TRIESTE

INFN Istituto Nazionale di Fisica Nucleare

CHEP'13 AMSTERDAM — CHEP2013: October 14 - 18, 2013 in Amsterdam, The Netherlands

Distributed storage and cloud computing: a test case

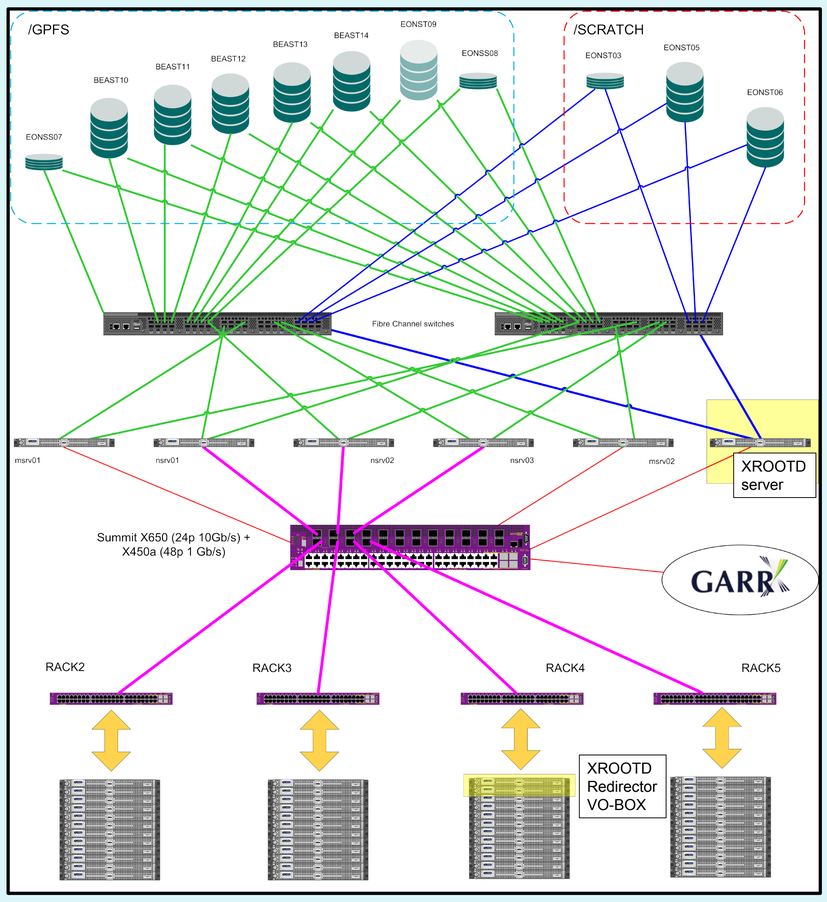
Stefano Piano[1], Giuseppe Della Ricca[1,2]

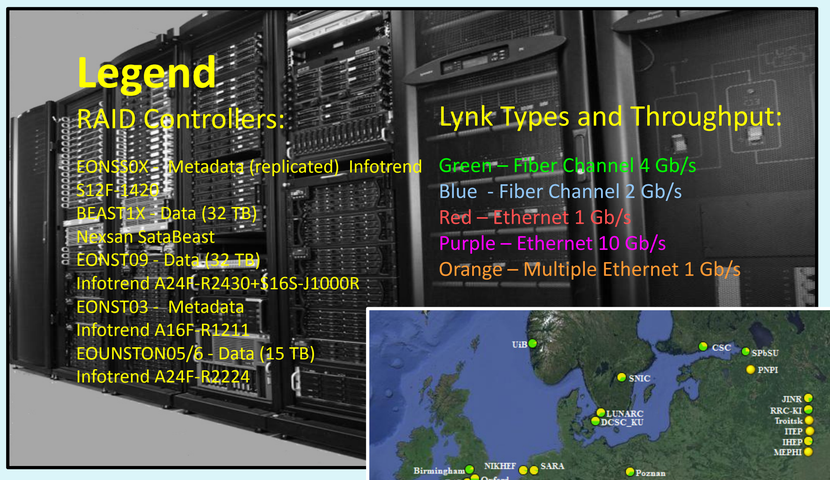
[1] INFN Sez. di Trieste, via A. Valerio 2, Trieste; [2] Dip. di Fisica, Univ. di Trieste, via Valerio 2, Trieste

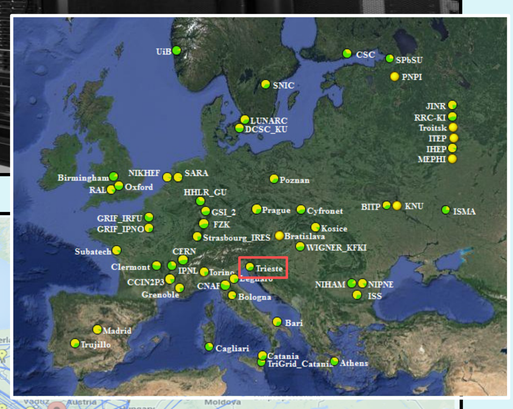
Introduction

Since 2003 the computing farm hosted by the INFN T3 facility in Trieste supports the activities of many scientific communities. Hundreds of jobs from 45 different VOs, including those of the LHC experiments, are processed simultaneously. The currently available shared disk space amounts to about 300 TB, while the computing power is provided by 712 cores for a total of 7400 HEP-SPEC06. Given that normally the requirements of the different computational communities are not synchronized, the probability that at any given time the resources owned by one of the participants are not fully utilized is quite high. A balanced compensation should in principle allocate the free resources to other users, but there are limits to this mechanism. In fact, the Trieste site may not hold the amount of data needed to attract enough analysis jobs, and even in that case there could be a lack of bandwidth for their access. The Trieste ALICE and CMS computing groups, in collaboration with other Italian groups, aim to overcome the limitations of existing solutions using two approaches. Sharing the data among all the participants, avoiding data duplication and taking full advantage of GARR-X wide area networks (10 GB/s) allows to distribute more efficiently the jobs according to the CPU availability, irrespective of the storage system size. Integrating the resources dedicated to batch analysis with the ones reserved for dynamic interactive analysis, through modern solutions as cloud computing, can further improve the use of the available computing power.

THE INFN TRIESTE SITE







References

ALICE: http://aliweb.cern.ch
CMS: http://cms.web.cern.ch
LHC: http://lhc.web.cern.ch/
LUSTRE: http://wiki.lustre.org/index.php/Main_Page
GARR-X: http://www.garr.it/rete/garr-x
GPFS: http://www03.ibm.com/systems/software/gpfs/
OPENSTACK: http://www.openstack.org/
PROOF: http://root.cern.ch/drupal/content/proof
XROOTD: http://xrootd.slac.stanford.edu/

Test Case for the TRIESTE CMS Computing Group:

Development and validation of a tool set for accessing remote data and for optimizing data storage resource utilization at the sites.

The aim is to develop a national federation data model with the verification of speed and reliability of a system with a single national namespace, with a fully distributed file catalog, possibly by means of the technology of the "Global Redirector":

1) A protocol that is independent of the underlying file system and that allows remote access of data is XROOTD:
   XROOTD gives the possibility of downloading any data from a remote storage when the file is opened, storing them to make later access quicker, and deleting the older data automatically when space is running out. In general, features supported by XROOTD are:
   - the capability to redirect requests from a client to several servers hosting the requested data;
   - the capability to read even small pieces of data without transferring the whole files;
   - the capability to implement dynamic cache nodes between server and client.

   The Trieste site makes available to the project one server for remote management of local data, in practice a server to interface the local parallel file system already existing with the remote data system.
2) The proprietary protocols such as GPFS, Lustre and Hadoop are also evaluated. These software solutions allow to manage geographically distributed storage among several sites but with the specific request to have the same storage management software installed in each of the sites. More advanced features of all protocols are tested such as:
   - the capability to have hierarchies of federations that allow remote access to data favoring those with lower latency;
   - the capability to define limits on the access number by a particular site or storage;
   - the capability to implement policies for authentication and authorization requests;
   - the capability to perform dynamic cache at a site based on the requests of the user job;
   - the capability to independently manage the available space for the cache;
3) The ultimate goal is to test the remote access by scientific applications, especially in all sites participating in this project, trying to satisfy at least the following use cases:
   - transient faults on the storage system of the site, e.g. files unavailable on the site where the job is running;
   - interactive access to the files: the user who is developing or debugging code can remotely access to files;
   - sites without local storage: a small site can remotely access the files necessary to run jobs.
   - dataset not found in sites of the federation of the project: If one or more users need to access for short time periods data that are not available on the sites of the federation, it will be possible to temporarily transfer this data in a storage area distributed among all sites thereby allowing access to them with good performance.
4) The development plan includes that:
   - all operations are done without the intervention of administrators;
   - the space is automatically released when it will be needed to accommodate new data.

Test Case for the TRIESTE ALICE Computing Group:

Development and validation of a virtual infrastructure for interactive analysis integrated into a multi-purpose computer center.

Interactive analysis requires quick allocation, for a limited time period, of computing resources with sufficient capacity to execute the user task in quasi real-time. The demand for interactive resources is unpredictable and uneven and cannot be handled efficiently by the allocation algorithms in a classical batch farm. We propose a novel approach to this problem, based on virtualization and Cloud tools and described below:

1) The interactive analysis model is based on PROOF, which is able to run over several events in parallel :
   - PROOF is able to distribute an analysis task to all the nodes of the cluster.
   - each node runs the analysis on one portion of the data, in parallel with the other nodes.
   - at the end all the single outputs are assembled together.

   The user sees the whole cluster as a single, extremely powerful, computer: he is able to develop and test the code on a single workstation, then run with minimal or no change on the virtual facility.
2) Access to interactive computing resources is not constant over time, but has typically access peaks alternating with periods of lower utilization. Moreover, computing centers such as the Trieste site primarily handle non-interactive applications. From the point of view of optimal utilization of resources:
   - the conversion of the infrastructure in a cloud infrastructure allows a more flexible management;
   - in a cloud infrastructure different types of applications are represented by different types of virtual machine;
   - unused resources could be allocate according to the needs;
   - OpenStack is chosen because it is generic and not geared toward a specific computing case;
   - the infrastructure virtualization is completely hidden to the end user: he uses resources (interactive analysis, Grid, various services) without the need to know the virtual nature of the underlying resources.
3) A model for implementation of the required resources through "provisioning" is evaluated:
   - a server stores different “software appliances” (software packages containing the application and just enough operating system to make it work).
   - depending on the needs these appliances are used to dynamically start virtual machines capable to satisfy needs more extensive than those so far taken into consideration, such as parallel sessions of PROOF, MPI sessions, etc..
4) From the results of research it is possible to achieve in cooperation with other computing sites the development of a national-wide interactive cluster:
   - PROOF with “multi-master”: a single national access point that communicates with the access points of different locations
   - “PROOF on Demand” which uses the underlying systems with proven scalability (resource management system) to manage a potential large amount of resources allocated to PROOF.

Acknowledgments

The present work is supported by the Istituto Nazionale di Fisica Nucleare (INFN) of Italy and is partially funded under contract 20108T4XTM of Programmi di Ricerca Scientifica di Rilevante Interesse Nazionale (Italy). The authors would like to acknowledge the support received from the team of the Trieste INFN computing team. The results of the project will be made public as a report at the project natural conclusion, in 2015.