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

[a]A. De Salvo, [b]D. Elia, [c]T.Boccali, [d]L. Perini

[a]: INFN Sezione di Roma 1 La Sapienza [b]: INFN Sezione di Bari [c]: INFN Sezione di Pisa [d]: Universita' degli Studi and INFN, Milano

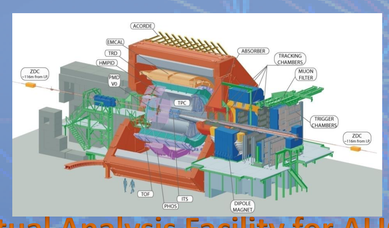
In 2012, 14 Italian Institutions participating in all major LHC Experiments won a grant from the ITALIAN MINISTRY OF RESEARCH (MIUR), to optimise Analysis activities and in general the Tier2/Tier3 infrastructure. We report on the activities being researched upon, on the considerable improvement in the 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.

The participating sites are:

- INFN (Pisa, Laboratori di Legnaro, Laboratori di Frascati)
- Univ. Genova
- Univ. Catania
- Univ. Milano
- Univ. Trieste
- Univ. Torino
- Univ. Bologna
- Univ. Perugia
- Univ. Roma 1 Sapienza
- Univ. Napoli
- Univ. Bari

ALICE – A Large Ion Collider Experiment

The ALICE experiment at the LHC



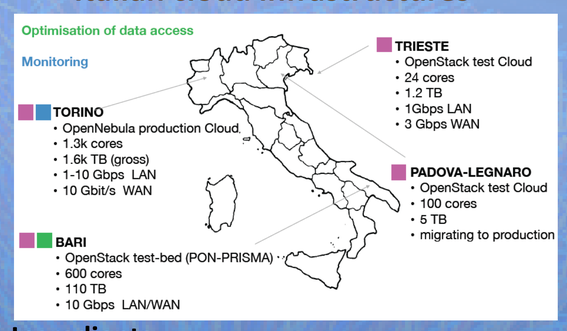
Virtual Analysis Facility for ALICE within the PRIN STOA-LHC Project

Main targets:

- Improve robustness of the existing LHC Italian infrastructure
- Global effort to ease data and resource access to LHC users
  - parallel and interactive analysis solutions (i.e. **V**irtual **A**nalysis **F**acility for ALICE)
  - standard access to interactive resources of different deployments
  - federation among different analysis facilities to optimize distribution and access to remote data
- Build a uniform environment capable of managing at once interactive and batch activities (Cloud Computing paradigm)
- Allow users outside high-energy physics to exploit LHC computing infrastructures

See also oral contribution [1] to this Conference.

Deploying VAFs on the Italian cloud infrastructures



Ingredients:

- Proof On Demand (PoD)
- **HTCondor** as batch system (cloud-aware)
- **Elastiq** daemon (optimization of resource usage)
- **CernVM Online** for cluster contextualization
- **XRootD** for distributed storage and data federation

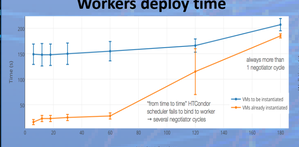
Activities:

- Benchmarking activities
- Test on local data storage access
- Application monitoring with ElasticSearch ecosystem
- Data federation

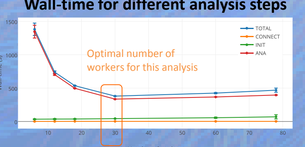
Testing VAF performance: preliminary results

Benchmark done analyzing about 200K Pb-Pb collision events:

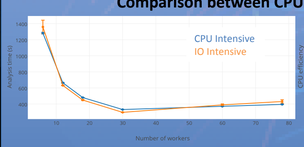
Workers deploy time

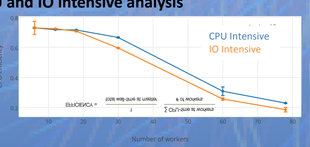


Wall-time for different analysis steps



Comparison between CPU and IO intensive analysis





Monitoring and accounting issues

The monitoring and accounting are based on **ElasticSearch**, **Logstash** and **Kibana** stack. A SQL database back-end for data preservation and to ensure flexibility to switch to different monitoring solutions.

The information is transferred from the database to the ElasticSearch engine via a custom Logstash plugin. A set of Kibana dashboards with pre-defined queries are provided.

See also oral contribution [3] to this Conference.

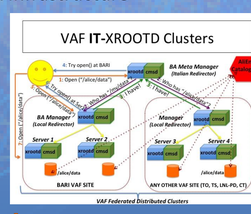
Distributed storage and data federation for the VAF

Basic idea:

- dedicated storage system shared among the VAF site federation needed to optimize resource usage
- usage of a unique **XRootD-based Italian redirector** to distribute and share relevant datesets encouraged by preliminary file access performance tests

First implementation:

- XRootD configuration tested in Bari using VMs provided by the PRISMA [2] Openstack Infrastructure
- hierarchy includes a local redirector (Manager) for each VAF site and a global Italian redirector (Meta-Manager) located in Bari



See also poster contribution [4] to this Conference.

Outlook

- complete VAF performance studies
- finalize data federation implementation and testing

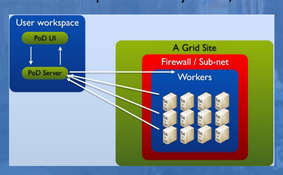
References

[1] https://indico.cern.ch/event/304944/contribution/447
[2] http://recas.ba.infn.it/recas1/index.php/recas-prisma
[3] https://indico.cern.ch/event/304944/session/7/contribution/389
[4] https://indico.cern.ch/event/304944/contribution/341

ATLAS A Toroidal LHC ApparatuS

Improving the analysis productivity

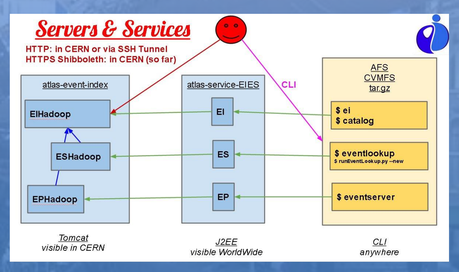
The Italian community is involved in the studies of possible improvements of the analysis techniques, via distributed tools and advanced data access technologies.
A particular focus has been given to the Proof-On-Demand facilities (PoD).
PoD is an extension of the standard Proof facilities, used to give 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.
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 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 and Http federations. 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 for example GPFS.



Improving the data access performance

ATLAS has been using for years standard DataBases to search the event in order to select only the interesting part of big amount of data. This database has been historically defined as Tag DB.
With the raise of new technologies like the Map-Reduce techniques and the Hadoop facilities, ATLAS has been redesigning its Tag DB, in order to evolve it to a more performing system, called EventIndex.
The EventIndex, now operational, is a collection of pointers to the events in the ATLAS datasets.
Its main features are:

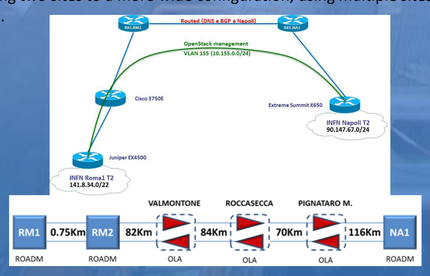
- Using the Hadoop facilities, no need for Oracle anymore
- Integration with the EventService, in order to move to the processing granularity of a single event, instead of a full dataset
- Full featured system, http(s) and CLI interfaces available



The Italian community has been coordinating the EventIndex activities since the beginning and is contributing both to the development and the operations of the system.

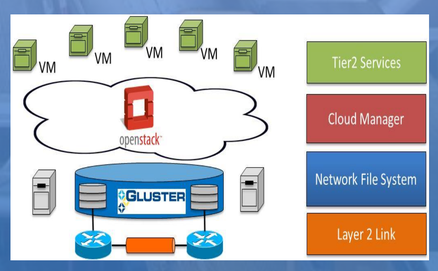
Improving the inter-site productivity

A pilot project of a Distributed T2 has been put in place between Napoli and Roma, using a dedicated Layer-2 link, provided by GARR, the Italian NREN.
This pilot is the prosecution of the studies carried out in Napoli to improve the reliability and to extend the functionalities to the Overlay network LHCONE.
The latencies of the link are such that it is possible to use synchronous storage replicas with distributed FileSystems like Gluster and CEPH.
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 transport.



Improving the intra-site productivity

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 has been 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.



CMS – Compact Muon Solenoid

Tests of new computing technologies

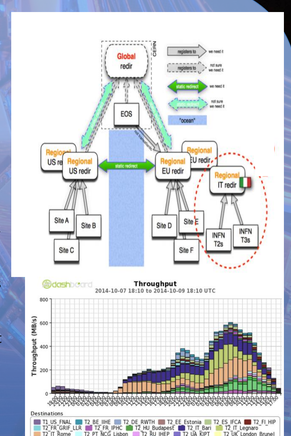
One strength of the **CMSSW** Software Stack is its complete independence from any proprietary code. A library / algorithm / tool, to enter the stack, must allow complete code distribution and patching.
In this way, we can recompile the full stack on **virtually any POSIX platform with a C++ compiler**, and even more easily on platforms which support g++.
We are currently performing benchmarking and porting activities on SoC architectures (Atom and ARM), also outside CMS boundaries.

(Italian) CMS Xrootd Federation

For the next LHC Run (2015-2017) the LHC Experiments, and CMS in particular, are planning to rely more and more on direct access to remote data. The **Italian federation of CMS Tier1/Tier2s** has been the second overall, and the first in Europe, to allow for remote Xrootd access.
Italy (Bari) hosts the European level redirector, which serves all European Tier2s and CERN. This is matched on the US side by the Nebraska redirector. Inside Europe, Italy has been configured as a sub-federation, and remote access are preferably done within its boundaries.

The Xrootd federation also provides a more robust computing infrastructure for CMS in Italy. A site's storage downtime does not automatically imply stop of the site activities: local CPUs can continue processing data received transparently from the other sites in the federation.

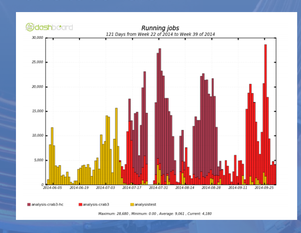
Tests have been done on the total possible throughput. Each italian T1/T2 is able to provide remote data access in excess of 500 MB/s



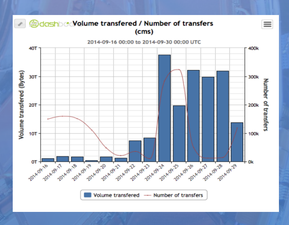
Pisa serving data to all CMS Tier2s at 600 MB/s

Tools for Distributed Analysis

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. Italy is responsible for the 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 bound to at least double by the Run2 start.
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



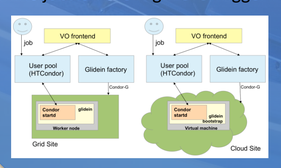
25k running jobs per day

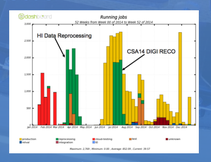


300k output files moved per day

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: instead of submitting pilot jobs, virtual machines (where the pilot jobs run) are created on demand.
At the Padova-Legnaro Tier2 a OpenStack Cloud based testbed has been set up, and here the model has been successfully demonstrated executing CMS CRAB analysis jobs. The same model is applied when running CMS offline jobs on the HighLevelTrigger resource at CERN





Jobs running at the HLT

Optimization of computing centres 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/PBS 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), 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.