Journal of Physics: Conference Series **664** (2015) 042013 doi:10.1088/1742-6596/664/4/042013

Local storage federation through XRootD architecture for interactive distributed analysis

F Colamaria[1], D Colella[1], G Donvito[2], D Elia[2], A Franco[2], G Luparello[3], G Maggi[4], G Miniello[1], S Vallero[5], G Vino[1]

[1] Università degli Studi di Bari and INFN Sezione di Bari
[2] INFN Sezione di Bari
[3] Università di Trieste and INFN Sezione di Trieste
[4] Politecnico di Bari and INFN Sezione di Bari
[5] Università di Torino and INFN Sezione di Torino

E-mail: Domenico.Elia@ba.infn.it

Abstract. A cloud-based Virtual Analysis Facility (VAF) for the ALICE experiment at the LHC has been deployed in Bari. Similar facilities are currently running in other Italian sites with the aim to create a federation of interoperating farms able to provide their computing resources for interactive distributed analysis. The use of cloud technology, along with elastic provisioning of computing resources as an alternative to the grid for running data intensive analyses, is the main challenge of these facilities. One of the crucial aspects of the user-driven analysis execution is the data access. A local storage facility has the disadvantage that the stored data can be accessed only locally, i.e. from within the single VAF. To overcome such a limitation a federated infrastructure, which provides full access to all the data belonging to the federation independently from the site where they are stored, has been set up. The federation architecture exploits both cloud computing and XRootD technologies, in order to provide a dynamic, easy-to-use and well performing solution for data handling. It should allow the users to store the files and efficiently retrieve the data, since it implements a dynamic distributed cache among many datacenters in Italy connected to one another through the high-bandwidth national network. Details on the preliminary architecture implementation and performance studies are discussed.

1. Introduction

The Bari ReCaS Data Center [1], previously known as the Bari Computer Centre for Science (Bc^2S), currently hosts the PRISMA [2] Openstack Cloud Infrastructure which has been set up jointly by the Italian Institute for Nuclear Physics (INFN) and the University of Bari. The infrastructure has the aim to offer to users a quota of virtual machines (VMs) providing de facto a computing service on demand (Infrastructure-as-a-Service, IaaS). All the core services are provided in High Availability mode: backup instances (*replica*) are automatically available if the primary ones become temporary unavailable.

In High Energy Physics (HEP) the Grid Computing Model is the most widespread model for running analysis jobs. Nevertheless, nowadays Virtualization and Cloud Computing are considered two suitable and reliable complementary technologies to fulfill the high-throughput computing tasks required by HEP. A Virtual Analysis Facility (VAF) for the ALICE experiment at the LHC [3] has been developed in Bari. Such a computing facility consists of a PROOF cluster of virtual machines dynamically deployed by the Openstack cloud infrastructure built in Bari for the PRISMA project. In this context Grid jobs are replaced by *flexible* VMs available *on demand* by users, ensuring more *interactivity* in running data analysis tasks and optimization in the usage of the computing resources. Similar facilities are currently running also in other Italian sites with the aim to create a federation of interoperating farms able to provide their computing resources for interactive distributed analysis. Since most of these Italian VAF sites are also LHC Tier-2 centers, the interest in the creation of a Distributed Storage System and a Data Federation among them comes from the necessity of sharing their powerful computing resources overcoming, at the same time, the limitations which arise from accessing files within the single site only through the AliEn [4] Catalogue. In this paper the elasticity and the scalability features of the VAF site at Bari along with the implementation of the Distributed Storage System and the ALICE Data Federation using the XRootD [5] remote file access protocol are presented.

2. The Virtual Analysis Facility

A VAF is a dedicated PROOF (Parallel ROOT Facility)-based [6] cluster able to provide a variable number of VMs to users. The possibility of creating and terminating the virtual instances on demand ensures the cluster to be flexible, scaling up when analysis jobs are running on PROOF workers or scaling down when they are idle. All the VMs are provided and deleted elastically by the python tool called elastiq [7]. This tool periodically checks the status of EC2 [8] VMs belonging to HTCondor [9] cluster to find and terminate the idle ones. The VM contextualization and the elastiq configuration are provided using the web interface of CernVM Online [10]. The latest VAF cluster configuration at Bari provides up to 56 PROOF workers. In Fig. 1 the time taken by the PROOF workers to join the cluster is reported as a function of the number of available PROOF workers themselves (from 12 to 44): it shows clearly the elasticity of the cluster, the 44 workers being totally ready to join the cluster in about 9 minutes only.

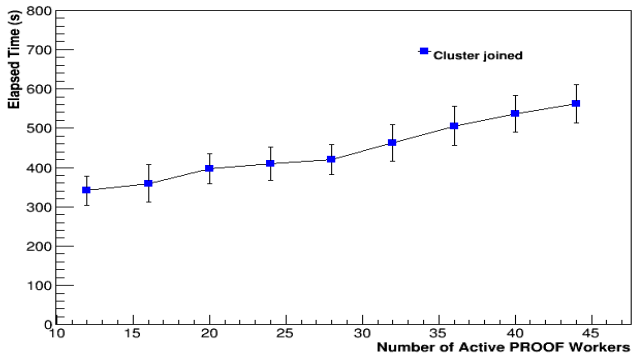

Fig. 1: Total time for the workers to join the cluster as a function of the number of workers.

The total analysis time reduction accessing files through AliEn Catalogue has been calculated to be approximately 80% just passing from a 4- to a 28-PROOF worker configuration (see section 5). In next sections an overview of the XRootD architecture used to implement the VAF Data Federation and the hierarchical configuration used at Bari is described, trying to make first comparisons to understand differences between data access via AliEn and that via the Distributed Storage System.

3. Distributed Storage and Data Federation using XRootD

The sharing of storage resources belonging to the Italian VAF sites creating a Data Federation has many advantages. It can provide a single shared disk space available to host analysis dataset files and then allow to use more efficiently the overall disk space itself. The XRootD remote file access protocol has been chosen for the creation of a unique distributed storage system starting from each single local VAF storage. Such a choice is able to provide a storage solution capable to export and aggregate filesystems from a distributed set of hosts in a very efficient way, assuring high performances and scalable fault tolerant access to data repositories. All these features make it suitable to be used for a Distribuited Storage. Preliminary studies on XRootD remote file access performances using an Italian redirector running both CPU intensive (≈ 83%) and I/O intensive (≈ 75%) analyses have been carried out, providing the following encouraging results: the wall clock time of an I/O intensive (CPU intensive) analysis job task accessing data via XRootD is about 19% (10%) greater than accessing files locally.
To provide such distribuited storage, a cluster must be set up using the server and the manager XRootD nodes. The first one contains the file resources to share whereas the manager node contains all the meta-data of the server nodes subscribed to it. A client must submit a file request to the manager node that selects the best available server by using the information from all server nodes. Then the manager tells the client to redirect the open request to that server. This policy allows to optimize the traffic and the machine usage for the Distributed Storage Cluster.

4. VAF Federated Distributed Cluster

The goal of the VAF Federated Distribuited 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 Distribuited 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.
Before running a given analysis campaign, the interested dataset must be staged in the DSC. This staging is performed by the system administrator directly from the AliEn Catalogue. The AliEn suite is installed only on the server nodes so that the administrators can download the interested dataset using both manual or automatic tools. This procedure allows to manage more efficiently the server storage resources in each site and prevents possible misuses which can also reveal to be unsafe. Once the dataset staging is completed, any client submitting its request to the global redirector can be successfully served. Fig. 2 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.
To keep the architecture more flexible and reliable, all the nodes run on different VMs provided by the Bari PRISMA Openstack Infrastructure. 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.

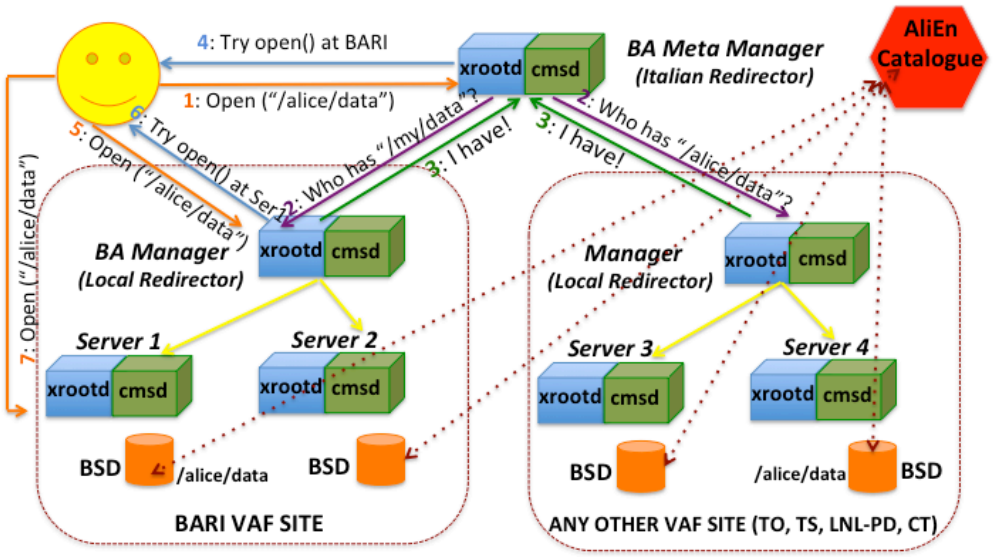

Fig. 2: Schematic picture of the VAF Federated Distributed Cluster design.

5. Benchmarking activities

Fig. 3 shows the total Wall Clock Time (WCT) of the benchmark analysis jobs as a function of the available number of PROOF-workers. The blue line refers to jobs accessing dataset files via AliEn Catalogue, the yellow line refers to jobs accessing files via VAF Data Federation (DF) from the Padova-Legnaro site, the green line refers to jobs accessing files via VAF DF from the Torino site and the red line refers to jobs accessing files via VAF DF in the Bari site. In the region corresponding to a number of worker nodes smaller than 30-35 an almost constant reduction of the total WCT when accessing data through the DF in PD-LNL and Torino sites compared to the access through the Alien Catalogue is observed. For an increasing number of worker nodes such a difference tends to vanish for PD-LNL due to the reduced bandwidth for that site (~1Gbps compared to ~10Gbps for Bari and Torino).

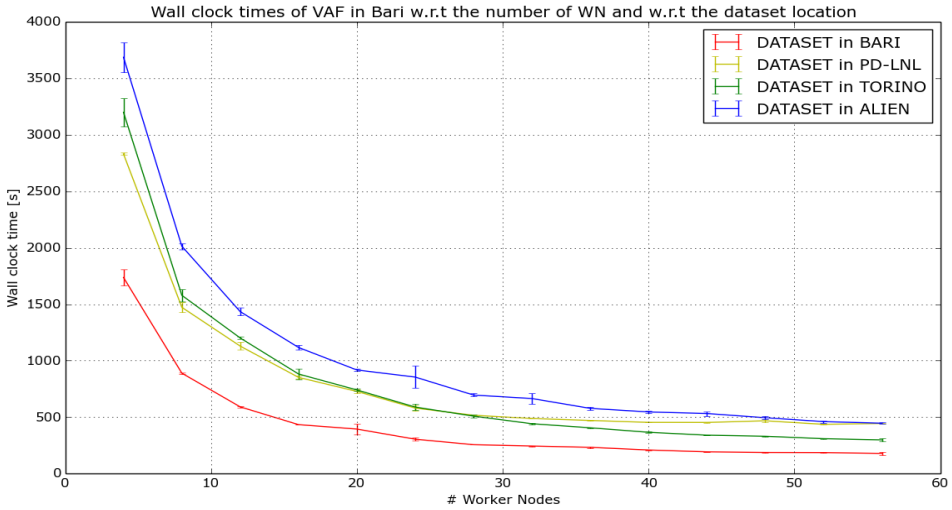

Fig. 3: Total WCT of the benchmark analysis for different data accesses via DF and via Alien.

6. Conclusions

A Virtual Analysis Facility for ALICE experiment exploiting PRISMA OpenStack Infrastructure has been setup and deployed in Bari. Elasticity and scalability performances have been studied: more than 40 PROOF workers can be available in less than 10 minutes and a meaningful reduction of the analysis job wall clock time has been observed increasing progressively the number of the active PROOF-workers.
The implementation of a Data Federation and a Distributed Storage System among italian VAF sites to preserve and share all the ALICE datasets needed to perform analysis campaigns has been also studied and set up. The Bari site currently hosts the national XRootD global redirector. Preliminary tests on the ALICE VAF DF have been carried out to check and optimize its performances. As well as for accessing files via AliEn Catalogue, the WCT for the benchmark analysis job still reproduces quite well the scaling as a function of the number of PROOF-workers available. The saturation visible when using a cluster configuration with more than 30 PROOF-workers when accessing data from the PD-LNL DF is strictly related to the temporary limits of the networking hardware resources in that site. In few months additional sites are expected to join the DF, with uniform network performance: this will allow to study the most efficient XRootD VAF cluster configuration, possibly including a redundancy of the global and the local redirector to ensure failover.

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

Bibliography

[1] http://recas.ba.infn.it/recas1/index.php/recas-ba/datacenter
[2] http://recas.ba.infn.it/recas1/index.php/recas-prisma
[3] http://home.web.cern.ch/about/experiments/alice
[4] http://alien2.cern.ch/
[5] https://root.cern.ch/drupal/content/proof
[6] http://xrootd.org/
[7] https://github.com/dberzano/elastiq
[8] http://aws.amazon.com/it/ec2/
[9] http://research.cs.wisc.edu/htcondor/
[10] https://cernvm-online.cern.ch/dashboard/