Optimization of Italian CMS Computing Centers via MIUR funded Research Projects

Optimization of Italian CMS Computing Centers via MIUR funded Research Projects

T.Boccali[1], G.Donvito[2], A.Pompili[2], G.Della Ricca[3], E.Mazzoni[1], S.Argiro'[4], C.Grandi[5], D.Bonacorsi[6], L.Lista[12],   F.Fabozzi[7],L.M.Barone[8], I.Talamo[3], A.Santocchia[9], H. Riahi[14], A.Tricomi[10], M.Sgaravatto[13], G.Maron[11]

[1] INFN Sezione di Pisa
[2] Politecnico di Bari
[3] Universita' di Trieste
[4] Universita' di Torino
[5] INFN Sezione di Bologna
[6] Universita' di Bologna
[7] Universita' di Napoli Federico II
[8] Universita' di Roma La Sapienza
[9] Universita' di Perugia
[10] Universita' di Catania
[11] INFN Laboratori di Legnaro
[12] INFN Sezione di Napoli
[13] INFN Sezione di Padova
[14] INFN Sezione di Perugia

Corrisponding author: Tommaso.Boccali@cern.ch

Abstract. In 2012, 14 Italian Institutions participating LHC Experiments (10 in CMS) have won a grant from the Italian Ministry of Research (MIUR), to optimize Analysis activities and in general the Tier2/Tier3 infrastructure. A large range of activities is actively carried on: they cover data distribution over WAN, dynamic provisioning for both scheduled and interactive processing, design and development of tools for distributed data analysis, and tests on the porting of CMS software stack to new highly performing / low power architectures.

1. Introduction
The Italian Institute for Nuclear Physics (INFN) finances the participation of Italian institutions to the four LHC experiments, from research & development to maintenance. A sizeable part of the budget is currently spent on the Distributed Computing Infrastructure, consisting in one Tier1 Center, 11 Tier2 centers and an handful of Tier3 centers. INFN contributions are used to meet the pledges Italy has agreed upon with the experiments, and scarce resources are left nowadays for activities like research and development of new computing solutions.

In a resource constrained environment, optimization of the computing architecture is the key to a better use of our resources, which directly translates into better or more physics results.

For this reason, 14 institutions participating the LHC Computing have submitted and won in 2012 a 3 years grant from the Italian Ministry if Research (MIUR); out of them, 11 are part of the CMS Experiment[1]. The complete list includes: Universita' di Torino, Universita' di Trieste, INFN Laboratori di Legnaro, Universita' di Bologna, INFN Sezione di Pisa, Universita' di Perugia, Universita' di Roma La Sapienza, Universita' di Napoli, Politecnico di Bari, Universita' di Catania.

We describe in this paper the development lines that are actively pursued, with the expected results.

2. Towards an Italian Xrootd Federation

The Computing infrastructure CMS has in Italy is based on a large multi-experiment Tier1 (at CNAF, Bologna), 4 large Tier2s (Bari, Legnaro, Pisa, Roma Sapienza) and a number of small Tier3s, mostly financed outside INFN budget (Trieste, Torino, Bologna, Perugia, Catania, Napoli). The physics interests of Italian physicists, combined with the size of out Tier2s, makes feasible the placing of all interesting recent data and Monte Carlo samples on their storage; the Tier1 is also able to use part of its disk resources for this, especially after the recent separation from the tape system. It is thus highly probable that popular file/dataset requests can be satisfied within the Italian resources. If we complement this with the very high quality networking we are given by our NREN (GARR[2]), with 40 Gbit/s T1 and 10-20 Gbit/s for Tier2s and most Tier3s, our infrastructure seems ideal to allow for an Italy-wide data streaming solution. Currently, CMS has decided to use Xrootd[3] as the protocol for geographical data access. Italy is hosting two major Xrootd redirectors in Bari: one serves the European institutions in CMS, and fallbacks to the CERN global redirector (see Figure 1). The second instead is supposed to serve only Italian institutions, thus allowing us to insert in the federation also sites which we do not want to expose outside national boundaries. Our Tier1, CNAF, is active part of the federation, and its storage is served via a special Xrootd configuration: in order not to expose its tape backend to Xrootd, a specific plugin was prepared, tested and put into production to allow masking of GPFS[4] files only residing on the tape backend. The plugin has then been re-engineered by Xrootd developers, and it is now in the official distribution.
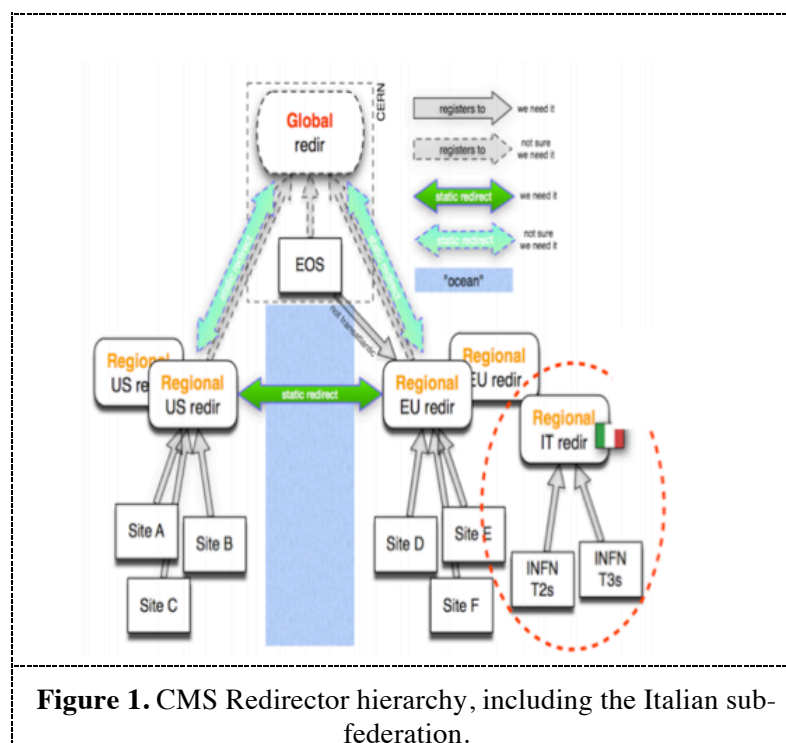

Figure 1. CMS Redirector hierarchy, including the Italian sub-federation.

3. Optimization of Computing Centers for Interactive Analysis

While GRID enabled access to the resources is well established in our sites, the final step of physics analyses is less specified in the CMS computing Model. The activities which are under study are:

- **"User Interface on demand"** via LSF[5]/PBS[6] sharing with Worker Nodes, to allow for a variable number of interactive machines depending on the request. This increases resource usage, since we can avoid to reserve a large number of User Interfaces, to stay mostly idle, and can use them as Worker Nodes for most of the time.
- **Italy-wide login on all User Interfaces**: this has been implemented via AAI (Authentication/ Authentication INFN system)[7], and is currently tested on a few sites. Every Italian user, registered centrally (at the INFN Administration) as a CMS member, can login on a selected number of User Interfaces without any direct interaction with the local site.
- **PROOF[8] deployment**: either on large (64 core) machines, or on the existing GRID clusters. Tests with Proof on Demand are being evaluated.
- **Xrootd caching servers** at the frontiers of small analysis centers: in centers with small storage systems, pre-allocating large data samples is unpractical, and Xrootd access is preferred. On the other hand, the final analysis step is often repeated many times, and a Geographical Xrootd access cannot be optimal. The solution we implemented is based on Xrootd caching servers: in these sites, the whole Xrootd Federation is faked as a "tape backend" to the local storage: if a file is not found locally, it is "staged in" via the Federation, and made to reside locally. Subsequent accesses will be local. Xrootd also takes care of purging the local storage when full, eliminating older files.

4. Tools for Distributed Analysis

Italy is committed to the design, development and integration of the next generation tool for CMS Distributed Analysis tool, CRAB3[9]. For the distributed analysis tool sustainability, CRAB3 is integrated with the distributed Analysis tool of ATLAS, PanDA[10], into a Common Analysis Framework. The efforts are undertaken by CMS, ATLAS and the CERN/IT department. Italy is responsible for design and development of main components in the framework:

- **TaskManager backend**: needed to provide into PanDA the concept of Analysis Tasks.
- **Users Data Management system (AsyncStageOut)**: it manages and monitors the transfer and publication of CMS analysis jobs outputs.
- **CMS component in PanDA**: it handles the CMS jobs metadata to allow the later transfer and publication of the outputs, and also to create analysis reports to end-users.

Italy has ensured also crucial activities for the integration of the framework:

- **AsyncStageOut scale tests**: during the commissioning of the framework, scale test of the AsyncStageOut independently of underlying services has been performed.
- **Alpha and Beta testing**.

In the future, we plan to add more commonalities to the framework for CMS and ATLAS, such as the user data management, the analysis job splitting or the framework deployment. Actually, the AsyncStageOut can interact only with the Workload management tools such as PanDA. The AsyncStageOut will be exposed also to users for the management of their files.

5. Dynamic Provisioning on GRID and Cloud

In the job submission framework of the CMS experiment, resource provisioning is separate from resource scheduling. This is implemented by **pilot jobs**, which are submitted to the available Grid sites to create an overlay batch system where user jobs are eventually executed. CMS is now exploring the possibility to use Cloud resources besides the GRID, basically considering the same architecture for what concerns the dynamic resource provisioning.

The submission workflows, in case of GRID and Cloud usage, are shown in Figure 3.

In the Grid scenario, the Glidein factory is the component responsible to submit, through Condor-G, pilot jobs (called "glideins") to the available Grid sites. Such pilot jobs are responsible to install and configure the allocated slot as an executing node of the overlay batch system (HTCondor[11] is used): the new worker node therefore joins the HTCondor pool, and can run user jobs.

When there are no more jobs to be executed (or when the site claims the resource) the execution of the glidein finishes and the worker node leaves the HTCondor pool.

In the Cloud scenario the very same approach is used: the only difference is that the Glidein factory, instead of submitting pilot jobs using the Grid interface, creates on demand Virtual Machines, which on boot starts the glideins. The VMs instantation is performed by the Condor-G component of the GlideinWMS[12] service, using the EC2 interface available on most Cloud implementations.

At the Padova-Legnaro Tier2 a OpenStack[13] Cloud based testbed has been set up, and here the model has been successfully demonstrated executing CMS CRAB analysis.
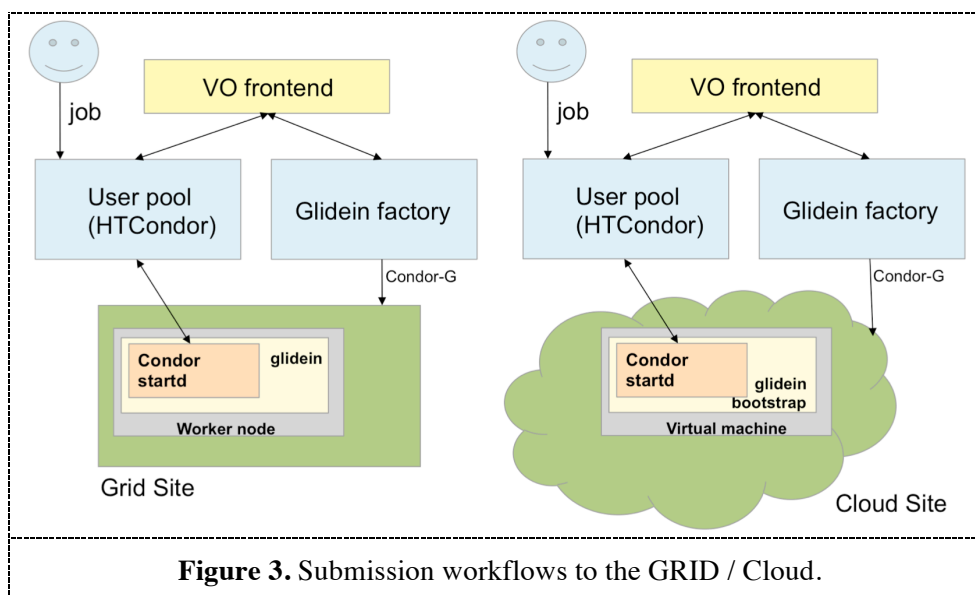

Figure 3. Submission workflows to the GRID / Cloud.

6. Tests of new Computing technologies

Italy are investing manpower and resources in the test of technologies which may become relevant to CMS Computing in the longer run. One strength of the CMSSW Software Stack is its complete independence from any proprietary code.

A library / algorithm / tool, to enter the stack, must allow complete code distribution and patching.In this way, we can recompile the full stack on virtually any POSIX platform with a c++ compiler, and even more easily on platforms which support g++.

We are currently performing benchmarking and porting activities on

- Xeon MIC (previously known as Phi)[14]: 3 machines are available in Pisa, Bologna, and are being used in initial tests of the architecture. At the moment, benchmarking under ROOT is

the main activity, but next tests will be about the offloading to the MIC of actual fragments of CMSSW.

- ARM architectures: the ARMv7 architecture, available now, is interesting as a testbed for future power-efficient data centers. We are currently testing
  - Single "consumer level" boards like the HardKernel Odroid-U2 (a naked Samsung Galaxy S3)[15];
  - Server-grade ARM cluster-in-a-box (Dell Copper[16]).
- We are waiting to get hold on the first 64 bit ARMv8 chips, which will be the first viable ARM solutions for us.

7. Acknowledgements

8. Bibliography

[1] The CMS Collaboration, http://cms.web.cern.ch/content/cms-collaboration

[2] GARR, http://www.garr.it/b/eng

[3] The Xrootd project, http://xrootd.slac.stanford.edu/

[4] The General Parallel File System, IBM, http://www-03.ibm.com/systems/software/gpfs/

[5] IBM Platform LSF, http://www-03.ibm.com/systems/technicalcomputing/platformcomputing/

[6] Portable Batch System, http://en.wikipedia.org/wiki/Portable_Batch_System

[7] INFN-AAI, http://web.infn.it/aai/

[8] The Parallel ROOT Facility, http://root.cern.ch/drupal/content/proof

[9] CRAB3: Establishing a new generation of services for distributed analysis at CMS, M Cinquilli *et al* 2012 *J. Phys.: Conf. Ser.* **396** 032026 doi:10.1088/1742-6596/396/3/032026

[10] PanDA: distributed production and distributed analysis system for ATLAS, T Maeno 2008 *J. Phys.: Conf. Ser.* **119** 062036 doi:10.1088/1742-6596/119/6/062036

[11] Thain D, Tannenbaum T and Livny M 2005 *Concurrency and Computation: Practice and Experience* 17 2-4 323-356 doi:10.1002/cpe.938

[12] Sfiligoi I, Bradley D C, Holzman B, Mhashilkar P, Padhi S and Würthwein F 2009 Comp. Sci. and Info. Eng., 2009 WRI World Cong. on 2 428-432 doi:10.1109/CSIE.2009.950

[13] OpenStack Cloud Software, http://www.openstack.org/

[14] Intel Many Core Architecture, http://www.intel.com/content/www/us/en/architecture-and-technology/many-integrated-core/intel-many-integrated-core-architecture.html

[15] HardKernel Odroid-U2, http://www.hardkernel.com

[16] Dell Copper ARM Architecture, http://www.dell.com/learn/us/en/555/campaigns/project-copper?c=us&l=en&s=biz