The Legnaro-Padova distributed Tier-2 challenges and results

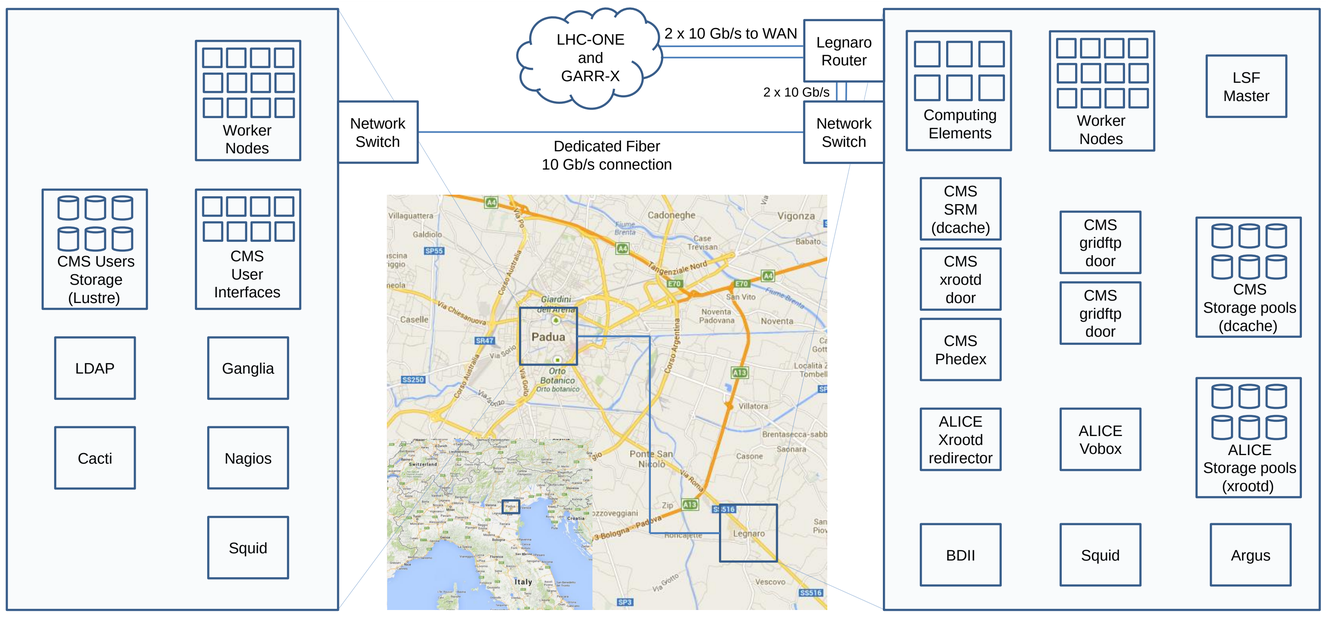
S. Badoer, M. Biasotto, F. Costa, A. Crescente, S. Fantinel, R. Ferrari, M. Gulmini, G. Maron, M. Michelotto, M. Sgaravatto, N. Toniolo

Introduction

The Legnaro-Padova Tier-2 is a computing facility serving in particular the ALICE and CMS LHC experiments. Its unique characteristic 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. Nevertheless these resources are seamlessly integrated and are exposed as a single computing facility.
The history of 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 of the two INFN units has been achieved exploiting a dedicated fiber link connecting the two sites, implementing therefore a distributed Tier-2: the services and resources, since then all located in Legnaro, have been deployed in both sites.

Distributed Tier-2 setup

Computing room in Padova



Computing room in Legnaro

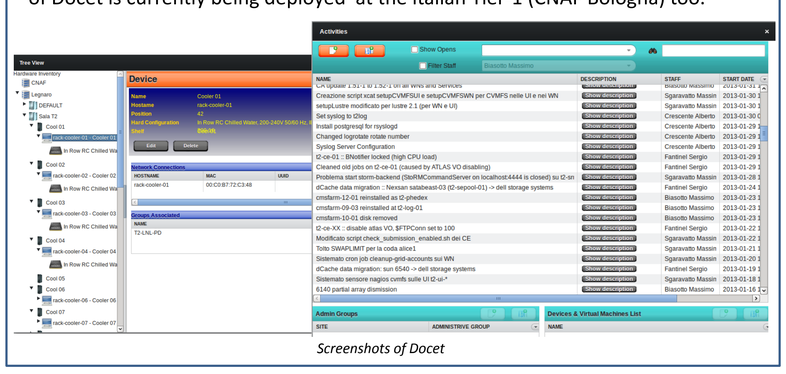
Tier-2 Resources	
Worker Nodes	190
Cores	2400
Computing Power	26000 HS06
Disk	1.5 PB

The Tier-2 resources and services are distributed between the two sites as represented in the image above. The deployment is not symmetric because the storage and the other critical services are all concentrated in one site only (Legnaro) in order to avoid exposing them to the downtimes of two sites for maintenance and failures. The bulk of servers is composed of 190 Worker Nodes which are splitted about equally between the sites and configured as a single cluster managed by LSF batch system. These computing resources are accessible using a Grid interface through as many as 6 Cream CEs providing a high level of scalability and reliability. The storage for the two main VOs is not shared but each one has its own dedicated storage system: dCache for CMS with 1.1 PB on 16 disk servers and native xrootd for ALICE with 390 TB on 7 servers. Along with the storage there are the services providing remote data access through different protocols (srm, gridftp doors and xrootd redirector) and the CMS dataset transfer agents (Phedex). One of the critical aspects in setting up the distributed Tier-2 was the networking. In its current configuration all the nodes in both sites are in the same class B private network and VLAN which are propagated across the dedicated link. All the outbound traffic goes through the Legnaro router which is connected at 2x10Gb/s to LHC-ONE and GARR-X networks.

Site administration tools and solutions adopted

Docet

Docet is the main tool used for the logging and synchronization of activities among the people working on the Tier-2 administration. It was designed and developed by Alberto Crescente, a member of the Legnaro-Padova Tier-2 staff and is the result of the strong know-how collected after many years of experience in large farm management by Alberto and his colleagues. It helps in tracking the activities, collecting all information about the devices (servers, storage, switches, etc.) and their location in the computing room, in gathering documentations, notes and "tips & tricks" covering different aspects that can help everyone involved in the center operations to maintain the history of device failures and so on. An instance of Docet is currently being deployed at the italian Tier-1 (CNAF Bologna) too.



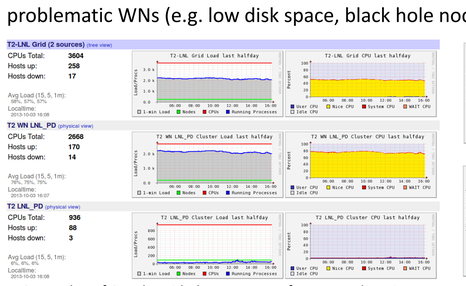
Screenshots of Docet

Monitoring tools

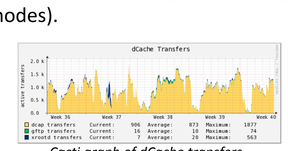
The core of the Tier-2 monitoring is based on three well known tools, all with customized scripts and configurations

- Ganglia is the main source providing the status and performance of the Tier-2 hardware resources, which are divided in two groups with different frequencies of metric collection: high (~30 sec.) for the storage servers and all services, and lower (~5 min.) for less critical machines like the worker nodes.
- Nagios collects all the site health information not only of internal resources but also from external views (SAM tests, CMS and ALICE specific monitors, etc.)
- Cacti is used to monitor all the network switches and appliances.

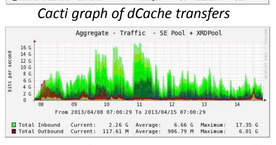
In addition to the monitoring there are several scripts and cron jobs developed to take corrective actions for the more common issues, for example automatically killing memory consuming jobs when a server starts swapping or closing problematic WNs (e.g. low disk space, black hole nodes).



Screenshot of Ganglia with the two groups for WNs and services



Cacti graph of dCache transfers

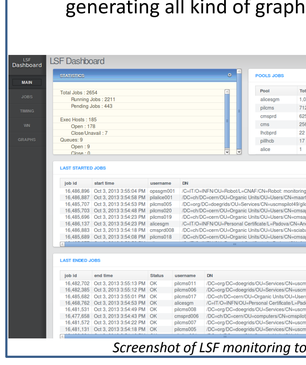


Cacti graph of aggregated disk-pools traffic

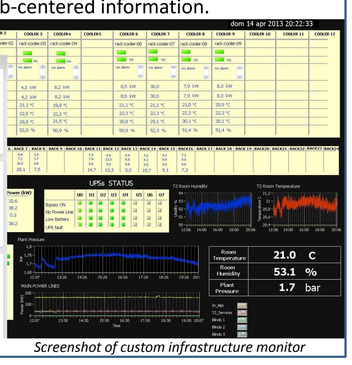
Custom monitoring tools

In addition to the standard tools commonly used for farm monitoring, a couple have been locally developed to meet specific needs.

- **Cooling and Power Infrastructure Monitor**: a custom application to monitor chillers, rack coolers, power distribution lines and UPS. Composed by a backend server collecting the data via OPC protocol and a LabView frontend for graphic presentation and alert notifications.
- **LSF Monitor**: a new software currently being developed by A. Crescente (author of Docet) to replace an old unmaintained tool used so far (LSFMon). It collects from the LSF batch system all available data for all jobs (even the past ones), stores them in a Postgres DB and offers a web graphical interface for generating all kind of graphs, views and job-centered information.



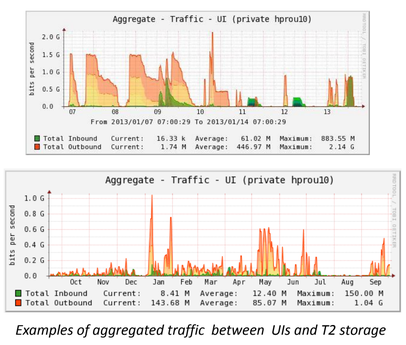
Screenshot of LSF monitoring tool



Screenshot of custom infrastructure monitor

CMS UI Cluster

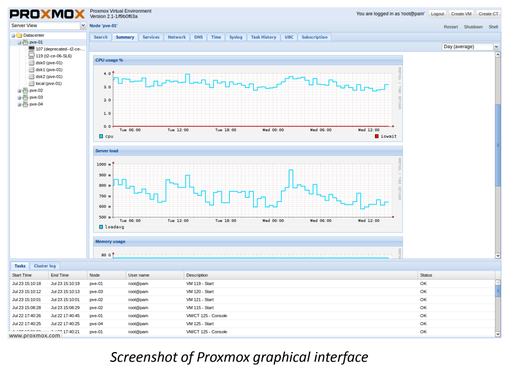
In Padova a special cluster of User Interfaces has been setup in order to provide an efficient system for the analisys activities of the local CMS community. It is currently composed by 13 servers which all mount a shared storage organized in home directories, a large working area and some scratch space. This UI storage is a mix of old and new disk systems, all aggregated in a Lustre distributed file-system, for a total of about 30TB.
From these UIs the physicists have direct read access to the whole CMS T2 storage through the native dcache protocol so from the same machine they can run directly and very efficiently the smaller analisys tasks or submit the larger ones to the Grid.



Examples of aggregated traffic between UIs and T2 storage

Virtualization Infrastructure

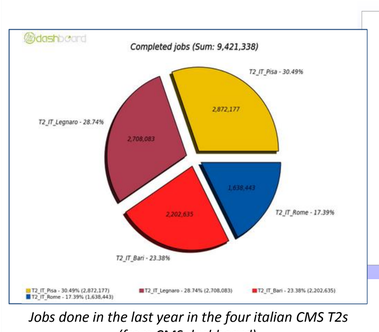
The ever growing requirement of dedicated servers for running specific services has motivated the adoption of a virtualization solution allowing a more efficient use of the hardware resources. A first setup in 2009 was based on a number of independent ESXi v4 VMWare virtualization servers, which accomplished its task for about two years even with the limitations of the “free” VMware version. In 2011, after a new evaluation of the available tools, it was decided to migrate to a ProxMox 2.x environment composed by three powerful servers with a shared SAS/FC storage backend. This new setup offers more advanced features like live migration of VMs and an easier maintenance of the virtualization infrastructure.



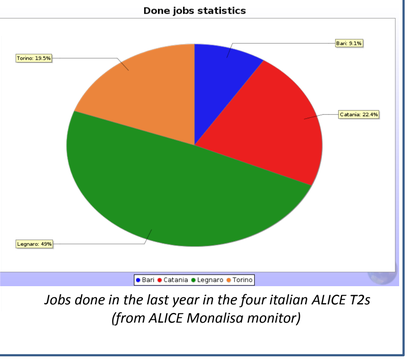
Screenshot of Proxmox graphical interface

Site performance and results

Despite the intrinsic complexity of its distributed architecture, the Legnaro-Padova Tier-2 proved to be very reliable: it has always been among the top sites in the availability ranking measured by CMS and ALICE and in the official WLCG “Availability and Reliability Report” the site averages for the last two years were respectively 99.7% and 99.2%.
The site contribution to the computing activities of the two VOs is usually larger than their quota of resources thanks to the dynamic sharing of WNs which allows the exploitation of extra slots whenever they are left unused by the other VO (this was especially true for ALICE in the last year, as can be seen in the plots below).



Jobs done in the last year in the four italian CMS T2s (from CMS dashboard)



Jobs done in the last year in the four italian ALICE T2s (from ALICE Monalisa monitor)