack to get familiar with all the parts, when there is a need to implement special advanced configurations. The greater flexibility and control in the choice of configuration are the main strengths of this method.

2.2. Installation with automation tools

Puppet is a tool that has the purpose to automate procedures within a data center. It provides modules for installation and automatic configuration of OpenStack.

After defining the configuration, Hiera, an add-on Puppet component that allows hierarchical configurations, the agent automatically boots the process of installation and configuration. At the end of the process it will produce a report. Errors can only be due to incorrect (or lack of) configuration.

The OpenStack Puppet modules [8] follow the normal development of OpenStack with a major new release for each version of OpenStack. Within each cycle version of the modules they are released 1/2 updates.

Pros and cons:

- Puppet is almost a programming language that needs to be learned; its "learning curve" depends on ones experience, which will also serve to handle errors.
- This method architecturally has no limits on the definition of configurations.

Work progress:

- at this moment all services are covered by OpenStack Puppet modules available.
- In the future we plan to add the ability to support more complex use cases such as CEPH on the hypervisors.

It is possible to manage the installation of OpenStack through Puppet also via a GUI through the use of the Foreman tool.

Pros:

- Use of a unique web interface to provision Linux hosts (Redhat, CentOS, Ubuntu, Debian, Suse) and manage both puppet modules and OpenStack components;
- Easy installation of components using OpenStack QuickStack modules - there is no need to have a deep knowledge of OpenStack to install and configure it;

Cons:

- The installation procedure is in continue evolution (as OpenStack) - it may therefore be necessary to apply minor changes and patches in order to have working modules.

- If the OpenStack infrastructure is very complex it should be assessed whether it is possible to finely parameterize via Foreman OpenStack components (keystone, nova, glance, cinder, etc.);
- Error messages are not easy to understand.

2.3. Semi-automatic installation using Fuel

Fuel is a tool for the installation and management of an OpenStack-based infrastructure. It allows to perform the installation and configuration of one or more OpenStack environments through a web-GUI (see Fig. 2), and it gives also the possibility of using the CLI (command line client). When we performed our tests it was available for Icehouse [9] on Centos 6.5 and Icehouse on Ubuntu 12.04.4.

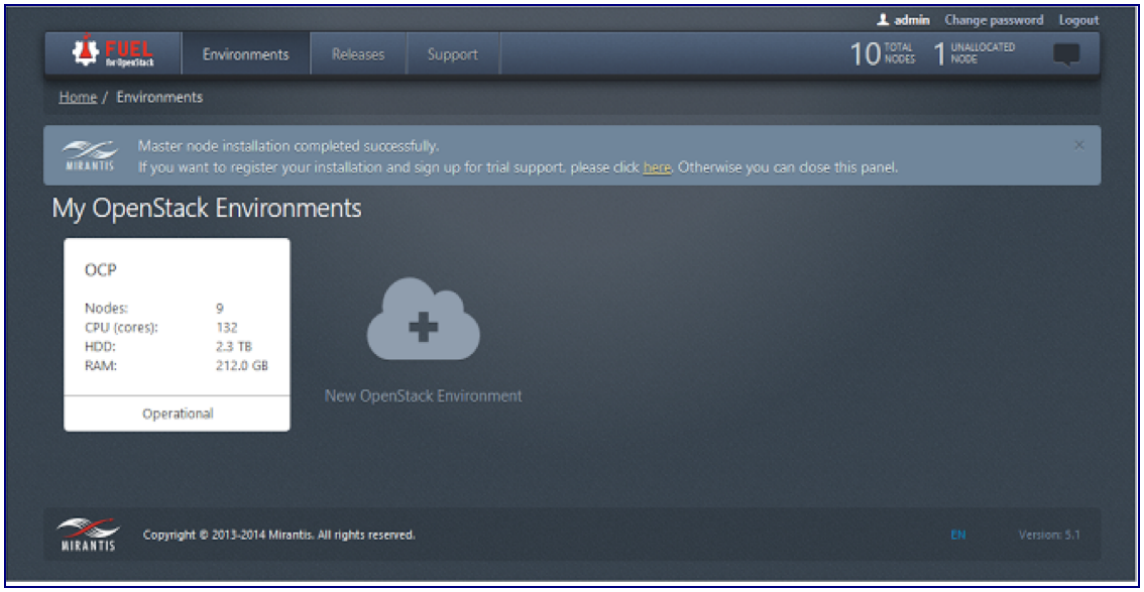


Figure 2. Fuel GUI

Functionality offered:

- Fuel discovers automatically each host (physical or virtual) configured to boot from network and present in its VLAN. Afterwards different roles can be assigned to each node (controller, storage etc.).
- There is a wizard for an easy installation
- Logs can be consulted via GUI, providing also control of the verbosity (see Fig. 3).
- It offers the possibility to perform tests on the health of the services and the correctness of network configuration.

The environment/infrastructure we deployed for testing purposes consisted of:

- Icehouse on CentOS 6.5
- Multi-node HA
- Neutron with GRE
- Nodes: 1 Fuel Master, 3 Cloud Controllers, 2 Compute nodes, 1 Cinder server, 2 Ceph servers and 1 Zabbix server (for monitoring)

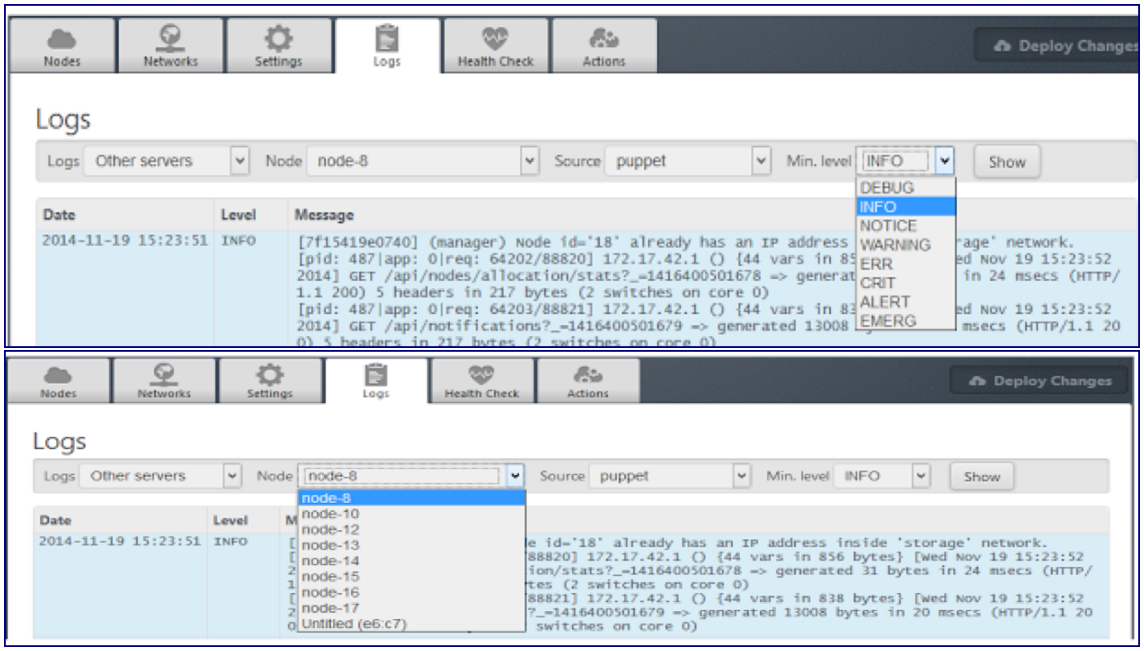


Figure 3. Fuel - Installation and Configuration Logs display

Some of the advantages and disadvantages that we identified are:

- Pros:
  - Easy installation through the GUI
  - It enables subsequent changes and new deployments (to change the additional services and add or remove nodes)
  - There is no need of custom adaptations.
- Cons:
  - The initial configuration cannot be changed (eg. move from a simple to a high-availability setup)
  - The OpenStack regions are not yet supported.
  - The various components of the Cloud Controller (eg. Keystone, Glance, Horizon, Neutron) are all installed on the same machine.
  - No installation options for an "all in one" setup (with only one server).

3. Future Work

One of the layers of the OCP platform is the IaaS, the base on which all the other layers build up. The chosen middleware is OpenStack, the Juno [10] version, and in order to ease its deployment in the testbeds available in the experimenter regions an activity of testing various deployment methods was carried out. The results were presented and based on them it was decided that in the first instance a manual instalation will be performed in some of the testbeds whereas future work will be invested in the preparation of an automatic tool based on Puppet and Foreman to be used in the remaining testbeds and new ones that will agree in using the results of the OCP project. This tool will also be used for performing the updates of the infrastructures to the new versions of Openstack.

References

[1] OpenCityPlatform Project, *http://www.opencityplatform.eu/*
[2] PRISMA,*http://www.ponsmartcities-prisma.it/*
[3] MCloud,*http://www.ecommunity.marche.it/AgendaDigitale/MCLoud/Obiettivi/tabid/206/Default.aspx*
[4] OpenStack, *http://www.OpenStack.org/*
[5] Puppet, *https://puppetlabs.com/*
[6] Foreman, *http://theforeman.org/*
[7] Fuel, *https://www.mirantis.com/products/mirantis-openstack-software/*
[8] Openstack Puppet, *https://wiki.openstack.org/wiki/Puppet*
[9] OpenStack Icehouse, *https://www.OpenStack.org/software/icehouse/*
[10] OpenStack Juno, *https://www.OpenStack.org/software/juno/*

Porting the Filtered Back-projection algorithm on low-power Systems-On-Chip

Elena Corni[1,2], Lucia Morganti[3], Maria Pia Morigi[1,2], Rosa Brancaccio[1,2], Eva Peccenini[2,4], Matteo Bettuzzi[1,2], Daniele Cesini[3], Giuseppe Levi[1,2] and Andrea Ferraro[3]

[1]Department of Physics and Astronomy, University of Bologna, Italy
[2]INFN (National Institute of Nuclear Physics), Section of Bologna, Italy
[3]INFN-CNAF, Bologna, Italy
[4]Enrico Fermi Center for Study and Research, Rome, Italy

E-mail: lucia.morganti@cnaf.infn.it

Abstract.

Among its various applications, X-ray Computed Tomography can be profitably applied in the field of Cultural Heritage to reconstruct the internal structure of art objects in a non-invasive way. In a collaboration between the X-ray Imaging Group of the Physics and Astronomy Department at the University of Bologna and INFN-CNAF, we investigated the possibility of porting the X-ray Computed Tomography reconstruction algorithm to new low power architectures typical of the embedded and mobile market, the so-called Systems-on-Chip (SoCs).

In particular, we exploited the Graphics Processing Unit (GPU) of the NVIDIA Tegra K1 SoC, and maximized the simultaneous use of CPU and GPU by combining a multi-threaded OpenMP version and a CUDA version of the reconstruction algorithm. Our main finding is that only three Tegra K1 boards, equipped with Giga ethernet interconnections, allow to reconstruct as many 2D slices (of a 3D volume) per unit time as a traditional high-performance computing node, using one order of magnitude less energy. These results seem very promising in view of the construction of an energy-efficient computing system of a mobile tomographic apparatus.

1. Introduction

We are embarking upon a new era in which scientific workloads that were traditionally confined to High Performance Computing (HPC) systems are starting to be ported to low power embedded architectures in order to improve energy-efficiency and power consumption. The interest of the scientific community for these low power, low cost Systems on Chip (hereafter SoCs) is mainly triggered by their ever-increasing computing performances.

In this paper, we focus on the porting of a real scientific application, and specifically the Filtered Back-projection algorithm from X-ray Computed Tomography, to low power SoCs.

The X-ray Computed Tomography analysis of large art objects of Cultural Heritage, carried on for both scientific investigations and restoration purposes, is typically time-consuming and power-consuming. Moreover, in most situations it is simply not possible to execute the reconstruction software directly where and when the X-ray measurements are acquired. Hence, a natural interest arises in exploring the possibility of running the reconstruction algorithm on a mobile, possibly battery-powered, device.

The chosen application for the present study is the C and MPI Filtered Back-projection algorithm for Computed Tomography reconstruction developed by the X-ray Imaging Group of the Physics and Astronomy Department at the University of Bologna.

The chosen SoC-based platform for the development, porting and testing of the application is the NVIDIA Tegra K1 SoC, which is available in a cluster of development boards located at INFN-CNAF.

All the results obtained with the low power architecture are compared with those obtained on a typical x86 HPC node accelerated with a recent NVIDIA GPU, belonging to a high-performance cluster which is also located at INFN-CNAF.

The paper is organized as follows.

In Section 2 we provide details on the experimental setup, i.e. the computing architectures, the apparatus used to measure the power consumption, and the Dataset adopted in the tests. In Section 3, after a brief explanation of the Filtered Back-projection algorithm, we describe the porting to OpenMP and CUDA that were performed for the present work. In the following Section 4 we evaluate the algorithm from the points of view of performances and energy consumption for both low power and traditional architectures, and propose a low power, portable solution for X-ray Tomography applied to art objects. Finally, we draw our conclusions in Section 5.

2. Experimental setup

2.1. Target architectures

For comparative purposes, all the tests presented in this work are performed using two different architectures.

The first one (hereafter Xeon) represents a typical node in a HPC cluster. It is equipped with two Intel Xeon E5-2620 CPUs, 6 physical cores each, HyperThread enabled (i.e. a total of 24HT cores in the single node), and with a Tesla K20 GPU accelerator from NVIDIA.

The second one represents a low power alternative: a cluster of NVIDIA Jetson Tegra K1 development boards, each one (hereafter TK1) equipped with a quad-core, 32-bit, Arm Cortex A15 processor, and a K1 GPU from NVIDIA.

The following Table summarizes the main features of the two architectures.

	CPU	Cores	RAM	Frequency	GPU	Cores	RAM	Frequency
Xeon	Intel Xeon E5-2620	24	48	2500	K20	2496	5	706
TK1	ARM Cortex-A15	4	2	2300	K1	192	2	852

Table 1. CPU (left) and GPU (right) specifications of the target architectures. RAM is given in GB; for the TK1 SoC, it is shared between CPU and GPU. Frequency is given in MHz.

2.2. Measuring power consumption

The experimental apparatus used to determine the electric power consumption of the considered architectures consists of a Tektronix DMM4050 digital multimeter for DC current measurements of TK1, and a Voltech PM300 Power Analyzer for AC power measurement of Xeon.

We stress that for the Xeon node we measured the power consumption upstream of the main server power supply (measuring on the AC cable), whereas for the TK1 board we measured the DC current absorbed. However, this difference in the measurements should not impact significantly on the obtained results, given the close to one cos phi factor of the server power supply.

2.3. Dataset

For the tests presented in this work, we used a real Dataset from the high-resolution tomographic analysis of the Kongo Rikishi, a Japanese wooden statue dating back to the XIII century [1, 3, 4].

We obtained three representative sets of images of increasing computational complexity, named Kongo_2048, Kongo_1024, and Kongo_256 hereafter, whose sides have a size of 2048, 1024 and 256 pixels, respectively. The number of angles at which radiographs have been acquired for each set is 720; only for the 256-side slice set, it is 360.

3. The Filtered Back-projection algorithm for Tomographic reconstruction

The process of Tomographic reconstruction (see e.g. [2]) starts with acquired radiographs (projections) of the art object at many different angles and for many different heights of the detector, and ends with a series of slices, i.e. 2D "reconstructed" images which correspond to internal sections of the investigated art object at the different heights. The full 3D volume of the object can then be reconstructed by superimposing a set of reconstructed slices.

For a given height on the detector, all the information needed to reconstruct a slice of the investigated object at fixed height is contained in the so-called *sinogram*, a 2D image in which each row contains the 1D attenuated projection of the object on the detector at fixed height and at many different angles.

Roughly speaking, the main steps of the Filtered Back-projection reconstruction algorithm are [2]:

(i) perform the 1D Fast Fourier Transform of each row of the sinogram in the frequency domain;

(ii) filter each row;

(iii) perform the 2D Fourier anti-transform;

(iv) "geometrically" back-project the anti-transformed space into the normal space.

Over the recent years, the X-ray Imaging Group of the Physics and Astronomy Department at the University of Bologna developed a C application performing all the above steps in order to reconstruct slices of large art objects from X-ray radiographs [1]. Since the amount of acquired data is typically very big, and the relative processing time consuming, a MPI version of the code has been used in HPC clusters [1, 3].

In order to run the Filtered Back-projection reconstruction algorithm in the target architectures described above, we developed both a multi-threaded OpenMP version and a CUDA (Compute Unified Device Architecture, from NVIDIA) version of the algorithm, as detailed below.

3.1. A multi-threaded OpenMP version of the Filtered Back-projection algorithm

For the porting of the Filtered Back-projection algorithm to OpenMP, we marked with specific preprocessor directives the following sections of the code to be executed in parallel by different *threads*:

- the zero padding of the slice
- the convolution
- the geometric reconstruction of the slice.

The latest section, which is the most time-consuming (around 99% of the total execution time for the original code), is the one in which back-projected values are assigned to each pixel of the slice. In the original C code, this step resulted in two nested loops on the x and y coordinates of only those pixels which lay inside the so-called *reconstruction circle*, i.e. a circle (inscribed in the slice, assuming the center of rotation placed at the center of the slice) where the reconstruction

process is defined. In order to parallelize the two original nested loops on the x and y directions of the slice, we used the OpenMP *collapse* clause.

Figure 1 shows the runtime, speedup and efficiency of the algorithm as a function of the number of threads for the Xeon architecture (hyperthreading enabled) and for a slice in the Kongo_256 Dataset.

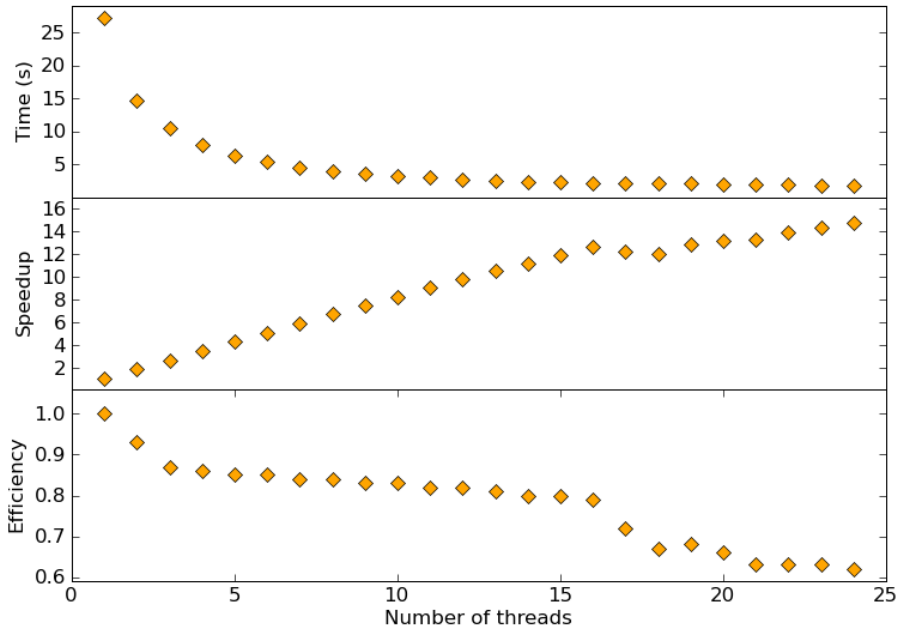


Figure 1. From top to bottom: runtime, speedup and efficiency of the OpenMP version of the Filtered Back-projection algorithm for increasing number of threads in the Xeon.

3.2. A CUDA version of the Filtered Back-projection algorithm

For the porting of the Filtered Back-projection algorithm to CUDA, we defined the following three *kernels*, i.e. functions called by the CPU (*host*) and executed on the GPU (*device*) by many *threads* in parallel:

- the zero padding of the slice
- the Fast Fourier Transform (FFT) and convolution
- the geometric reconstruction of the slice.

For the typical sizes of the images in our Dataset, offloading the zero-padding to the GPU is always timesaving.

With respect to the kernel devoted to FFT and convolution, instead, we compared complex-complex single precision FFT speeds for NR-f from Numerical Recipes [5] (the one adopted in the original code), FFTW and the CUDA CUFFT library for arrays of increasing size, as shown in Figure 2, and found that the CUDA CUFFT library becomes more efficient only when the size of the slice exceeds 4096 pixel, i.e. arrays are bigger than 8192 (2^{13}). Such big sizes are not present in our Dataset, and so the FFT-kernel is actually never called by the CPU.

Finally, we wrote the geometric reconstruction-kernel so that every thread processes one pixel of the slice, and we restricted the computation to those pixels belonging to the square in which

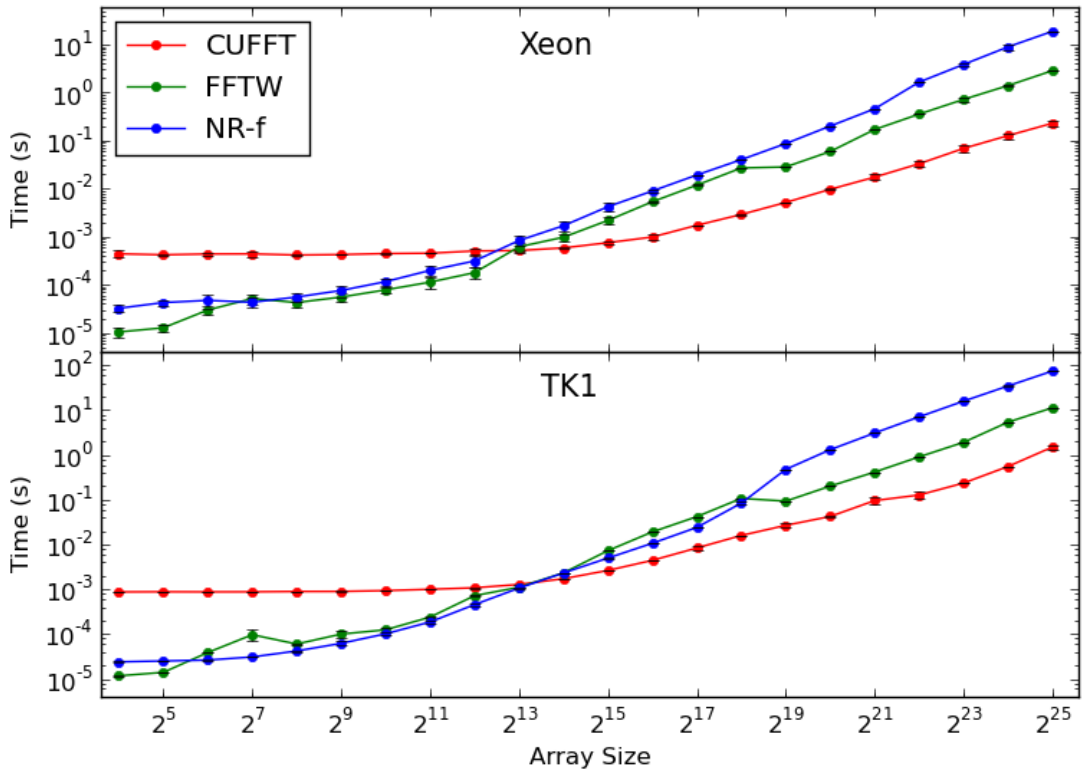


Figure 2. Complex-complex single precision FFT execution time for NR-f, FFTW and CUFFT library for increasing array size (see similar findings by, e.g., [6], [7]).

the reconstruction circle is inscribed. As naturally expected for such image-processing problem, this part of the code is remarkably accelerated by the CUDA implementation. Of course, the speed of the reconstruction process depends on the adopted number of threads and blocks[1]. After some experiments with increasing values of *dimBlock*, defined as the size of each block of threads in the grid, we found that a minimum value in the execution time was reached when *dimBlock* equals 16 for both Xeon and TK1 architectures. Hence, this value of *dimBlock* is adopted as optimal in what follows.

For reference, the execution time is 0.5s for the reconstruction of one image of the Dataset Kongo_256 (smallest) and 7s for one image of the Dataset Kongo_2048 (largest) on the Xeon, while the corresponding values on the TK1 SoC are 0.9s and 24s.

4. Results

A natural way to evaluate the speed of the process of Computed Tomography reconstruction is the number of 2D slices reconstructed per time unit.

Nonetheless, in this study we do not restrict ourselves to time-to-solution, and we also wish to consider energy-to-solution metrics.

Hence, in Figure 3 we show the number of slices per time unit (left plot) and per energy unit (right plot) reconstructed using the CPU (OpenMP version) and the GPU (CUDA version) for the two different architectures and for three characteristic images from our Datasets.

[1] The batch of threads that executes a kernel is organized in a *grid* of thread *blocks*.

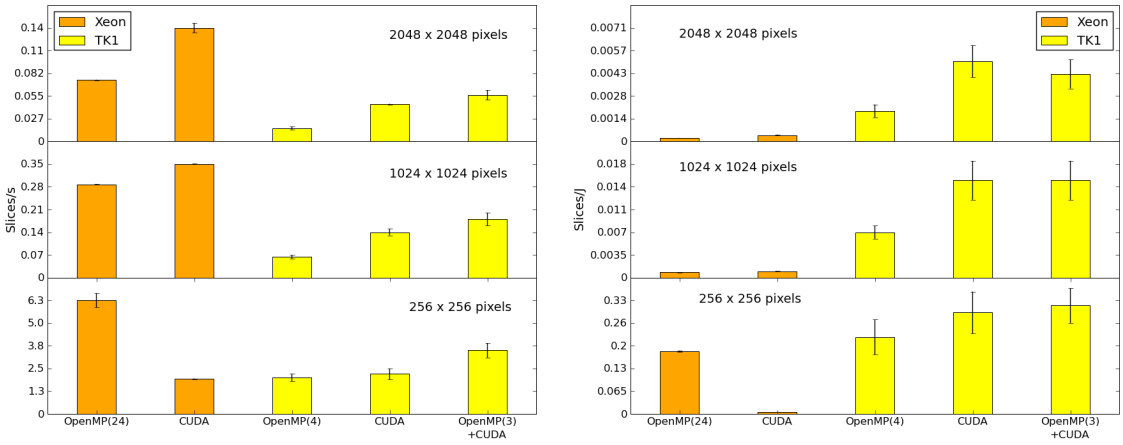


Figure 3. Slices per second (left) and slices per Joule (right) of OpenMP and CUDA versions of the Filtered Back-projection algorithm executed on Xeon (orange) and TK1 boards (yellow). The reconstructed images belong to the Kongo_2048 (top), Kongo_1024 (middle) and Kongo_256 (bottom) Datasets. The rightmost bar shows the combined GPU+CPU solution on TK1, i.e. only 3 CPU threads are used for the OpenMP version (see Section 4.1). Otherwise, the OpenMP version uses all the available threads, i.e. 24 threads on Xeon and 4 threads on TK1.

Of course, the HPC node guarantees a higher speed than the low power SoC, and thus a bigger number of reconstructed slice per second. However, when it comes to energy efficiency the bars in Figure 3 show the opposite behavior, and the SoC TK1 performs much better than the Xeon.

Incidentally, the left plot in Figure 3 also shows that the GPU-version of the algorithm is generally faster than the multi-threaded version, due to the higher number of available threads. Still, when processing images of small sizes (see the bottom row), performance bottlenecks in the data transfer to and from CUDA device arise, and the OpenMP version allows to reconstruct more slices per second than the CUDA version.

4.1. The proposed solution: CPU and GPU combined, portable and low power

Based on our tests, we speculated that the best performances could be achieved using both the CPU and GPU of a TK1 board, i.e. executing the OpenMP version and the CUDA version of the algorithm in parallel on the SoC.

In practice, this can be done running the multi-threaded OpenMP version of the algorithm on three of the four available CPU cores in the TK1, while the fourth and last available core executes the GPU version of the algorithm. The slices reconstructed per time unit and per energy unit with such configuration are shown by the rightmost bars in both the left and right plots of Figure 3.

With such approach, indeed, only three TK1 boards allow to reconstruct as many images per time unit as a traditional server, but consuming one order of magnitude less electrical power.

4.2. Correctness of the reconstructed images

The images reconstructed with the different versions of the Filtered Back-projection algorithm and for the different architectures were compared in terms of pixel-by-pixel standard deviations with the image reconstructed using the original, serial code.

In particular, we assumed that for values of the standard deviations smaller than the minimum detectable value in a grayscale, i.e. the ratio between the maximum intensity in an image and the total number of distinguishable levels in the grayscale, two slices can safely be considered equivalent (see Dynamic Range, e.g. [8]). In this way, we checked that the slices reconstructed with all the implementations of the algorithm used in this work, and for both Xeon and TK1, were consistent with the slices reconstructed using the original code.

5. Conclusions and future work

In this paper, we presented our experience of porting the Filtered Back-projection algorithm used for 3D Computed Tomography reconstruction to a low power, low cost system-on-chip, the NVIDIA Tegra K1 (TK1). The porting, which was done in two programming languages, OpenMP and CUDA, so to exploit both the CPU and GPU available on the system-on-chip, was straightforward. The correctness of the output of the application was always checked.

Performances were measured in terms of number of 2D slices (of a 3D volume) reconstructed per time unit and per energy unit, and compared with those obtained on a traditional x86 HPC node (Xeon) accelerated with a NVIDIA K20 GPU.

If on one hand the HPC node provides better absolute performance, i.e. number of reconstructed slices per second, than the SoC architecture, on the other hand, the SoC results up to 30 times more energy-efficient, i.e. number of reconstructed slices per Joule.

Finally, we proposed a combined OpenMP/CUDA approach to run the application, that allows to obtain the same performance (slices/s) of the HPC node using only three TK1 boards in a portable, cheap, fast, reliable and power saving device.

We are well-aware of the slightly unfair comparison described here, between a development board and a server built for harsh production environments, with multiple power supplies, redundant fans, multiple hard disks, etc. However, we believe that this work provides good indication of the potential of modern SoCs for specific kinds of applications, in particular those that manage to exploit the power of GPU.

In the future, we will work on improving the CUDA version of the algorithm using the so-called streams, in order to better exploit the GPU, which in our tests was never fully loaded, with a usage efficiency close to 60%. Moreover, we will try to improve the data management for the array transfers between host and device, as the current implementation is not optimized for small sizes of the images.

References

[1] Brancaccio, R., Bettuzzi, M., Casali, F., Morigi, M.P., Levi, G., Gallo, A., Marchetti, G., and Schneberket, D., IEEE TRANSACTIONS ON NUCLEAR SCIENCE, 58, 4, 2011

[2] Kak, A.C., and Slaney, M., Principles of Computerized Tomographic Imaging, IEEE Press, 1988

[3] Brancaccio, R., Bettuzzi, M., Casali, F., Morigi, M.P., Levi, G., Gallo, A., Marchetti, G., and Schneberket, D., ART'11 10th International conference on non destructive investigations and microanalysis for the diagnostics and conservation of cultural and environmental heritage., FLORENCE, AIPnD, 1-8, 2011

[4] Casali, F., Morigi, M.P., Bettuzzi, M., Berdondini, A., Brancaccio, R., and D'Errico, V., Restaurare L'Oriente - Sculture lignee giapponesi per il MAO di Torino, Nardini Editore, Collana Cronache 1, 38- 43, 2008

[5] NUMERICAL RECIPES IN C: THE ART OF SCIENTIFIC COMPUTING (ISBN 0-521-43108-5)

[6] Fast Fourier Transform (FFTs) and Graphical Processing Units (GPUs), slide 19, Kate Despain, CMSC828e.

[7] `http://www.sharcnet.ca/~merz/CUDA\_benchFFT/`

[8] Bettuzzi, M., Brancaccio, R., Morigi, M.P., and Casali, F., Effective dynamic range measurement for a CCD in full-field industrial X-ray imaging applications, Proc. SPIE 6616, Optical Measurement Systems for Industrial Inspection V, 2007

Knowledge Transfer

External Projects and Technology Transfer

M C Vistoli, A Ferraro and B Martelli

1. Introduction

The transfer of knowledge and technology from publicly-funded research institutions to society at large has attracted increasing attention in recent years.

In parallel the search for funding beyond the ordinary budget allocated to the Institute represents a strategic activity in order to sustain new and existing technological research by the centre.

In recent years a new activity has been opened at CNAF, focused on the above goals. The activity has been consolidated during 2014, with the plan to fully integrate it in the CNAF organization.

2. Activities in 2014

The group assigned to the unit "External Projects and Technology Transfer" is composed of three people. During the year they have followed primarily the following activities:

- Management of the MIUR-funded Open City Platform (OCP). The project, approved by the Ministry at the beginning of 2014, was followed in all the stages: definition of the budget and of its amendments, definition of the contract, management of the relashionship with the project "Social Innavation ICT for Citizen" when the Ministry associated it to OCP, recruitment of personnel at the participating INFN sites, interfacing with the Ministry, presentation of the project activities, e.g. at the CCR meetings.
- Presentation of several project proposals in response to EU calls in the field of Research and Development in ICT.
- Contribution to the revision of the national computing infrastructure, as part of the EU-T0 initiative.
- Opening of a new line of research with the purpose to investigate the use of low-power computing devices (Systems on Chip) to be applied to the typical applications developed and used by INFN researchers. The activity was then included in a dedicated INFN project called "Computing on SoC architecture" (COSA), within the context of the V National Scientific Commission of INFN (CSN5).
- Participation to the activities of the INFN-wide Commission on Technology Transfer. Within this context in particular we are involved in the definition of a policy, based on open-source principles, on how to properly managed the intellectual property derived from all the software developed within the Institute, with the dual objective to allow it to be easily shared for research purposes and, when possible, even become object of technology transfer for business reasons.
- Involvment in the "Rete Alta Tecnologia" (High Technology Network) of the Emilia-Romagna region. The network groups together private and public laboratories and

innovation centres, with the purpose to offer competences, instruments and resources to the local economic system. In 2014 INFN has become a shareholder of ASTER, which is the agency that coordinates the Network.
As a specific contribution to the initiatives promoted by ASTER, we have been involved in the drafting of the Smart Specialization Strategy (S3), which represents the foundation for the plan of POR/FESR for the years 2014-2020. Additionally we have actively participated to the Research-to-Business (R2B) fair in Bologna.

- Together with Regione Emilia Romagna we have also started the work for the foundation of a laboratory of industrial research, putting together the competences available in all the three INFN sites on the regional territory, namely Bologna, Ferrara and CNAF itself.
- Involvement in the coordination of the "Cluster Nazionale Fabbrica Intelligente".
- Definition, together with ASTER and CINECA, of the regional computing infrastructure, which has been subsequently identified by the Region and by the EU as one of the main three research infrastructures available in the regional territory and, as such, has been included in the "Piano Nazionale della Ricerca".

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*

User Support

Head: D. Cesini

M. Tenti
A. Falabella
L. Morganti
S. A. Tupputi
S. Taneja

Tier1

Head: L. dell'Agnello

Farming	**Storage**	**Networking**	**Infrastructure**
A. Chierici	V. Sapunenko	S. Zani	M. Onofri
S. Dal Pra	A. Cavalli	L. Chiarelli[1]	M. Donatelli
G. Misurelli	D. Gregori	D. De Girolamo	A. Mazza
A. Simonetto	M. Pezzi	F. Rosso	
S. Virgilio	A. Prosperini		
	P. Ricci		

[1] GARR employee relocated at CNAF

R&D Service

Head: D. Salomoni

D. Andreotti	M. Bencivenni	A. Ceccanti	V. Ciaschini
G. Dalla Torre	E. Fattibene	M. Favaro	F. Giacomini
M. Manzali	D. Michelotto	A. Paolini	E. Ronchieri
P. Veronesi	E. Vianello	G. Zizzi	

National ICT Services

Head: R. Veraldi

S. Antonelli

Technology Transfer

Head: M. C. Vistoli

A. Ferraro	B. Martelli

Hardware and Software Support

Head: G. Vita Finzi

Information System

Head: G. Guizzunti

S. Bovina	M. Canaparo	E. Capannini	F. Capannini
S. Cattabriga	C. Galli	S. Longo	C. Simoni

Director Office

Head: A. Marchesi

Expenditure Centralization Office[2]

Head: M. Pischedda

[2]The office is under the INFN Director General.

Seminars

Jan. 9^{th}	Filippo Mantovani **Supercomputing based on Mobile Processors**
Jan. 15^{th}	Diego Michelotto, Marco Bencivenni **Un Portale Web per Comunità Scientifiche**
Mar. 4^{th}	Paolo Veronesi, Giuseppe Misurelli **Gestione del ciclo di vita dei server con Foreman e Puppet**
Mar. 14^{th}	Pedro Andrade **Exploiting open source tools to realize a new monitoring infrastructure at CERN**
Apr. 4^{th}	Davide Salomoni **Report da ISGC 2014**
Jul. 16^{th}	Matteo Favaro **Ceph, architettura e utilizzo**
Jul. 23^{th}	Fabio Capannini **Orchestration in OpenStack with Heat**
Sep. 19^{th}	Andrea Petrucci **Experience with Infiniband for CMS Event Building**
Oct. 3^{rd}	Alberto Di Meglio **CERN openlab, un modello di collaborazione tra ricerca e industria**
Oct. 7^{th}	Elisabetta Ronchieri **Report dalla conferenza RPSD 2014**
Oct. 7^{th}	Salvatore Tupputi **Report dalla conferenza GPU in HEP**
Oct. 20^{th}	Tim Mattson **Programming Extreme Scale Computers**
Nov. 18^{th}, 20^{th}	Francesco Giacomini **Modern C++ – From Pointers to Values**

Dec. 2^{nd}	Davide Salomoni **Report da SuperComputing 2014**
Dec. 2^{nd}	Alessandro Paolini, Enrico Fattibene **Report da OpenStack Summit 2014**
Dec. 4^{th}	Rene Meusel **Introduction to the CernVM-File System**
Dec. 9^{th}	Matteo Panella **Casi d'uso di infrastrutture OpenStack: file system on-demand e database su container**
Dec. 19^{th}	Lorenzo Dini **Building Software at Google Scale**
Dec. 22^{nd}	Leonardo Testi **ALMA: alla scoperta dell'Universo freddo**