Distributed storage and cloud computing: a test case

Distributed storage and cloud computing: a test case

S Piano1, **G Della Ricca**1, 2

1 INFN Sez. di Trieste, via A. Valerio 2, Trieste, Italy
2 Dip. di Fisica, Univ. di Trieste, via A. Valerio 2, Trieste, Italy

E-mail: `stefano.piano@ts.infn.it`

Abstract. Since 2003 the computing farm hosted by the INFN Tier3 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. 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 taking full advantage of GARR-X wide area networks (10 GB/s) and integrating the resources dedicated to batch analysis with the ones reserved for dynamic interactive analysis, through modern solutions as cloud computing.

1. Introduction

The experience acquired during the first year of data taking at CERN with the Large Hadron Collider [1] (in particular, by the CMS [2] and ALICE [3] experiments) has shown that event reconstruction as well as data analysis can be done entirely via the computational grid (GRID) [4]. In particular, GRID technologies for the LHC experiments (LCG, LHC Computing GRID) interconnects the various sites hierarchically being organized as a tree. The first level, Tier0, seats next to the experiments. Tier0 is designed to keep a copy of the raw data and to provide enough computing resources to perform the initial reconstruction of events. In the hierarchy, Tier0 is followed by Tier1's, which are nation-wide computational centers. A Tier1 fulfills the requirements of simulating and reconstructing raw data. Further computational centers, i.e., Tiers2's and Tier3's which are commonly hosted by Departments and Institutes, are dedicated to the study of both particle and nuclear physics via the final analysis of the reconstructed data. In order to avoid a chaotic data transfer among centers, a computational grid requires that the data analysis is performed were the data are stored in a DataGRID paradigm, which tailors the Tiers2's and Tier3's design. As an example, a Tier2(3) storing a large amount of data naturally involves a large amount of analysis, which requires its storage capability to be balanced by its computational CPU (and vice versa). Computing centers with large amount of processing power but poor amount of disk storage are not efficiently used. Conversely, computing centers with large amount of disk storage but low computing power are swamped by jobs that remain pending for a long time before running. The Tier3 Trieste site supports the activities of many scientific communities, and hundreds of jobs from 45 different Virtual Organizations

(VO), including those of the LHC experiments, are processed simultaneously. The requirements of different computational communities are not synchronized, and it can happen quite naturally that some of the CPU resources owned by one of the participants are used, while at the same time the resources of some other participant are free. These conditions call for a balanced compensation, which should allocate free resources to other users, but there are limits to the compensation mechanism if the site, as Trieste, cannot hold enough data to attract analysis jobs. Even if the data are enough there could be a lack of bandwidth for the data access. This may lead to a not optimal distribution of jobs for analysis, because the most important parameter is the presence of data and not the number of free CPUs (DataGRID). The decision to implement GRID in a DataGRID manner comes directly from the estimated cost of WAN transfers. There are other parameters into play, which are constantly evolving. In recent years many of the expected parameters have been confirmed (e.g., the growth of CPU power). The network aspect has probably exceeded the expectations of the '90s. Today, all Italian Tier2 CMS sites are connected at 2 Gbit/s through the Consortium GARR network, and a full upgrade to 10 Gb/s (GARR-x) [5] is expected within few years. With the prediction of 7/Mbit/s/kSi2K for the CMS experiment, a 10 Gb/s connection can handle 1.5 MSI2K which means about 800 jobs that read the data remotely. The use of pre-processed data (AOD) in the final phase of the CMS experiment analysis leads to a lower request of bandwidth for one job. A 10 Gb/s connection can support more than 1000 remote jobs if they work with AOD. Through remote access, each site would provide the CPU power not efficiently used in order to analyze the data hosted by other sites. To take full advantage of wide area networks the protocols should respond well to the increasing latency between the two connection endpoints. Moreover, their functionality should withstand the mistakes and failures, and they should facilitate not only the copy of files from one site to another, but also the chance to read well-defined file chunks without transferring unnecessary data. The protocols, born in different contexts, are either open as HTTP [6], XROOTD [7] and NFSv4.1 [8], or proprietary as Lustre [9] and GPFS [10]. All these protocols are characterized by functionality that allow their usage through routers and firewalls. Besides, they allow to redirect the connection to dedicated servers to provide a given service or data of interest. Specifically, protocols that allow access to WAN data such as HTTP and XROOTD are already used by the scientific community and in particular by LHC experiments.

A further implication of using the WLCG infrastructure for data analysis is the need to have a batch management for the data processing. This management is not so efficient for the final stages of analysis, typically operated by small groups of physicists or by individual researchers. 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 ($\sim$100 TB) involves the step to merge the partial results, this step is dominated by the I/O and file transfer, which lowers the CPU efficiency of this kind of activity. In order to provide our communities with resources for interactive parallel analysis, the ALICE collaboration has defined a standard for the deployment of Analysis Facility based on PROOF (Parallel ROOT Facility) [11], an extension of the ROOT framework. Currently, there are seven facilities, the main one is located at CERN. The physicists of the ALICE collaboration have had great advantage beyond their expectations by analyzing data on the Analysis Facility: the first published works were fully based on the analyzes run on the Analysis Facility at CERN, which quickly became inadequate for the needs of the experiment. That called for the creation of other similar structures. However, there is still no adequate solution to integrate the interactive analysis on WLCG infrastructure: the Analysis Facilities built so far are using dedicated resources. In Italy the CPU resources are funded exclusively to build computing farm accessible through batch jobs, i.e. computing nodes of WLCG. Most of the Italian sites are too small to dedicate nodes to interactive analysis, on the model of the CERN Analysis Facility. Besides, they are too small to host all AOD data necessary to perform the analysis on the whole collected data sample. The Trieste ALICE computing group, in collaboration with other groups in this project, aims to overcome the limitations of existing solutions in two ways: the first way is to integrate resources dedicated to batch analysis with dynamic interactive analysis through modern solutions as cloud computing. The second way is to federate computing centers in order to share data with all sites, avoiding data duplication. Sharing storage resources within a wide federation will help to overcome the limitation of local data access and to minimize data replication.

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

The aim of the Trieste CMS computing group is to contribute to the studies [12] [13] for the development of 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". An interface that is independent of the underlying file system and that allows remote access of data is XROOTD, a proven system of tools for accessing data that was originally developed for the BaBar experiment. 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. Another possible choice is the development that EMI (European Middleware Initiative) is conducting, i.e. a dynamic catalogs using the HTTP protocol. This approach will make possible to federate different types of Storage Elements supporting the HTTP protocol. In this project the solution released by EMI will be tested in order to evaluate the features and performances. The interest in the HTTP protocol to access files is due to its public availability and its extensive usage in many contexts. In general, the features that these protocols should support 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.

For both protocols the Trieste CMS computing group will make available 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. In particular, the Bari site will maintain a server with an up-to-date file index of all the storage systems participating to the federation. This server is commonly called Global "redirector" and is responsible for dispatching the user requests to the node that is physically capable of providing the requested file. In collaboration with the Bari site, the proprietary protocols such as GPFS, Lustre, Hadoop, etc., will also be evaluated. These softwares 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, which are considered usable, will be tested such as:

- the capability to have hierarchies of federations that allow, in a transparent way for the user and automatically for the administrator of resources, remote access to data favoring those with lower latency;
- the capability to define limits on the access number by a particular site or storage, to avoid overloading the servers that provide the data;
- the capability to implement policies for authentication and authorization requests from scientific communities and users who will use the system;
- 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 so that it can be cleaned up when new disk space is needed.

In collaboration with the Bari site the ultimate goal of this part of the Trieste CMS computing group will be 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: if some files are unavailable on the site where the job is running, then the job remotely accesses files by selecting the site that has the lowest latency network;
- interactive access to the files: the user who is developing or debugging code can remotely access to files of interest from any host, including your desktop / laptop, without knowing the location of the file;
- sites without local storage: a small site that cannot manage a complex storage system, can remotely access the files necessary to run jobs. That greatly reduces the human costs to manage a site of modest size;
- 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.

All operations will be done without the intervention of administrators. The space will be automatically released when it will be needed, for example to accommodate new data in place of those no longer accessed.

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

The interactive analysis does not consider waiting time but allocation of computing resources immediately to the user who requests them. The resources will be used for a limited period, typically repeating the analysis process several times on the same data. The interactive analysis model, well established and considered complementary to the batch model (GRID) creates some problems, that the Trieste Alice computing group aims to settle. The first problem is the use of computing resources, the second is the access to data. The interactive analysis model is based on PROOF. In a nutshell, PROOF is able to distribute an analysis task, developed using a PROOF-aware framework, to all the nodes of the cluster. Each node will run the analysis on one portion of the data, in parallel with the other nodes. At the end, all the single outputs (typically in the form of histograms) will be assembled together. PROOF is thus able to run over several events (namely p-p and/or Ion-Ion collisions in our cases) in parallel: this is sometimes referred to as a case of “embarrassingly parallel workload” [14], given the complete independence of the events to be analyzed. The user will see the whole cluster as a single, extremely powerful, computer: he will be able to develop and test the code on a single workstation, then run with minimal or no change on the virtual facility. 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 computing team will primarily handle non-interactive applications. From the point of view of optimal utilization of resources, the conversion of the whole infrastructure in a cloud infrastructure will allow a more flexible management. In a cloud infrastructure different types of applications will be represented by different types of virtual machines that could be allocate according to the needs, converting unused resources in interactive resources and vice versa to maximize their use. Among the various technologies for the management of a virtual or cloud computing center, composed by heterogeneous resources, OpenStack [15] was chosen because it is generic and not geared toward a specific computing case. The use of a “mainstream” technology does not restrict in any way the analysis interactive use case, but it will improve the quality and at the same time it will ease the management of other services that are not related to computing applications. It is worth noting that the infrastructure virtualization is completely hidden to the end user, who will use resources (interactive analysis, Grid, various services) without the need to know the virtual nature of the underlying resources. Evaluation of a installation model through “provisioning” (technical term that refers to the process by which resources are dynamically configured and assigned to different services according to needs). In view of doing a more general and flexible virtualization for the High Energy Physics use case, a model for implementation of the required resources through “provisioning” will be evaluated, provided that the milestones relating to the above points are reached. In this model, 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 will be 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 or MPI sessions. From the results of the two lines of research it will be possible to achieve in cooperation with other computing centers the development of a national-wide federated interactive cluster: from the point of view of PROOF this would mean assessing the feasibility and scalability of both approach “multi-master” (a single national access point that communicates with the access points of different locations) and an approach “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. The researchers developing the various use cases for interactive analysis will need early access to the prototype. It must be noted that early use and response by skilled researchers is essential to develop efficiently a useful tool. Thus we foresee to give access to the prototype in the first year to a limited number of users, who are supposed to connect from remote. The needed documentation will be made available on the web in a wiki-page format. We expect that initially the access will be quite rudimentary and the service, even though usable, will not be at a production quality level, since the prototype will evolve during the life span of this project.

4. 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 (PRIN, Italy). The authors would like to acknowledge the support received from the Trieste INFN computing team, the ALICE and CMS central computing teams. The results of the project will be made public as a report at the project natural conclusion, in 2015.

References

[1] The Large Hadron Collider (LHC) http://lhc.web.cern.ch/
[2] The Compact Muon Solenoid Experiment (CMS) http://cms.web.cern.ch
[3] A Large Ion Collider Experiment (ALICE) http://aliweb.cern.ch
[4] The Worldwide LHC Computing Grid (WLCG) http://wlcg.web.cern.ch/
[5] GARR-X http://www.garr.it/rete/garr-x
[6] RFC 2616: Hypertext Transfer Protocol – HTTP/1.1 http://tools.ietf.org/html/rfc2616
[7] The eXtended Root Daemon (XROOTD) http://xrootd.slac.stanford.edu/
[8] Network File System (NFS) version 4 Protocol http://www.ietf.org/rfc/rfc3530.txt
[9] The Lustre File System http://wiki.lustre.org/index.php/Main_Page
[10] IBM General Parallel File System (GPFS) http://www03.ibm.com/systems/software/gpfs/
[11] The Parallel ROOT Facility http://root.cern.ch/drupal/content/proof
[12] Grandi C *et al.* 2012 *J. Phys. Conf. Ser.* **396** 032053
[13] Bauerdick L *et al.* 2012 *J. Phys. Conf. Ser.* **396** 042009
[14] Foster I 1995 *Designing and Building Parallel Programs* (Reading, MA: AddisonWesley) section 1.4.4
[15] OpenStack A Cloud Operating System http://www.openstack.org/