Journal of Physics: Conference Series

PAPER • OPEN ACCESS

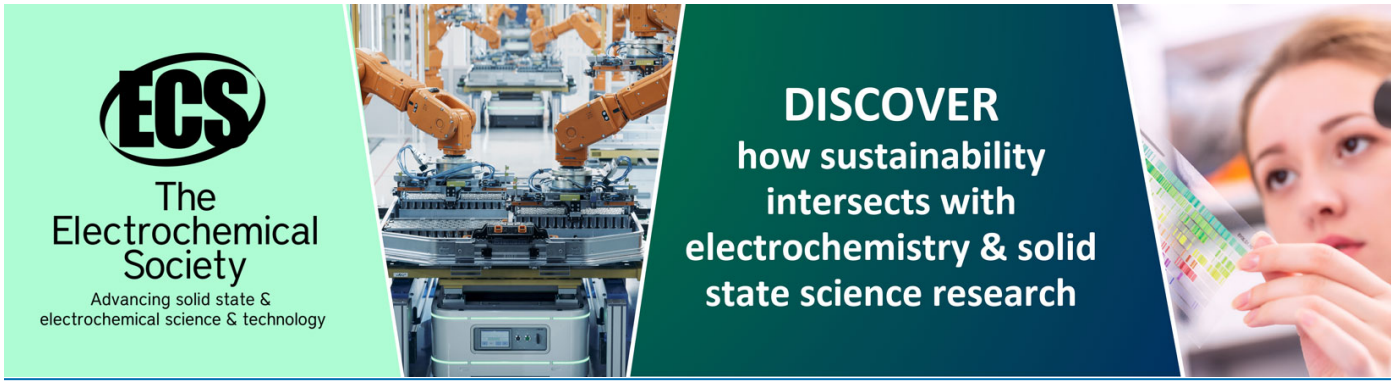
Extending the farm on external sites: the INFN Tier-1 experience

To cite this article: T Boccali *et al* 2017 *J. Phys.: Conf. Ser.* **898** 082018

View the article online for updates and enhancements.

You may also like

- A Review: Different Challenges in Energy-Efficient Cloud Security
Poonam Kumari and Meeta singh
- Development of Computer Network Security Based on Cloud Computing
Yizhi Li
- Belle II public and private cloud management in VMDIRAC system.
Rafa Grzymkowski, Takanori Hara and on behalf of the Belle II computing group



This content was downloaded from IP address 3.135.216.174 on 25/04/2024 at 23:32

IOP Conf. Series: Journal of Physics: Conf. Series **898** (2017) 082018 doi:10.1088/1742-6596/898/8/082018

Extending the farm on external sites: the INFN Tier-1 experience

T Boccali[1], A Cavalli[2], L Chiarelli[3], A Chierici[2], D Cesini[2], V Ciaschini[2], S Dal Pra[2], L dell'Agnello[2], D De Girolamo[2], A Falabella[2], E Fattibene[2], G Maron[2,4], A Prosperini[2], V Sapunenko[2], S Virgilio[2] and S Zani[2]

[1] INFN Sezione di Pisa, Largo B. Pontecorvo 3, Pisa 56127, IT
[2] INFN-CNAF, v.le B. Pichat 6/2, Bologna 40100, 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

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

Abstract. The Tier-1 at CNAF is the main INFN computing facility offering computing and storage resources to more than 30 different scientific collaborations including the 4 experiments at the LHC. It is also foreseen a huge increase in computing needs in the following years mainly driven by the experiments at the LHC (especially starting with the run 3 from 2021) but also by other upcoming experiments such as CTA[1] While we are considering the upgrade of the infrastructure of our data center, we are also evaluating the possibility of using CPU resources available in other data centres or even leased from commercial cloud providers. Hence, at INFN Tier-1, besides participating to the EU project HNSciCloud, we have also pledged a small amount of computing resources ($\sim$ 2000 cores) located at the Bari ReCaS[2] for the WLCG experiments for 2016 and we are testing the use of resources provided by a commercial cloud provider. While the Bari ReCaS data center is directly connected to the GARR network[3] with the obvious advantage of a low latency and high bandwidth connection, in the case of the commercial provider we rely only on the General Purpose Network. In this paper we describe the set-up phase and the first results of these installations started in the last quarter of 2015, focusing on the issues that we have had to cope with and discussing the measured results in terms of efficiency.

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 by more than 20 divisions dislocated at the main Italian University Physics Departments, 4 Laboratories and 3 National Centres dedicated to specific tasks. CNAF is the National Center of the INFN "for the Research and Development in INFN Information and Communication Technologies": it participated as a primary contributor in the development of Grid middleware and then in the operation of the Italian Grid infrastructure. Since 2003, CNAF hosts the Italian Tier-1 for the high-energy physics experiments at the Large Hadron Collider (LHC) in Geneva, ALICE, ATLAS, CMS, and LHCb, providing the resources, support and services needed for all the activities of data storage and distribution, data processing, Monte Carlo production and data analysis. Nowadays, besides the four


Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

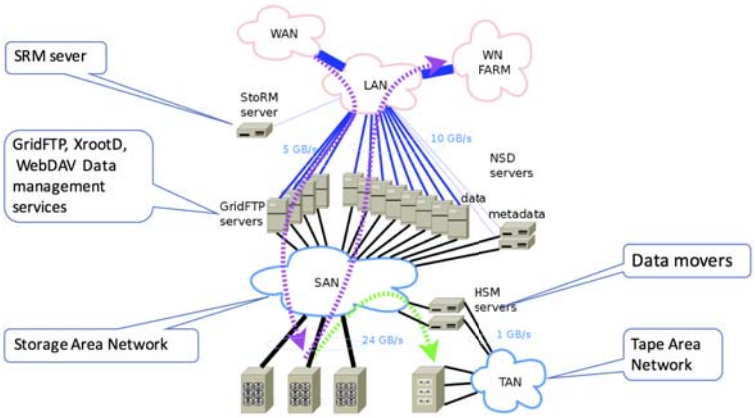


Figure 1: The INFN Tier-1 HSM layout

LHC experiments, the INFN Tier-1 provides services and resources to $\sim$ 30 other scientific collaborations, including BELLE2 and several astro-particle experiments.

2. INFN Tier-1 current status and configuration

Currently, the INFN Tier-1 hosts nearly 1,000 worker nodes (WNs) for a total amount of about 21,500 computing slots and a power capacity of 200k HS06[4]. All the computing resources are centrally managed by a single batch system, IBM Platform LSF[5] while HTCondor[6] is being evaluated as a possible alternative. The resources are dynamically allocated according to a fair-share policy, which prevents resources underutilization and user starvation. Also a small ($\sim$ 33 TFlops) HPC cluster with nodes interconnected via Infiniband is available for special applications.

CNAF also operates a large storage infrastructure based on industry standards: all disk servers and disk enclosures are interconnected through a dedicated Storage Area Network (SAN) and the data are hosted on several IBM GPFS[7] file systems, typically one per major experiment. These choices allowed the implementation of a completely redundant data access system. Currently there are $\sim$ 22 PB of net disk space in 15 file systems. Also a tape library, interconnected via a dedicated SAN, is available with, currently, $\sim$ 45 PB of used tape space. The tape system is managed by IBM TSM[8], integrated with GPFS to build a Hierarchical Storage Manager system[11] (Fig.1). The disk-servers are connected to the farm through multiples of 10 Gbps links: the aggregate bandwidth between the farm and the storage is 100 GByte/s. Besides POSIX, the data can be accessed through standard protocols and interfaces, as defined by the WLCG/EGI projects (GridFTP and SRM, XRootD, and WebDAV).

The data center is interconnected to LHCOPN[9] and LHCONE[10] networks with a dedicated link (60 Gbps where 20 Gbps are reserved for LHCONE) and has access to the General Internet with a 20 Gbps link. An upgrade of the link to LHCOPN and LHCONE to 2x100 Gbps is foreseen in 2017.

3. Computing cluster workload

The computing resources are used all the time at CNAF (see for example the usage of farm in 2016 depicted in Fig.2 where the sharp decreases are due to security upgrades or other

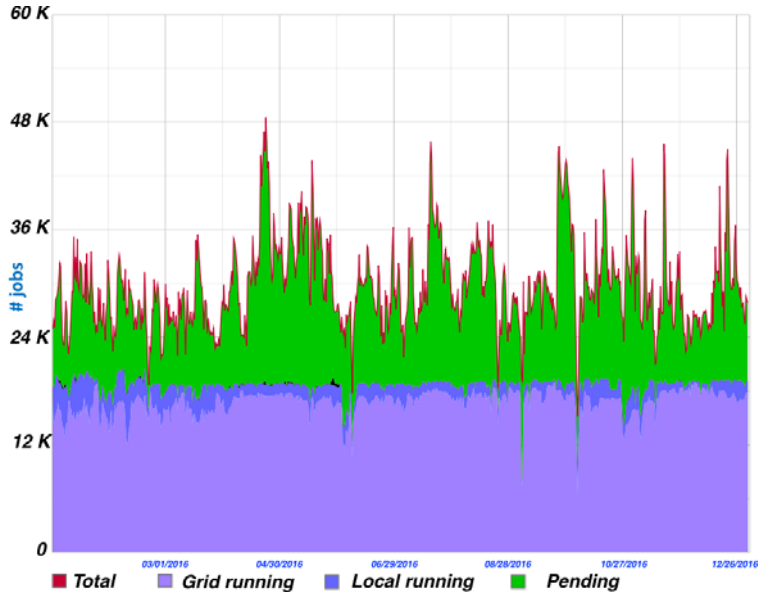


Figure 2: The INFN Tier-1 farm usage in 2016

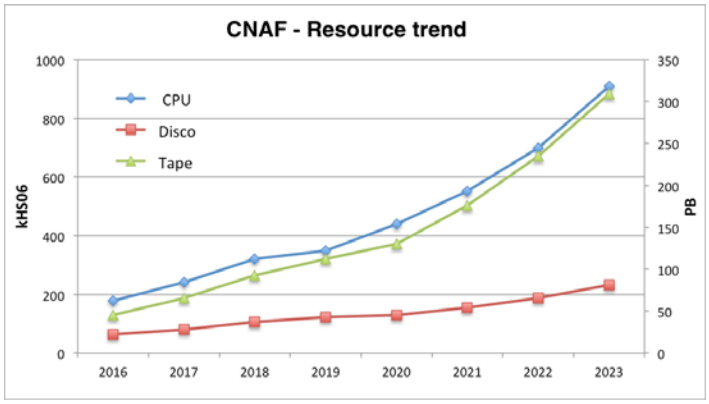


Figure 3: The foreseen trend of resources at INFN Tier-1 up to 2023

maintenance related tasks) with a large amount of waiting jobs ($\sim$ 50% of the running jobs).[1] This is paired by a good efficiency in data access and hence in job processing.

In the next years, a huge increase of resource request is foreseen, primarily due to the WLCG experiments, and a non-negligible contribution will come from astro-particle ones. By the end of 2023, i.e. at the end of LHC Run3, the amount of needed CPU power at INFN Tier-1 will be a almost a factor 5 larger than the quantity currently installed; roughly, the same scaling factor applies for disk and tape storage (Fig.3). Even if, according to our best estimates, the CNAF data center will be able to host the requested resources up to the end of Run 3, we started to test the usage of remote farms as a transparent (and possibly dynamic) extension of the Tier-1. In this way we will also be able to cope with unplanned requests of CPU resources, i.e. to cover peaks via the so-called cloud-bursting.

[1] The irregular profile visible on the first three months is due to "short jobs flooding" whose negative effect has been addressed and prevented as described in [12].

The tests are performed according to two different (but complementary) scenarios:

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

The first case is being addressed through tests with the Italian Cloud provider Aruba [13] while for the second one, a test has been performed during the 2016 extending INFN Tier-1 farm to Bari-ReCaS data center.

INFN Tier-1 is also participating to the European Pre Commercial Procurement project HelixNebula Science Cloud [14].

In the following sections, the setup and the results of a pre-production extension of the INFN-Tier-1 to the Bari-ReCaS data center will be presented.

4. Farm extension to Bari-ReCaS

The data center Bari-ReCaS is a common effort of INFN and Università degli Studi di Bari Aldo Moro. Active from July 2015, it is composed of 128 Worker Nodes (for a total of $\sim$ 8200 computing slots equivalent to $\sim$ 100k HS06), a small HPC Cluster (800 cores) with Infiniband, 3.6 PB of disk space and 2.5 PB of space on a tape library. A part of these resources ($\sim$ 25k HS06, 1.1 PB of disk) are allocated to CMS and Alice Tier-2. The distance between Bari-ReCaS and CNAF is $\sim$ 600 km with a Round-Trip-Time of $\sim$ 10 ms (Fig. 4).

Following an agreement with Bari-ReCaS, 48 servers, for a total amount of $\sim$ 20k HS06 of CPU power, were allocated to the INFN Tier-1 as extra-pledge resources for WLCG experiments. These servers provide $\sim$ 13% of additional computing power to WLCG experiments at the INFN Tier-1.

The primary goal of our test was to establish an extension to the INFN Tier-1 which would be transparent to users, meaning that the entry point for jobs would remain CNAF, while they could be dispatched to Bari-ReCaS nodes. A similar case is the one between CERN and Wigner[15] with the notable difference being the absence of storage in Bari-ReCaS (excluding a small cache for LSF and transient data).

Figure 4: Relative positions of CNAF and Bari-ReCaS

4.1. Network setup

The first step was the setup of a dedicated connection between INFN Tier-1 and Bari-ReCaS data center. The network link was dimensioned allowing $\sim$ 1 MB/s per core (on the Tier-1 LAN this value is slightly higher, 5 MB/s per core). Hence for 2,000 cores a 2x10 Gbps link has been established (in Fig.5 the commissioning of the link performed during the last quarter of 2015). This link is configured as a Level 3 VPN and two /22 CNAF subnets (one for the public IP addresses of WNS and the other one for their management interfaces) are being routed over this link and assigned to Bari-ReCaS nodes (obviously only a small fraction of these addresses is currently used). Even if connected to the common core switch in Bari-ReCaS, these nodes are isolated from the other ones in the data center: they access the WAN, including LHCONE and LHCOPN, through CNAF (see the layout of the connection in Fig.6).

4.2. Farm setup

The main goal of the testbed is to exploit, as already outlined, the remote WNs in Bari-ReCaS in the same way as those in CNAF, granting transparent access to the users, with no need of

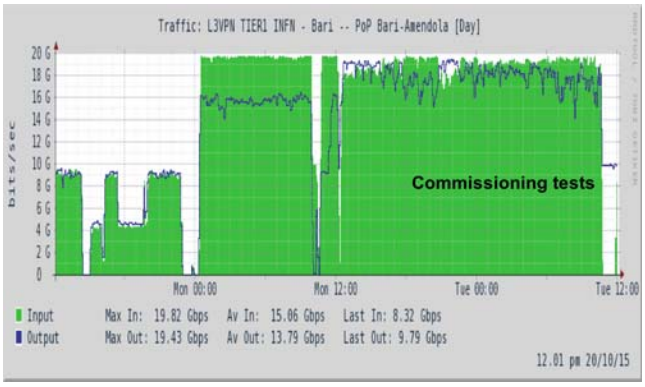


Figure 5: Commissioning of the VPN

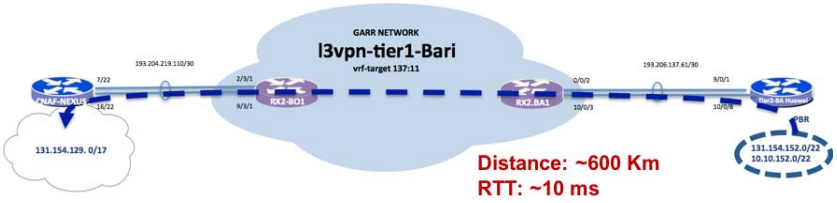


Figure 6: Layout of the VPN

any special configuration on their side. Hence WNs have been configured as part of CNAF LSF cluster and the entry points for the users are the standard Computing Elements of INFN Tier-1. A cache for the shared file-system for LSF has been configured. Some auxiliary services (namely the CVMFS Squid servers for software distribution and the Frontier Squid servers for the condition databases of ATLAS and CMS) have been replicated in Bari-ReCaS. Using different "views" in the DNS zone, WNs in Bari-ReCaS are redirected to the local instances instead of the ones at CNAF.

For all other services, including Kerberos for authentication and LDAP for authorization, the WNs access the instances at CNAF; also the installation and configuration of the WNs is performed from CNAF puppet/foreman servers.

4.3. Data access

To match the requirement of transparent use of remote resources, jobs need to access data the same way as they do at CNAF.

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

For this reason we studied how to implement the Posix access to a local cache for WNs in Bari-ReCaS. The natural choice was to exploit a native extension of GPFS, AFM[16][17], capable to implement a caching system (Fig. 7). Two 10 Gbps servers with $\sim$ 330 TB-N of disk (initially 100 TB–N only, exhibiting however too poor performances) have been allocated in Bari-ReCaS for AFM; the cache is configured in read–only mode to improve performances, as the output is written directly to CNAF using StoRM[18]. Alice data are not cached since their

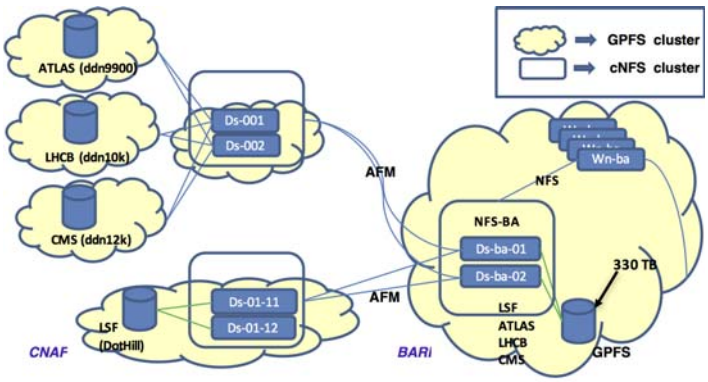


Figure 7: Layout of cache system in Bari-ReCaS

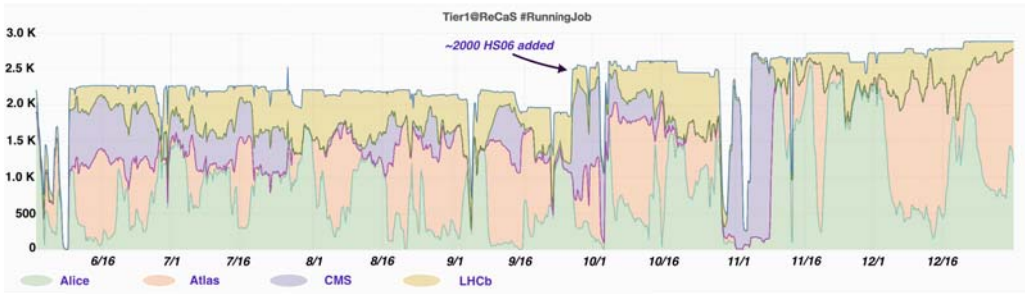


Figure 8: Farm usage in Bari-ReCaS in the second half of 2016

jobs only use remote access via Xrootd. One more small AFM cache, decoupled from the one for the data, has been configured for the LSF shared file-system.

During the first half of 2016, we have performed several tunings of the cache, including the enlargement of the storage space and a reconfiguration of GPFS to better fit with the underlying hardware: however the current infrastructure (network and storage system) does not allow for further improvements. Limiting factors are given by the size of the cache (it can contain up to 3% of the amount of data at CNAF), the maximum fill-in and fill-out speed (1 GB/s r/w aggregated) and obviously the pattern of data access which is not tunable at our side.

5. Results

The configuration of the farm and the caching system in Bari-ReCaS is stable since June 2016: since then the farm has been extensively and steadely used (Fig. 8) with swinging results (Fig. 9). Obtaining a reliable comparison between CNAF and Bari-ReCaS is quite difficult because the workload on the WNs of the two partitions is different: CNAF nodes are concurrently used by several VOs, with usually different mix, while only WLCG VOs can use WNs at Bari-ReCaS. Hence the efficiency of jobs (i.e. the ratio between the used CPU and Wall-clock time) can be penalized, expecially at CNAF, from occasional CPU overload due to jobs of other experiments running on the machine, resulting in a temporarily higher efficiency for Bari-ReCaS jobs.

On the long term, we have anyway observed that efficiency is better for I/O demanding jobs running at CNAF, noticeably for Atlas and CMS. The main reason is the low speed of the cache: in certain cases we have observed the data flowing continuously from CNAF to the cache and

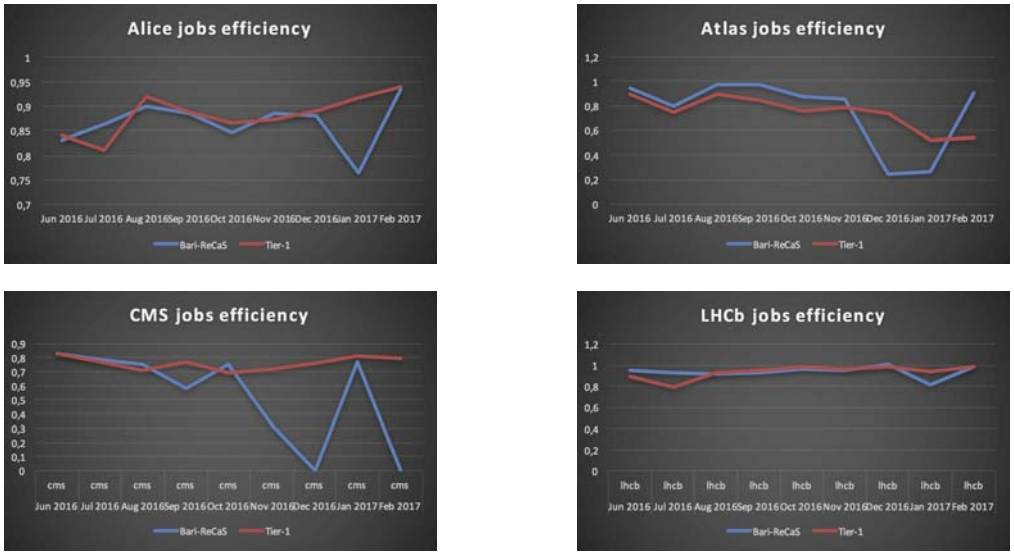


Figure 9: Tier-1 vs Bari-ReCaS Comparison

Table 1: Overall comparison of jobs efficiency (second half of 2016)

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

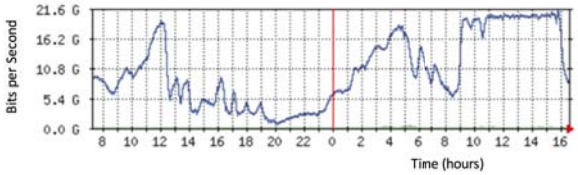


Figure 10: Saturation of VPN due to Alice jobs

then to the WNs, with the cache introducing a delay. As an attempt to overcome the problems with the cache, we plan to test a new and more performant storage system during the first semester of 2017. In the meanwhile, we have mitigated the issue by defining an ad-hoc queue for Atlas in LSF to dispatch low-I/O jobs only to Bari-ReCaS WNs; moreover we have had at times to inhibit the access to CMS jobs. Noticeably, LHCb jobs show a high efficiency also in Bari-ReCaS. On the other hand, Alice jobs, that do not use the cache, have an efficiency comparable to that of jobs running at CNAF (the difference is minimal); indeed remote data access could be a viable option, probably having a dedicated channel with a bandwidth larger than 20 Gbps. In any case, this method could not be extended to non-LHC experiments which need Posix access to the data.
Given these constrains for the use of Bari-ReCaS partition, we have, with the exception of CMS, comparable efficiency with CNAF (Tab.1).
Furthermore, with the current configuration, the remote access exploited by Alice jobs can occasionally saturate the VPN for some hours, hence interferring with the cache system (see for example Fig.10). For these reasons, we are also considering to increase the bandwidth of the VPN: this will be possible starting from the second half of 2017.

6. Conclusions

INFN Tier-1 has a strong commitment and is fully addressing the computing requirements of physics experiments. Even if the upgrade of the Data Center is planned to host resources at least until the end of LHC Run 3 (2023), we are testing (elastic) extensions of our Data center to other sites, including commercial clouds, since in this way it could be possible to address other use cases such as temporary peak requests in a cheaper way than with flat provisioning. Besides some small scale tests on commercial providers, we have performed a scalability test integrating into CNAF farm $\sim$ 20k HS06 provided by Bari-ReCaS. These resources are currently being used for LHC jobs only. Results obtained are promising even if the current infrastructure penalizes I/O intensive jobs running in Bari-ReCaS. An enhanced and more performant cache system is expected to replace the current one; network connection enhancement will be also considered.

References

[1] The Cherenkov Telescope Array URL `https://www.cta-observatory.org/`
[2] Bari ReCaS Data Center URL `https://www.recas-bari.it/index.php/en/`
[3] Gruppo per l'Armonizzazione delle Reti della Ricerca URL `http://www.garr.it/`
[4] HEPiX Benchmarking Working Group HEP-SPEC06 Benchmark URL `http://w3.hepix.org/benchmarks/doku.php`
[5] Etsion Y and Tsafrir D 2005 *School of Computer Science and Engineering, The Hebrew University of Jerusalem* **44221** 2005–13
[6] Thain D, Tannenbaum T and Livny M 2005 *Concurrency and computation: practice and experience* **17** 323–356
[7] Schmuck F B and Haskin R L 2002 GPFS: A Shared-Disk File System for Large Computing Clusters. *FAST* vol 2
[8] Brooks C, McFarlane P, Pott N, Tomaz E and Trcka M 2006 *IBM Redbooks, June*
[9] The Large Hadron Collider Optical Private Network URL `http://lhcopn.web.cern.ch/lhcopn/`
[10] The Large Hadron Collider Open Network Environment URL `http://lhcone.web.cern.ch/`
[11] Ricci P P, Bonacorsi D, Cavalli A, dell'Agnello L, Gregori D, Prosperini A, Rinaldi L, Sapunenko V and Vagnoni V 2012 The Grid Enabled Mass Storage System (GEMSS): the Storage and Data management system used at the INFN Tier1 at CNAF. *Journal of Physics: Conference Series* vol 396 (IOP Publishing) p 042051
[12] Dal Pra S 2016 Adjusting the fairshare policy to prevent computing power loss "CHEP 2016 proceedings"
[13] Dal Pra S, Ciaschini V, dell'Agnello L, Chierici A, De Girolamo D, Sapunenko V, Boccali T and Italiano A 2016 Elastic CNAF DataCenter extension via opportunistic resources *International Symposium on Grids and Clouds* vol 13
[14] Helix Nebula Science Cloud URL `http://www.helix-nebula.eu/`
[15] The Wigner Data Center URL `http://wigner.mta.hu/wignerdc/index_en.php`
[16] Quintero D, Ceron R, Dhandapani M, da Silva R G, Ghosal A, Hu V, Li H C, Marthi K, Shi S F, Velica S *et al.* 2013 *IBM Technical Computing Clouds* (IBM Redbooks)
[17] Sapunenko V, D'Urso D, dell'Agnello L, Vagnoni V and Duranti M 2015 An integrated solution for remote data access *Journal of Physics: Conference Series* vol 664 (IOP Publishing) p 042047
[18] Carbone A, dell'Agnello L, Forti A, Ghiselli A, Lanciotti E, Magnoni L, Mazzucato M, Santinelli R, Sapunenko V, Vagnoni V *et al.* 2007 Performance studies of the StoRM storage resource manager *e-Science and Grid Computing, IEEE International Conference on* (IEEE) pp 423–430