Journal of Physics: Conference Series **513** (2014) 032090 doi:10.1088/1742-6596/513/3/032090

# The Legnaro-Padova distributed Tier-2: challenges and results

**Simone Badoer[a], Massimo Biasotto[a], Fulvia Costa[b], Alberto Crescente[b], Sergio Fantinel[a], Roberto Ferrari[b], Michele Gulmini[a], Gaetano Maron[a], Michele Michelotto[b], Massimo Sgaravatto[b], Nicola Toniolo[a]**

[a] INFN Laboratori Nazionali di Legnaro, Viale dell'Università 2, I-35020 Legnaro PD, Italy
[b] INFN Padova, Via Marzolo 8, I-35131 Padova, Italy

E-mail: simone.badoer@lnl.infn.it, massimo.biasotto@lnl.infn.it, fulvia.costa@pd.infn.it, alberto.crescente@pd.infn.it, sergio.fantinel@lnl.infn.it, roberto.ferrari@pd.infn.it, michele.gulmini@lnl.infn.it, gaetano.maron@lnl.infn.it, michele.michelotto@pd.infn.it, massimo.sgaravatto@pd.infn.it, nicola.toniolo@lnl.infn.it

**Abstract.** The Legnaro-Padova Tier-2 is a computing facility serving the ALICE and CMS LHC experiments. It also supports other High Energy Physics experiments and other virtual organizations of different disciplines, which can opportunistically harness idle resources if available.

The unique characteristic of this Tier-2 is its topology: the computational resources are spread in two different sites, about 15 km apart: the INFN Legnaro National Laboratories and the INFN Padova unit, connected through a 10 Gbps network link (it will be soon updated to 20 Gbps). Nevertheless these resources are seamlessly integrated and are exposed as a single computing facility. Despite this intrinsic complexity, the Legnaro-Padova Tier-2 ranks among the best Grid sites for what concerns reliability and availability. The Tier-2 comprises about 190 worker nodes, providing about 26000 HS06 in total. Such computing nodes are managed by the LSF local resource management system, and are accessible using a Grid-based interface implemented through multiple CREAM CE front-ends.

dCache, xrootd and Lustre are the storage systems in use at the Tier-2: about 1.5 PB of disk space is available to users in total, through multiple access protocols.

A 10 Gbps network link, planned to be doubled in the next months, connects the Tier-2 to WAN. This link is used for the LHC Open Network Environment (LHCONE) and for other general purpose traffic.

In this paper we discuss about the experiences at the Legnaro-Padova Tier-2: the problems that had to be addressed, the lessons learned, the implementation choices. We also present the tools used for the daily management operations. These include DOCET, a Java-based webtool designed, implemented and maintained at the Legnaro-Padova Tier-2, and deployed also in other sites, such as the LHC Italian T1. DOCET provides an uniform interface to manage all the information about the physical resources of a computing center. It is also used as documentation repository available to the Tier-2 operations team.

Finally we discuss about the foreseen developments of the existing infrastructure. This includes in particular the evolution from a Grid-based resource towards a Cloud-based computing facility.

## 1. Introduction

The history of the Legnaro-Padova Tier-2 goes back to 2001, when it started as a collaboration between INFN Legnaro National Laboratory and INFN Padova to setup a prototype computing farm, located in Legnaro, for CMS MonteCarlo productions. Since then the two sites have always been involved in several Grid related activities and in other computing activities of the LHC experiments, in particular ALICE and CMS.

In 2008 a tighter integration between the two sites has been achieved exploiting a dedicated fiber link connecting them. This allowed the implementation of a distributed Tier-2: the services and resources, since then all located in Legnaro, have been deployed in both sites.

The Legnaro-Padova center is an official Tier-2 for the CMS experiment since 2006, and a Tier-2 for ALICE since 2011. It also supports other experiments and other virtual organizations (VOs) of different disciplines in an opportunistic way.

## 2. Description of the Legnaro-Padova Tier-2

The resources and services of the Tier-2 are distributed in the two sites, as shown in fig. 1.

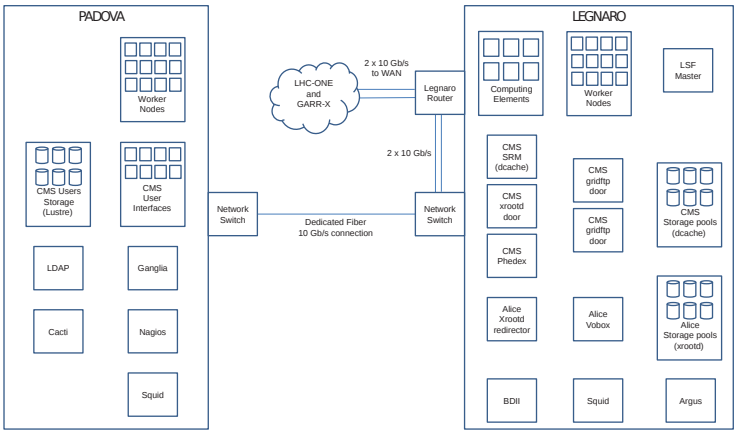


**Figure 1.** Deployment of resources and services in the Legnaro-Padova Tier-2.

There are about 190 worker nodes (WNs), deployed in both sites, providing about 26000 HS06 in total. They are managed by the LSF resource management system, and configured as a single cluster. These WNs haven't been statically partitioned among the supported VOs: instead the relevant pledges on resource usage by the experiments are implemented properly configuring fairshare scheduling at the batch system level. This proved to be very useful to have an efficient overall usage of the resources: the resources assigned to a certain experiment not fully used in a certain period can be used by another VO.

Experiment software is provided through CVMFS for the VOs ready to use it (these include CMS and now also ALICE, for which BitTorrent has been used for a while). CVMFS uses two squid servers (one located in Legnaro and the other one in Padova) in load-balancing configuration and with fail-over in case one of them is offline. These two squids are also used to allow CMS jobs to access the conditions data by means of Frontier. For the other experiments not using yet CVMFS, the software is provided through a file system which is NFS mounted on all WNs.

Each WN has its own local disk which is used, besides for the operating system installation, for the working directories of the running jobs and for the local CVMFS cache.

These WNs are accessible using a Grid interface through 6 CREAM CEs. The choice to deploy so many CE frontends was done in particular to cope with some past instabilities of the

CREAM service (now instead the CREAM CEs are quite reliable). The presence of multiple CEs, besides assuring an high level of scalability, proved to be very useful to properly manage interventions (such as OS or middleware updates) involving the WNs and/or the CEs: these operations can be done gradually without disrupting the operations.

Glexec is deployed since June 2011 on all WNs and already used in production for the jobs of the CMS experiment for which therefore identity switch is fully implemented. Glexec relies on an ARGUS server deployed in Legnaro.

For what concerns the storage, the first adopted solution for the CMS experiment was DPM. Since DPM was not well supported within the CMS experiment, and since at that time the evolution plans were not clear (in particular for what concerns the scalability), in 2007 it was decided to migrate to another system, namely dCache, which is still used. The dCache system for CMS, which provides in total about 1.1 PB, is now implemented by a dCache central host, 16 pool nodes, and 2 nodes where the gridftp doors are deployed. Read access to this storage from the Tier-2 nodes is implemented through the dcap protocol, while write access (from everywhere) is provided through the srm protocol. This storage system is also part of the CMS xrootd federation: this was implemented deploying a xrootd dCache door and a local redirector which "registers" this storage system to the CMS European local redirector located in Bari. The relevant xrootd monitor plugins had also to be installed on all pool nodes.

The same dCache installation was initially chosen as storage solution also for the ALICE experiment. Unfortunately this didn't work properly because the system was not able to properly manage the relevant data: the overall data volume was quite limited but the number of files was huge. Moreover at that time the xrootd support in dCache was not very efficient. These issues also caused problems to the CMS experiment (which was using the same dCache installation). It was therefore decided to implement a specific system for the ALICE data: right now the storage available to the ALICE experiment is implemented by 7 xrootd servers plus a redirector, all located in Legnaro, providing a total of about 390 TB. "Native" xrootd is used without any intermediate layer (i.e. there is not a distributed file system such as gpfs or lustre).

Legnaro also hosts the site BDII and some experiment specific services: Phedex (CMS) and a VOBOX (ALICE).

While physical resources are used for the worker nodes of the Tier-2, most of the services are hosted on virtual machines. In fact, because of the ever growing requirement of dedicated servers for running specific services, it was decided to set up a virtualization solution to allow a more efficient use of the hardware resources. A first virtualization system based on ESXi v4 VMWare virtualization servers was deployed in 2009. In 2011, after an assessment of the available virtualization tools, it was decided to migrate to a ProxMox 2.x based environment composed by three powerful servers with a shared SAS/FC storage backend. This new setup, besides simplifying the overall maintenance of the virtualization infrastructure, offers more advanced features, such as the live migration of VMs.

In a distributed system, such as the Legnaro-Padova Tier-2, the setting of an efficient and reliable networking infrastructure is clearly of paramount importance. As already mentioned, the two sites are connected through a 10 Gbps fiber link, used only for the Tier-2 related activities.

In the first implemented network layout the resources in each site had a local setup: different subnets for public IPs, different subnets for the private network, and each site was linked to the wide area through separated links. This setup revealed to be too complicated from a management point of view. Moreover the two class C private networks used in Legnaro were getting full.

Because of these problems, in 2011, it was decided to harmonize the layout creating a single network zone with a unique class B private subnet to be used for the worker nodes, and a single subnet for public IPs. The connectivity to the LHC-ONE/GARR-X wide area network from all

the Tier-2 resources (including the ones located in Padova) was implemented through a double 10GB/s link using the Legnaro frontier router.

## 3. Integration of the CMS analysis cluster

In Padova a special cluster of 13 User Interfaces has been setup and is available to the local CMS community, in particular for the end-user analysis. All these machines mount a shared storage, implemented by a Lustre distributed file-system, for a total of about 30 TB.

This cluster is seamless integrated with the Legnaro-Padova Tier-2. In particular from these user interfaces direct read access to the whole CMS Tier-2 storage is provided through the native dCache dcap protocol, using the 10Gbps Tier-2 link. This allows running very efficiently analysis tasks that need to process data stored in the Tier-2.

The cluster of user interfaces is managed using the same tools and services used in Tier-2 operations. For example the CMS experiment software is provided through CVMFS, relying on the same set of squid servers used in the Tier-2

## 4. Operations at the Tier2: procedures and tools

The distributed Tier-2 is managed by a team of people also distributed in the two sites. Apart from the operations requiring manual interventions on the physical resources (which are done by the local personnel) every other operations can be done by any person of the team, even though the relevant service is hosted in the other site.

Activities at the Tier-2 are planned and coordinated in weekly phone-conferences while face-to-face meetings take place whenever needed. Besides these formal meetings, the persons involved in the operations at the Tier-2 use to meet together quite often.

The sharing of all the relevant information and the proper management of the documentation are always important points in the operations of a complex system, but this is even more crucial in a distributed infrastructure, such as the Legnaro-Padova Tier-2. Docet [1] is the main tool used for this purpose in the Tier-2. It was designed and developed initially for the activities related to the Babar Tier-A located in Padova, and it is now heavily used in the context of the Tier-2 activities, also due to the fact that its designer and implementor is a member of the Tier-2 operations team. Docet is used first of all to manage all the information about the physical resources (computing nodes, storage servers, network devices, etc.). It is also used as a logbook for the performed activities, and to gather documentation and notes that can help everyone involved in the center operations.

An instance of Docet is currently deployed and used also at the INFN Tier-1.

A special focus was given to the deployment of an efficient monitoring infrastructure: the early detection of problems or performance degradations are indeed very important points to guarantee a high level of reliability of the infrastructure. This was implemented using several tools.

Ganglia is one of them: it is used to detect the status and performance of the resources. Cacti is instead used to monitor all the network switches and appliances. Nagios is heavily used to check the health of the overall infrastructure, and to notify the Tier-2 operations personnel in case of problems. It is used not only to monitor the physical resources but also to control all the services, such as the batch system, the middleware and also the experiment specific services. Nagios is configured to collect information also from external views, in particular the CMS and ALICE specific monitoring and testing systems. Specific custom NAGIOS sensors were developed if not available already. Basically every time that a new problem in the operations is found, a specific new NAGIOS sensor is implemented to prevent the problem in the future or at least to early detect it. The Ganglia, Cacti and Nagios services are hosted in Padova.

Besides these tools, several other custom scripts have been developed In particular tools have been implemented to detect problems on the worker nodes (swapping, disk space being filled, etc.): corrective actions, such as killing problematic jobs, are taken whenever possible, otherwise the relevant WNs are simply closed to prevent black-hole behaviors.

The cooling and power infrastructure is monitored by an application that has been locally implemented. It allows the monitoring of the chillers, of the rack coolers, the power distribution lines and the UPS. This monitoring system consists of a backend server collecting the relevant using the OPC protocol, while LabView is used in the frontend, for the graphic view and for alert notifications.

The operations of the LSF batch system are currently monitored by a tool called *LSFMon* [2]. This allows to check the current status or the past behavior of the system, monitoring several items such as the number of jobs per virtual organization, their efficiency, etc. Since this tool is not maintained anymore and since it lacks some functionality, it was decided to develop a new LSF monitoring system whose implementation is being finalized. Besides providing all the functionality of the old system, the new one allows also to have per-job information, while the old system can manage at most VO aggregated information.

## 5. Conclusions and future activities

In this paper we described the activities at the Legnaro-Padova Tier2, whose unique characteristic is that it is spread in two different sites, but exposes itself as a single computing facility. We discussed also about how this distributed infrastructure is managed by a team of people, also located in the two sites.

Despite the intristic complexity of its distributed architecture, the Legnaro-Padova Tier-2 proved to be very reliable and quite performant. For example in the last two years the average WLCG measured reliability and availability [3] were respectively 99.71 % and 99.21 %.

In the last year the Legnaro-Padova Tier-2 run about 49 % of the jobs executed in the four ALICE Italian Tier-2 centers, and about 30 % of the CMS jobs executed in the four Italian CMS Tier-2s.

For what concerns the future, the focus will be likely on the evolution from the Grid paradigm towards Cloud computing.

Since the Tier-2 proved to be very performant and reliable, a distributed private Cloud facility, including resources in Padova and in Legnaro, available to the users of these two sites, is being set up. Besides providing an interactive computing on demand service, the idea is to use the cloud paradigm also for the batch-like activities of the Tier-2. Some work has already been done in particular for the CMS experiment, which is exploring the possibility to use Cloud resources in addition to the Grid. At the Padova-Legnaro Tier2 an OpenStack Cloud based testbed has been set up, and here the execution of CMS CRAB analysis jobs has been successfully demonstrated.

## Acknowledgments

The work presented in this paper has been partially funded under contract 20108T4XTM of "Programmi di Ricerca Scientifica".

## References

[1] S. Dal Pra and A. Crescente, *The data operation centre tool. Architecture and population strategies*, 2012 J. Phys.: Conf. Ser. 396 042014 doi:10.1088/1742-6596/396/4/042014
[2] LSF Job Summary and Accounting Monitor project Website, http://sarkar.web.cern.ch/sarkar/doc/lsfmon.html
[3] WLCG Tier-2 Availability and Reliability Reports, http://sam-reports.web.cern.ch/sam-reports