INFN Tier-1: a distributed site

Luca dell'Agnello[1,*], *Tommaso* Boccali[2], *Daniele* Cesini[1], *Lorenzo* Chiarelli[3], *Andrea* Chierici[1], *Stefano* Dal Pra[1], *Donato* De Girolamo[1], *Antonio* Falabella[1], *Enrico* Fattibene[1], *Gaetano* Maron[1,4], *Diego* Michelotto[1], *Lucia* Morganti[1], *Andrea* Prosperini[1], *Vladimir* Sapunenko[1], and *Stefano* Zani[1]

[1]INFN CNAF, v.le B. Pichat 6/2 - 40100 Bologna, IT
[2]INFN Sezione di Pisa, Largo B. Pontecorvo 3, 56127 Pisa, IT
[3]GARR Consortium, c/o INFN-CNAF, v.le B. Pichat 6/2, Bologna 40100, IT
[4]INFN Laboratori Nazionali di Legnaro, Via dell'Università 2, 35020 Legnaro, IT

Abstract. The INFN Tier-1 center at CNAF has been extended in 2016 and 2017 in order to include a small amount of resources (~ 22 kHS06 corresponding to ~ 10% of the CNAF pledges for LHC in 2017) physically located at the Bari-ReCas site (~ 600 km distant from CNAF). In 2018, a significant fraction of the CPU power (~ 170 kHS06, equivalent to ~ 50% of the total CNAF pledges) is going to be provided via a collaboration with the PRACE Tier-0 CINECA center (a few km from CNAF), thus building a truly geographically distributed (WAN) center. The two sites are going to be interconnected via an high bandwidth link (400-1200 Gb/s), in order to ensure a transparent access to data residing on CNAF storage; the latency between the centers is small enough not to require particular caching strategies. In this contribution we describe the issues and the results of the production configuration, focusing both on the management aspects and on the performance provided to end-users.

1 Introduction

The National Institute for Nuclear Physics (INFN) is the research agency, funded by the Italian government, dedicated to the study of the fundamental constituents of matter and the laws that govern them. The INFN is composed of more than 20 divisions dislocated at the main Italian University Physics Departments, four Laboratories and three National Centers dedicated to specific tasks (Fig. 1).

CNAF is the National Center of the INFN "for the Research and Development in Information and Communication Technologies": it participated as a primary contributor in the development of Grid middleware and in the initial setup and maintenance of the Italian Grid infrastructure. The first incarnation of the CNAF data center dates back to 2003 as computing facility for the BaBar, CDF and Virgo experiments as well as a prototype for the INFN Tier-1 for the future high-energy physics experiments at the Large Hadron Collider (LHC) in Geneva: ALICE, ATLAS, CMS, and LHCb.

Nowadays, besides Virgo and the four experiments at LHC, the INFN Tier-1 provides resources, support and services needed for all the activities of data storage and distribution, data

*e-mail: luca.dellagnello@cnaf.infn.it

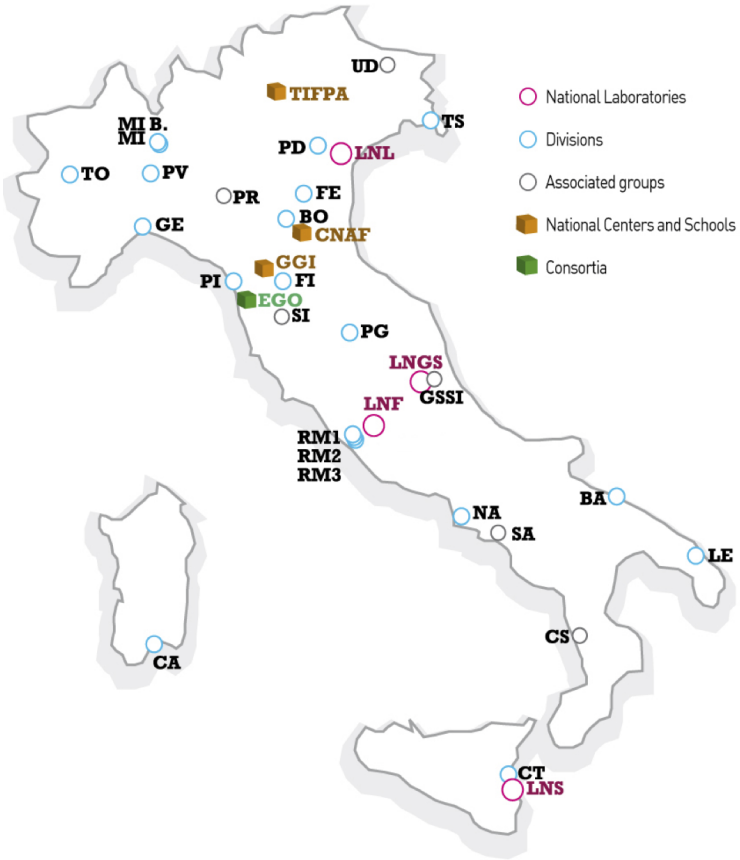

Figure 1: Dislocation of the INFN divisions

processing, Monte Carlo production and data analysis to ~ 30 other scientific collaborations, including Belle II and several astro-particle experiments.

The needed resources for the experiments in 2018 amount to 330 kHS06 for the computing farm, ~ 34 PB of disk and ~ 93 PB of tapes.

2 INFN Tier-1 beyond CNAF data center

Since 2015 INFN-CNAF has started a R&D program aiming to the utilization of remote CPU resources to extend our data center beyond CNAF premises. In addition to contingent economic savings with the long-term lease of CPU resources available at other sites, we wanted to evaluate new operational models (for example, possibly to accommodate unexpected CPU requests, through the so-called "cloud bursting" on commercial providers) and look for a solution to overcome the infrastructural limits of the current data center location, especially in view of the foreseen increase of resources needed for HL-LHC. Our main goal for any extension of the farm has always been, as much as possible, to make it transparent to users: i.e. the access to these resources should be possible through the usual 'gates' (local scheduler

and CEs) and their use should not introduce inefficiencies of any sort. The key issue for the last point is the remote access to data, possibly eased by a cache located in the remote farm.

3 Tests in the R&D program

We are testing several types of data center extensions, some of them already reached a production quality level:

- we performed functional tests of opportunistic usage of commercial clouds;
- we are part of the European project HNSciCloud for the adoption of commercial clouds in research environments;
- since 2017, we are using in production a static set of remote resources located in Bari;
- since March 2018, we are using in production a set of resources located at CINECA[1];
- we are planning to test also an extension of our HPC farm on CINECA.

3.1 Opportunistic computing on commercial clouds

We have been testing opportunistic computing via the dynamic extension of the farm on two different clouds:

- Aruba cloud (2015-2016): a small scale test (up to 150 cores max) on idle resources (see [1])
- Azure cloud (2017): grant for the utilization of 25,000 $ given by Microsoft to CMS to use MS Azure Cloud Infrastructure

Both these tests have been performed in collaboration with CMS experiment. The setup was made with an in-house developed application, `dynfarm`, responsible for the setup of a VPN tunnel between the remote cloud hosts and the local farm elements (CEs[2], the batch system and Argus service for authentication). Through the VPN, the LSF binaries and the log files are synchronized with a local cache too (see [2] for further details). Experimental data were not made accessible through the cache: CMS jobs can retrieve needed files via XRootD protocol from remote locations (an ad hoc XRootD redirector can be added in the remote cloud) using the General Purpose Network (hence not through the VPN link). Mainly due to the use of General Internet to retrieve data, the efficiency of jobs is good only for MonteCarlo ones (otherwise the overall efficiency[3] is ~ 0.4, very low if compared with ~ 0.8 for jobs running at CNAF).

3.2 Farm extension on HNSciCloud resources

In the framework of HNSciCloud European project[4] INFN has performed extensive scalability tests both with full site instantiation (i.e. using DODAS [3] for CMS and Vcycle [4] for Belle II) and elastic extension of the Tier-1 farm [5]. The configuration for the latter case is similar to the one described in Par. 3.1 (use of dynfarm and small cache for the remote nodes): hence it has been used mainly for MonteCarlo jobs (for both Atlas and CMS). The availability of these resources gave us the possibility to completely automate the process of

[1] Located in Bologna such as CNAF, CINECA is the Italian Tier-0 for PRACE.

[2] CE is the acronym for Computing Element, the grid element acting as gateway to a computing farm.

[3] The efficiency of a completed job is defined as the ratio CPT/WCT.

[4] Helix Nebula – The Science Cloud with Grant Agreement 687614 is a Pre-Commercial Procurement Action funded by H2020 Framework Programme.

node creation and contextualization, interfacing not only with OpenStack APIs[5], already well known on our site, but also with CloudStack APIs[6]. During the extended test period, more than 70 VMs, corresponding to $\sim 7kHS06$ (nominal) HS06 were available on a separate queue, providing multicore resources.

The most important outcome of this test has been the evaluation of potential dynamic extensions of our computing farm to face sudden burst requests. We succeeded in instantiating tens of virtual nodes that reliably extended the farm but this gave also us the idea of the many limitations that this technology right now imposes. The main pro is of course the possibility to fulfill sudden peaks of requests with extreme ease and reduced effort. Anyway the cons right now appear still to be too many. One of the outcomes of HNSciCloud is that precisely estimating the cost of resources used on cloud is difficult, due to the particular utilisation of resources done by the LHC experiments (that is 100% resources utilisation 24h/7) and due to the possible overcommitment of resources the cloud provider may apply in order to reduce its costs. In addition, significant discounts are normally applied by cloud providers only when it comes to long-term contracts (e.g. at least 1 year): this does not correspond to the case of using a sudden peak of resource demand (cloud bursting).

3.3 Farm extension to Bari-ReCaS

Since the beginning of 2017, a fraction (~22 kHS06) of the pledged computing resources for the WLCG experiments at the INFN Tier-1 is actually located in Bari-RECAS data center. Unlike previous cases (that can be merely considered only tests), we have had to implement a setup to accommodate as many workflows as possible, in order to make it as if belonging to the internal farm at CNAF. Since Bari is $\sim$ 600 km from Bologna, the latency of the network connection is not negligible[7]: to allow the jobs on the remote farm partition to transparently access the data from the storage at CNAF, we set up a Layer 3 VPN between Tier-1 and Bari-ReCaS data centers with a dedicated 20 Gbps link. After that we configured a cache in Bari to extend the CNAF file-systems, allowing jobs to access data directly through Posix interface. It's important to notice that Alice jobs do not need this cache since data is accessed via XRootD (CMS does the same as a fallback). Moreover, some other auxiliary services (e.g. squids) have been replicated in Bari. Resources both at CNAF and Bari are accessed through the same entry points, CNAF CEs and LSF server: moreover, all traffic with farm in Bari is routed via CNAF [6]. Despite the cache and the good connectivity with storage at CNAF, we have registered low values for the efficiency of jobs with intensive throughput, forcing Atlas experiment to differentiate the use of Bari-ReCaS farm partition from CNAF one. On the other hand, for certain workflows of the CMS experiment, we have observed better performances using direct access via XRootd to the data at CNAF instead of accessing them through the cache. The counter-intuitive result observed for CMS jobs is probably due to one or more of the following factors: the optimization of the code to deal with the latency on the WAN, an improper configuration of the cache (with a continuous flow of data coming in and out, suggesting therefore the uselessness of the cache itself) or, more probably, latency imposed by the hardware itself. To disentangle these effects, a larger capacity system would have been needed in order to allow some period of persistence of the data on the cache and, hopefully, observe some improvements.

[5]https://developer.openstack.org/api-guide/quick-start/
[6]https://cloudstack.apache.org/api.html
[7]The Round Trip Time from CNAF to Bari is $\sim 10ms$.

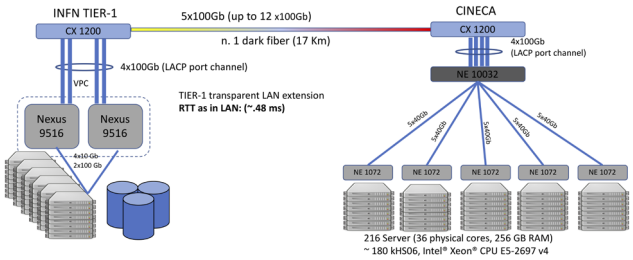

Figure 2: Schematic view of the CINECA - INFN Tier-1 interconnection

4 Farm extension to CINECA

The real turning point for the extension of CNAF data center has been the 'relocation' at CINECA of half of the farm, after the stop due to the flooding [7]. CINECA, located in Bologna ~ 17 km from CNAF, is the Italian supercomputing center and Tier-0 for PRACE. In 2017, INFN and CINECA signed an agreement for the long-term lease of three racks of computing nodes, dismissed from the 'A1 partition of Marconi[8]'. These three racks account for a computing power of ~ 180 kHS06 distributed among 216 servers.

All the nodes been refurbished with additional memory and a standard network interface: each server has now a 10 Gbit uplink connection to the rack switch while each of them, in turn, is connected to the aggregation router with 4x40 Gbit links. Like in the case of Bari-ReCaS, the logical network of the farm partition at CINECA is set as an extension of INFN Tier-1 LAN: a dedicated fiber couple interconnects the aggregation router at CINECA with the core switch at the INFN Tier-1 (see Fig. 2). The fiber is managed by a couple of Infinera Data Center Interconnect (DCI)[9] systems, able to multiplex several wavelengths thus providing a 500 Gbps (upgradable to 1.2 Tbps).

Users access CINECA partition transparently through the batch system and the CEs at CNAF. Since the latency between storage at CNAF and farm at CINECA is comparable to the internal farm (RTT: 0.48 ms vs. 0.28 ms on Tier-1 LAN), there is no need for a cache and so the servers at CINECA access directly the file-systems at CNAF.

These nodes, are in production since March 2018 for WLCG experiments (see Fig. 3) and are being gradually opened to other collaborations.

Like Bari-ReCaS, the servers at CINECA are fully managed (installation, configuration and management) by INFN Tier-1 staff while the physical layer (i.e. network configuration, hardware support) is controlled by CINECA staff. After some initial hiccups and iterations, an escalation procedure has been completely refined and a very good collaboration has been established.

However, this partition has undergone several reconfigurations due to both the hardware and the type of workflow of the experiments. In April we had to upgrade the BIOS to overcome a bug which was preventing the full resource usage, limiting at ~ 78% of the total what we were getting from the nodes. Moreover a reconfiguration of the local RAID configuration of disks is ongoing[10] as well as tests to choose the best number of computing slots.

[8]MARCONI is the Tier-0 system in production at CINECA since 2016

[9]https://www.infinera.com/wp-content/uploads/infinera-ds-Cloud-Xpress-2.pdf

[10]The initial choice of using RAID-1 for local disks instead of RAID-0 has been proven to slow down the system even if safer from an operational point of view.

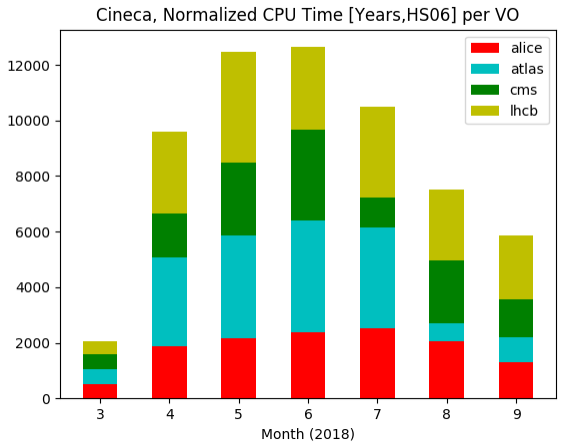

Figure 3: CPU time per month of jobs running at CINECA

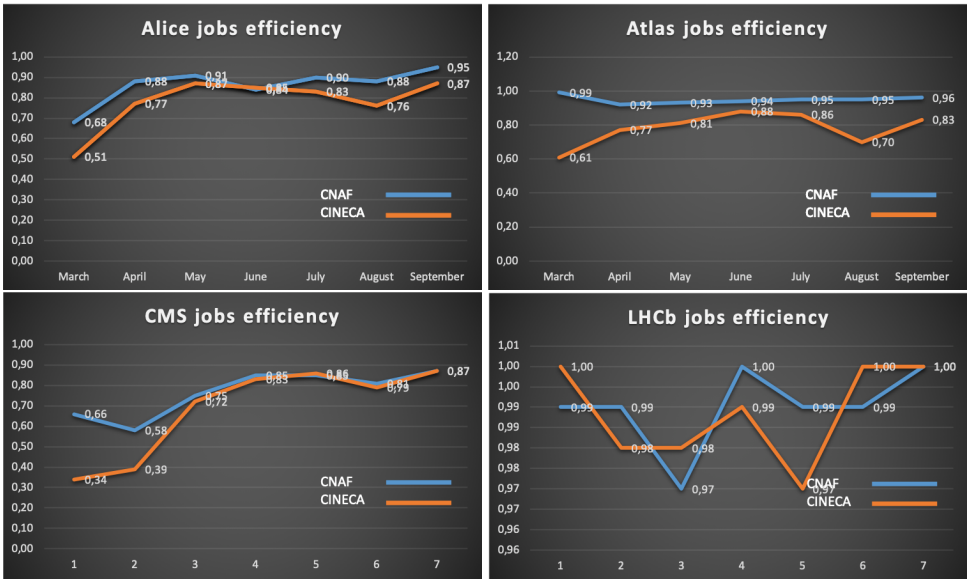

Figure 4: Comparison between job efficiency running at CNAF and CINECA

In parallel, since this partition has been installed with CentOS7, tests with singularity [8] for some experiments (mainly CMS and Atlas) have been done.

With the exception of CMS, the efficiency of the jobs of the main experiments (see Fig. 4) is a bit lower than that experienced at CNAF but this also reflects the various reconfigurations performed: we expect to achieve optimal performances only after the end of the hardware reconfiguration process.

4.1 HPC

Another opportunity, for the computing of the LHC experiments, is using the nodes from the CINECA Marconi A2 partition[11]. These nodes are completely controlled by CINECA and are accessible either opportunistically or with a grant. Technically, the main differences with the previous case are the type of the processor (KNL for a total of 68x4 cores per node, 96 GB of RAM) and the lack of external network connectivity.

Functional tests have been completed, proving the possibility to submit jobs to the CINECA local batch sysyem (Slurm) via the CNAF CE.

The Marconi A2 hosts have been configured by the CINECA sysadmins in order to support the most important requests by the LHC experiments: CVMFS has been added in order to allow the access to the software and calibrations, singularity has been installed in order not to require the installation of GRID middleware at the host level, and already present squids at CINECA have been allowed to cache CVMFS and Frontier payload. The standard A2 nodes at CINECA do not allow for external network connectivity, preventing LHC experiments to access remote data-sets. This has been solved by allowing CINECA nodes to route directly with CERN and CNAF networks. The former is needed in order to access calibrations and communicate back to central experiment service: the latter has allowed to install at CNAF an XrootD proxy which makes the full XrootD federation visible to the experiments, via an Xcache setup [9]. With such a setup, there is no limitation in the type of workflows which can technically be executed on A2 nodes. For what concerns the network bandwidth, the nodes are planned to be reconfigured in order to utilize the VPN to CNAF and share a fraction of the bandwidth available through it.

Since the functional tests have been positive, the LHC Italian computing groups are submitting a PRACE request for grant in order to be allowed to use up to 100 Million core hours on Marconi A2 Prace system.

5 Conclusions

The increasing demand of computing resources led to the investigation of several techniques to extend the existing farm of INFN Tier-1. This R&D program aims to allow for extensions transparent to the users, on a variety of systems.

Commercial Clouds are a possibility WLCG and the experiments are evaluating, mainly in case of burst requests that can't fit our current computing power. In this respect, tests made with Aruba, Azure and mainly with HNSciCloud project, proved this option to be feasible.

For what concerns the utilization of remote private resources, like WLCG Tier-2s or other academic centres, the mechanism has proved to be realistic: indeed it's been in operations with INFN-Bari for more than one year The Bari-Bologna distance has required research on caching technologies, which proved to be a pillar in the design of every long latency/low bandwidth scenarios.

The large fraction of INFN Tier-1 resources now at CINECA shows the possible operational model for the years to come, where CPU resources will be obtained from specialized centers, requesting a LAN-like network interconnection with CNAF storage. To the authors' knowledge, it is the first time in Italy two academic centres are connected with a DCI technology.

Without the utilization of grants on HPC centres, beyond year 2020, LHC Computing is probably going to face a big issue in obtaining all the resources required. Our experimentation was in the direction of showing how a major (PRACE Tier-0) HPC center can be used

[11] https://www.cineca.it/en/content/marconi

for LHC computing with small technical changes, mostly involving resource policies. No showstoppers have been found with small scale tests; in case the proposal is granted, further tests will try to show the usability of the solution for real mission critical LHC activities.

References

[1] S. Dal Pra, V. Ciaschini, L. dell'Agnello et al., "Elastic CNAF Data Center extension via opportunistic resources", PoS(ISGC 2016)031 International Symposium on Grids and Clouds (ISGC) 2016 - Infrastructure Clouds and Virtualisation, (2016), doi: 10.22323/1.270.0031

[2] V. Ciaschini and D. De Girolamo, "Dynfarm: A Dynamic Site Extension", J. Phys.: Conf. Ser., **898**, 082017 (2017), doi: 10.1088/1742-6596/898/8/082017

[3] D. Spiga et al., "DODAS: How to effectively exploit heterogeneous clouds for scientific computations", PoS(ISGC 2018 & FCDD) 024 (2018), doi: 10.22323/1.327.0024

[4] https://www.gridpp.ac.uk/vcycle/

[5] https://www.hnscicloud.eu/

[6] T. Boccali et al, "Extending the farm on external sites: the INFN Tier-1 experience", J. Phys.: Conf. Ser. **898**, 082018 (2017), doi:10.1088/1742-6596/898/8/082018

[7] L. dell'Agnello, "The flood" in INFN-CNAF Annual Report 2017, 154-162 (2017), ISSN 2283-5490 (online)

[8] http://singularity.lbl.gov

[9] http://slateci.io/XCache/