Journal of Physics: Conference Series **664** (2015) 032006 doi:10.1088/1742-6596/664/3/032006

Improvements of LHC data analysis techniques at Italian WLCG sites. Case-study of the transfer of this technology to other research areas

L Alunni Solestizi[1], S Argiro'[2], S Bagnasco[3], D Barberis[4], L M. Barone[5], T Boccali[6], D Bonacorsi[7], V Candelise[8], G Carlino[9], E Casula[10], D Ciangottini[1], A De Salvo[11], G Della Ricca[8], G Donvito[12], A Doria[9], R Di Nardo[13], D Elia[12], F Fabozzi[14], A Favareto[4], C Grandi[15], L Lista[9], G Luparello[8], G Maron[16], E Mazzoni[6], L Merola[14], G Miniello[17], D Mura[10], M Perez Villaplana[18], S Piano[19], A Pompili[17], D Rebatto[20], A Santocchia[1], M Sgaravatto[21], I Talamo[5], A Tricomi[22], S Vallero[2], M Venaruzzo[16], E Vilucchi[13]

[1] Università di Perugia and INFN Sezione di Perugia
[2] Università di Torino and INFN Sezione di Torino
[3] INFN Sezione di Torino
[4] Università di Genova and INFN Sezione di Genova
[5] Università di Roma La Sapienza and INFN Sezione di Roma I
[6] INFN Sezione di Pisa
[7] Università di Bologna and INFN Sezione di Bologna
[8] Università di Trieste and INFN Sezione di Trieste
[9] INFN Sezione di Napoli
[10] Università di Cagliari and INFN Sezione di Cagliari
[11] INFN Sezione di Roma I
[12] INFN Sezione di Bari
[13] INFN Laboratori Nazionali di Frascati
[14] Università di Napoli Federico II and INFN Sezione di Napoli
[15] INFN Sezione di Bologna
[16] INFN Laboratori di Legnaro
[17] Università di Bari and INFN Sezione di Bari
[18] Università degli studi di Milano and INFN Sezione di Milano
[19] INFN Sezione di Trieste
[20] INFN Sezione di Milano
[21] INFN Sezione di Padova
[22] Università di Catania and INFN Sezione di Catania

Abstract. In 2012, 14 Italian institutions participating in LHC Experiments won a grant from the Italian Ministry of Research (MIUR), with the aim of optimising analysis activities, and in general the Tier2/Tier3 infrastructure. We report on the activities being researched upon, on the considerable improvement in the ease of access to resources by physicists, also those with no specific computing interests. We focused on items like distributed storage federations, access to batch-like facilities, provisioning of user interfaces on demand and cloud systems. R&D on next-generation databases, distributed analysis interfaces, and new computing architectures was also carried on. The project, ending in the first months of 2016, will produce a white paper with recommendations on best practices for data-analysis support by computing centers.

1. Introduction

The Italian Institute for Nuclear Physics (INFN) funds the participation of Italian institutions to the four LHC experiments, from research & development to maintenance. A sizable part of the budget is currently spent on the Distributed Computing Infrastructure, consisting in one Tier1 Center, 11 Tier2 centers and more than 10 Tier3 centers. INFN contributions are used to meet the pledges Italy has agreed upon with the experiments, while scarce resources are left 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 of Research (MIUR). 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, Universita' di Genova, Universita' di Milano Statale, Universita' di Cagliari. We describe in this paper the development lines that are actively pursued, with the first results already put into production for the LHC Run 2 (2015-2018).

2. Geographical Data Access

Historically, LHC Experiments' Computing Models were designed following MONARC [1] recommendations, as a Distributed GRID infrastructure where jobs are sent where data have been previously placed. For the new LHC Run, thanks to the much improved general purpose connectivity, the LHC Experiments are moving towards a model in which at least a fraction of the accesses to data is performed over the WAN. A concept common to all the models is that of a data federation, where single storage servers can connect and disconnect elastically from the federation, without any central database keeping track of actual file locations.

a. ALICE

The Italian computing infrastructure for the ALICE experiment at the LHC is mainly based on a large national center in Bologna (CNAF) acting as Tier-1 site and four Tier-2 centers located in Bari, Catania, Padova-Legnaro and Torino, respectively. Whereas in the original schema each tier was assigned a different role and purpose, in ALICE such distinctions have been fuzzy since the very beginning and sites have been assigned to Tier-1 or Tier-2 essentially according to their size and the availability of custodial storage for a second collective copy of the full raw data sample. Smaller centers in Trieste and Cagliari also contribute as additional ALICE sites in the WLCG. Connectivity between 10 and 20 Gbit/s (Tier-2) and 40 Gbit/s (Tier-1) is guaranteed by the Italian Reserarch & Education Network provided by the GARR Consortium.

b. ATLAS

All the Italian Tiers of ATLAS (Frascati, Milano, Napoli and Roma) and the Italian T1 (CNAF) have been included in the Overlay Network called LHCONE [2]. The Atlas community has been involved in the definition and the refinements of the deployment of LHCONE, since the beginning. The activity of the Italian community was very important for the deployment of the monitoring systems, called PerfSonar PS, the extensions with the Software Defined Networks and the OpenFlow protocol and the integration of the Italian sites in the overall infrastructure.

The studies on LHCONE have been also the starting point for a new activity on a prototype of a distributed T2, that we'll discuss later.

c. CMS

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 the Tier2s, makes it feasible to place 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 2013 separation from the tape system. Italy already hosting 2 copies of the European Regional Redirector for the CMS Xrootd Federation, has implemented an Italian sub-federation. The idea is that, when a remote file access is requested by a processing task running in Italy, data sources located in Italy are preferentially chosen over regional ones (Europe in this case). In this way, Italy has put in place an highly connected environment, with interconnections between 10 and 40 Gbit/s as guaranteed by the national NREN (GARR), where analysis activities can be seamlessly carried on using a global space of the order of 10 PB, sufficient to host the majority of interesting analysis data files.

3. Distributed Computing Tools

a. ALICE

Within WLCG, computing activities and particularly the access to the data are carried out through batch programs. This kind of data processing and access is fully adequate for simulation activities, data reconstruction and several analysis tasks. Running on the computational GRID has been proven to be a fair option when relatively long lasting jobs are involved (several hours). However, such batch-based management is not the optimal solution for the final stages of analysis, typically operated by small groups of physicists, or for other use cases like optimization of algorithms, code debugging, fast data quality monitoring where the need for a fast and interactive access to the data is relevant. For these applications, in fact, the latency time between analysis job submissions and their actual executions on the GRID is comparable with the time of execution. Moreover, the analysis of large samples of ALICE experiment data (~100 TB) involves the step to merge the partial results, which is dominated by the I/O and file transfer thus lowering the CPU efficiency. To devote computing resources to parallel and interactive analysis, a standard deployment of analysis facilities based on PROOF (Parallel ROOT Facility) [3] has been defined within the ALICE Collaboration and successfully applied in several sites. The Italian community, in particular, is involved in the deployment and federation of elastic farms, where resources expand and reduce automatically, depending on the load, by varying the number of running virtual machines. The first Virtual Analysis Facility (VAF) has been deployed in Torino a few years ago and similar but smaller cloud-based test infrastructures have been recently setup in Bari, Padova-Legnaro and Trieste, where various connected activities are currently ongoing as schematically illustrated in Fig. 1.

The main developments are concentrated on the following items: benchmarking of the VAF performances, monitoring and data access/federation issues. To create a reliable benchmark, an ALICE analysis macro was extended in order to return the Wall and CPU Clock Time (WCT and CCT) of the different phases of the instance life and of the analysis steps. The main results obtained in Torino site are summarized in Fig. 2, where the different contributions to the total deploy time of the PROOF analysis is measured with increasing number of the workers: the init and connect parts are negligible with respect to the true analysis and an optimal number of 30 workers is suggested. Comments on the monitoring and data access/federation are included in the next section, while further details on the overall activity can be found in [4].

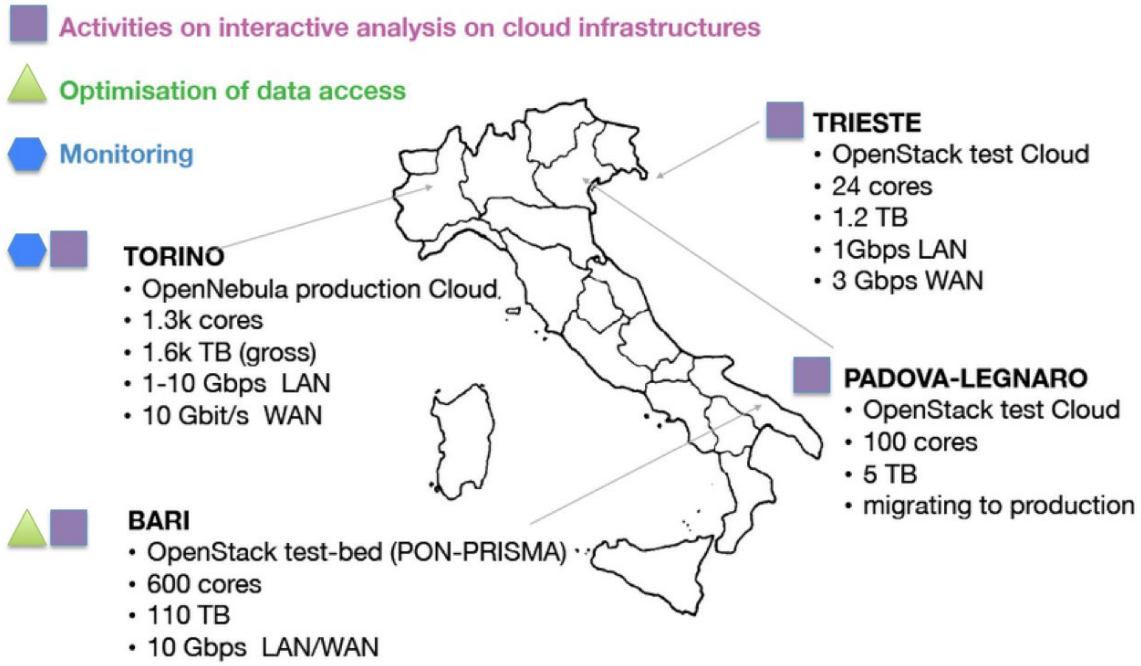

Fig. 1: Actual distribution of the main activities and infrastructures for the Italian VAF sites.

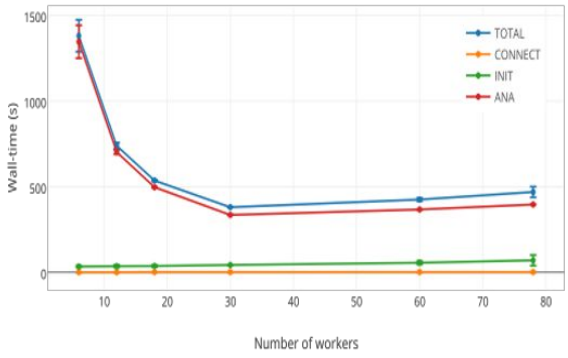

Fig. 2: Total WCT for benchmarking analysis as a function of the number of PROOF workers.

b. ATLAS

The Italian community is involved in the studies of possible improvements of the analysis techniques, via distributed tools and advanced data access technologies. A particular focus has been given to the Proof-On-Demand[5] facilities (PoD). PoD is an extension of the standard Proof facilities, used to give modularity, elasticity and more interactivity to the whole infrastructure.

The Italian community has been studying the possibility to port PoD to the ATLAS Grid infrastructure, since the age of the WMS. The system has been improved and extended more, up to the use of PanDA, the ATLAS production and analysis system. The PoD scheme is shown in Fig. 3. PoD is an easy to use tool, mostly available to all the machines without any need to install it. This task is achieved by exploiting the CVMFS[6] exports of PoD already included in the ATLAS central repository. PoD has been proved to be a versatile and useful tool, being able to use new computing

technologies like the Grid and Cloud ones, and being able to access data via standard filesystems or storage federations.
The data access has been tested in different environments, ranging from plain filesystems to Xrootd federations [7] and Http federations [8]. The data processing performance is scaling very well with the number of nodes dynamically aggregated by the facility. The data access performance is also comparable between storage federations and parallel filesystems, like GPFS.

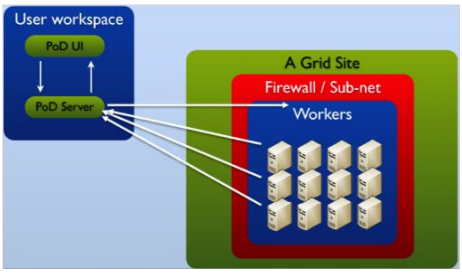

Fig. 3: The Proof-On-Demand scheme, divided in user space and execution space in the sites.

Besides the analysis job execution, the Italian Community is also involved in the DataBase infrastructure improvements, and in particular for what concerns the search of events in order to select only the interesting part of big amount of data. This database has been historically defined as "Tag DB" [9][10][11][12]. With the rise of new technologies, like Map-Reduce and Hadoop, ATLAS has been redesigning its Tag DB, in order to evolve it to a more performing system, called EventIndex [13]. The EventIndex, now operational, is a collection of pointers to the events in the ATLAS datasets. The system uses Hadoop, is integrated with other systems, allowing to enable the processing granularity of a single event, and it's a full featured system, with http(s) and CLI interfaces. The operational scheme of the system is show in Fig. 4.

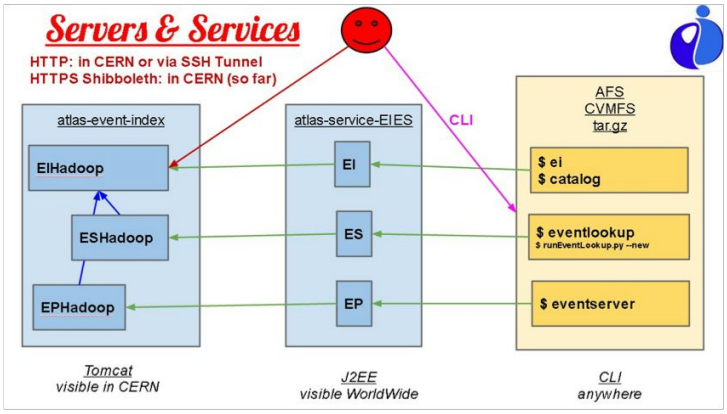

Fig. 4: The ATLAS Event Index scheme.

c. CMS

Italy is committed to the design, development and integration of the next generation tool for CMS Distributed Analysis tool, CRAB3. The new tool, the standard for LHC Run2, consists in a thin client and a server, which dispatches jobs via a glideInWMS [14]. Italy is involved in CRAB3 development, and for its transition to production use. Recent tests have shown CRAB3 scalability up to 200k jobs/day (with more than 20k running jobs). These numbers are expected to at least double by the start of Run2. Another component where Italy is investing a lot of effort is the AsyncStageOut, which

handles the transfer of analysis output to the submitter's site. During the same tests, the tool has been able to achieve up to 300k transfers per day. Fig. 5 shows the outcome of the these tests.

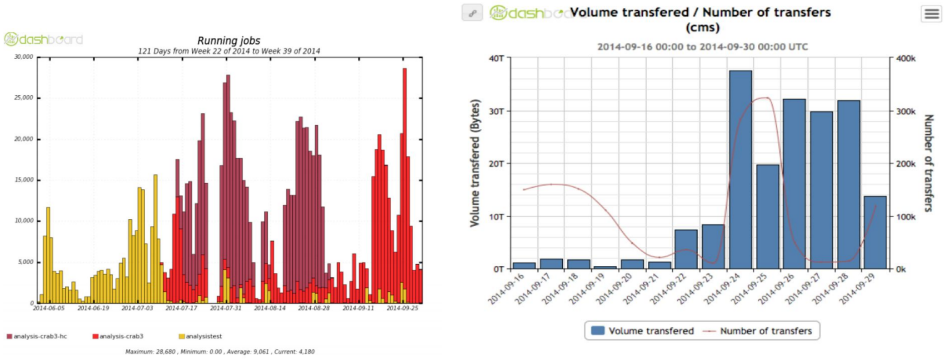

Fig. 5: Number of simultaneous CRAB3 running jobs (left); Number of AsyncStageOut transfers, and relative volume (right).

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.

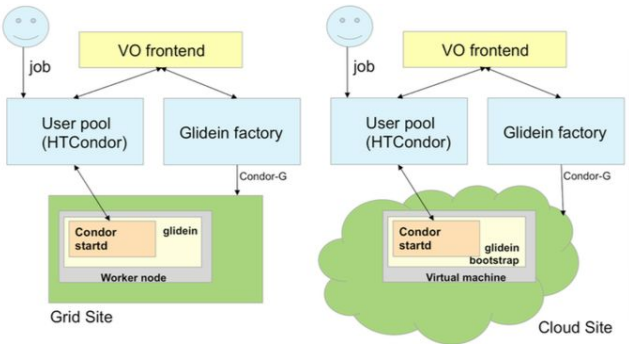

Fig. 6: CMS submission strategy to GRIDs and Clouds.

The submission workflows, in case of GRID and Cloud usage, are shown in Figure 6. 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[15]): 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 start the glideins. The VM instantiation is performed by the Condor-G component of the glideinWMS service, using the EC2 interface available on most Cloud implementations.
At the Padova-Legnaro Tier2 a OpenStack[16] Cloud based testbed has been set up, and here the model has been successfully demonstrated executing CMS CRAB analysis.

4. Computing Centers' Optimization

a. ALICE

Among the activities carried out for the deveopment of the interoperating VAF infrastructures, the development of a monitoring, accounting and billing infrastructure able to consolidate data from all levels of the stack, from the IaaS up to the application is currently ongoing in Torino site. The system is based on the ElasticSearch ecosystem, composed by ElasticSearch (ES), Logstash and Kibana and generally referred to as the 'ELK stack' [17]. In order to monitor the Virtual Analysis Facility application, the system relies on the Proof plugin TProofMonSenderSQL for collecting accounting data from the facility. At the end of each user query, data are gathered and sent to the database with a standard MySQL client/server protocol. In this case, the complex string processing capabilities of ES allows monitoring some additional observables such as e.g. the number of workers, the specific datasets analysed (LHC period, run, etc.) or the number of events processed. Further details can be found in [18].

A prototype architecture of VAF Data Federation has been also implemented. The goal of the VAF Federated Distributed Cluster is to create an alternative Storage System which can host the analysis datasets allowing the requests submitted by VAF users to be satisfied, bypassing the Alien Catalogue. To reach this goal the VAF technology has been combined with a Distributed Storage Cluster (DSC) solution based on Xrootd and including all the Italian VAF sites. The storage elements elected to join the DSC are linked to the Xrootd server nodes. To improve the scalability of the overall system, all these server nodes are not directly connected to a single Xrootd manager node. There is an intermadiate level composed by manager nodes, one for each VAF site, which has the task to manage all the server nodes belonging to that site. Using these intermediate nodes, file requests rectricted to a given VAF site can be handled. These manager nodes are also called *local redirector*: they are connected to a global manager node (*global redirector*) that can serve all the requests of files belonging to the whole Italian DSC. Fig. 7 schematically illustrates the Italian VAF Federated Distributed Cluster design, including all the possible request and response steps among the different parts involved in the data handling and usage

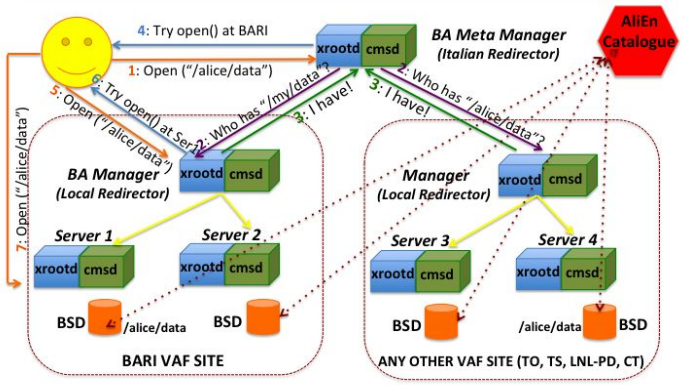

Fig. 7: Schematic picture of the VAF Federated Distributed Cluster design.

To keep the architecture more flexible and reliable, all the nodes run on different VMs provided by the Bari PRISMA Openstack Infrastructure [19]. The Xrootd Protocol imposes by construction that all nodes need to be network accessible from the client, hence the importance to manage this issue. In

order to prevent any problem concerning the VMs composing the VAF DSC, the dataset must not be located in the virtual disks of the server VMs but rather on the Block Storage Devices (BSDs) linked to them. This also provides the possibility to restore the server functionality in few minutes if something wrong happens to the VMs. Currently the size available for the Italian VAF BSDs ranges from 10 to 100 TB. Benchmarking activities are currently ongoing between Bari, Padova-Legnaro and Trieste sites: further details can be found in [20].

b. ATLAS

Grid sites are completely decentralized systems, but have by default no embedded High Availability functionalities. One of the goals of the R&D projects in Italy has been the addition of HA functionalities to standard centers, in particular for the Tier-2 centers of ATLAS. The traditional Grid Services were isolated and encapsulated in an HA envelope, by means of more modern techniques like the Cloud Computing facilities. This approach is adding an HA layer to the existing services while also providing native Cloud interfaces to be used by the ATLAS collaboration.

The Cloud Computing infrastructure is based on OpenStack and mainly using Gluster as backend filesystem. The cloudified systems are also used to extend the concept of HA to multiple sites, by federating more Cloud Systems and exposing the federated infrastructure as a single site. In this view, a pilot project of a Distributed T2 has been put in place between Napoli and Roma, using a dedicated Layer-2 link (Fig. 8), provided by GARR. The latencies of the link are such that it is possible to use synchronous storage replicas with distributed FileSystems like GlusterFS [21] and CEPH[22].

The Italian community has been testing the synchronous storage replication over WAN in extreme conditions, altering the link latency up to a factor ~7, simulating two site at the opposite sides of the country. Still the performance of the overall system is acceptable and not breaking the infrastructure integrity or disrupting the services. The replicated storage in the key point of a distributed set of centers: services can easily be migrated from a site to another one by exploiting the common, replicated storage facility and Cloud Computing infrastructures tailored to cope with both service continuity and disaster recovery, in order to achieve a full HA solution.

The plans of the distributed T2 experimentation are to expand the testbed from the existing two sites to a more wide configuration, using multiple sites and MPLS [23] transport.

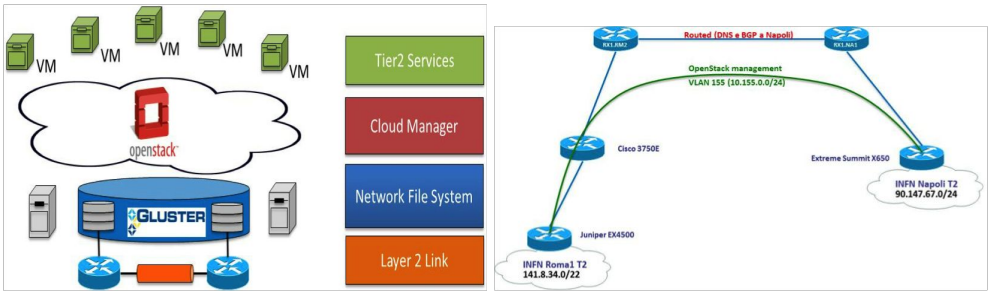

Fig. 8: The scheme of the Distributed Tier2 infrastructure between Napoli and Roma.

c. CMS

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[24]/PBS[25] sharing with Worker Nodes, to allow 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)[26], 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 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.

5. Conclusions

The Italian researchers part of the LHC Communities have started one year ago an R&D program in order to ease the use of national computing centers as analysis facilities.
Many of the solutions developed, either in their totality or in collaboration with other countries, are already in production, and are available to all the users of those centers, even from different research communities.

The present work is partially funded under program PRIN "/STOA-LHC 20108T4XTM/", /CUP: I11J12000080001,/.

6. Bibliography

[1] http://www.cern.ch/.MONARC
[2] http://lhcone.net
[3] N Xu, W Guan, S L Wu and G Ganis, *Data-oriented scheduling for PROOF*, Proceedings of CHEP 2010, 2011 J. Phys.: Conf. Ser. 331 032009.
[4] S Bagnasco et al., "Interoperating Cloud-based Virtual Farms", these Proceedings.
[5] Vilucchi E et al. 2014 PROOF-based analysis on the ATLAS Grid facilities: first experience with the PoD/PanDa plugin *J. Phys.: Conf. Ser.* **513** 032102 doi:10.1088/1742-6596/513/3/032102
[6] De Salvo A, 2012 Software installation and condition data distribution via CernVM File System in ATLAS *Journal of Physics: Conference Series* **396** (2012) 032030
[7] Bauerdick L et al. 2012 Using Xrootd to Federate Regional Storage *J. Phys.: Conf. Ser.* **396** (2012) 042009
[8] Furano F, 2013 The Dynamic Federations: federate Storage on the fly using HTTP/WebDAV and DMLite *The International Symposium on Grids and Clouds (ISGC) 2013*
[9] Cranshaw J et al. 2008 Building a scalable event-level metadata service for ATLAS *J. Phys. Conf. Ser.* **119**:072012, http://iopscience.iop.org/1742-6596/119/7/072012
[10] Cranshaw J et al. 2008 Integration of the ATLAS tag database with data management and analysis components *J. Phys. Conf. Ser.* **119**:042008, http://iopscience.iop.org/1742-6596/119/4/042008
[11] Cranshaw J et al. 2008 A data skimming service for locally resident analysis data *J. Phys. Conf. Ser.* **119**:072011, http://iopscience.iop.org/1742-6596/119/7/072011
[12] Mambelli M et al. 2010 Job optimization in ATLAS TAG-based distributed analysis *J. Phys. Conf. Ser.* **219**:072042, http://iopscience.iop.org/1742-6596/219/7/072042

[13] Barberis D et al. 2014 The ATLAS Eventindex: an event catalogue for experiments collecting large amounts of data *J. Phys.: Conf. Ser.* **513** 042002 doi:10.1088/1742-6596/513/4/042002
[14] http://www.uscms.org/SoftwareComputing/Grid/WMS/glideinWMS/doc.prd/index.html
[15] Douglas Thain, Todd Tannenbaum, and Miron Livny, "Distributed Computing in Practice: The Condor Experience" *Concurrency and Computation: Practice and Experience*, Vol. 17, No. 2-4, pages 323-356, February-April, 2005.
[16] http://www.openstack.org/
[17] https://www.elastic.co/products
[18] S Vallero, "Integrated Monitoring-as-a-service for Scientific Computing Cloud applications using the ElasticSearch ecosystem", these Proceedings.
[19] http://recas.ba.infn.it/recas1/index.php/recas-prisma
[20] F Colamaria et al., "Local storage federation through Xrootd architecture for interactive distributed analysis", these Proceedings.
[21] http://www.gluster.org
[22] http://www.ceph.com
[23] Network Working Group 2001 Multiprotocol Label Switching Architecture *rfc3031* https://www.rfc-editor.org/rfc/pdfrfc/rfc3031.txt.pdf
[24] http://www-03.ibm.com/systems/platformcomputing/products/lsf/
[25] http://www.mcs.anl.gov/research/projects/openpbs/
[26] https://wiki.infn.it/cn/ccr/aai/home