Local storage federation through XRootD architecture for interactive distributed analysis

F. Colamaria[1,2], D. Colella[1,2], G. Donvito[2], D. Elia[2], A. Franco[2],
G. Maggi[2,3], G. Miniello[1,2], S. Piano[4], S. Vallero[5,6], G. Vino[1,2]

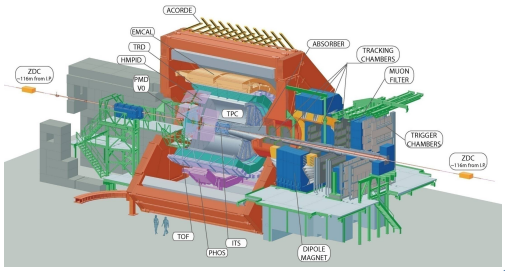
[1]Università degli Studi di Bari - Italy, [2]INFN Sezione di Bari - Italy, [3]Politecnico di Bari - Italy,
[4]INFN Sezione di Trieste - Italy, [5]Università degli Studi di Torino - Italy, [6]INFN Sezione di Torino - Italy

The present work is supported by the Istituto Nazionale di Fisica Nucleare (INFN) of Italy,
developed in the framework of contract 20108T4XTM of Programmi di Ricerca Scientifica di Rilevante Interesse Nazionale (STOA-LHC PRIN, Italy) and partially funded by it.

The ALICE Experiment

ALICE (A Large Ion Collider Experiment) is one of the four major experiments at the Large Hadron Collider (**LHC**):

- world-wide collaboration, composed of over **1500 members**, from 154 physics institutes in 37 countries
- general-purpose experiment, composed of **18** different **detectors**
- optimized to investigate the properties of strongly interacting matter in high-energy collisions between lead ions, where a new phase of matter (Quark-Gluon Plasma) is produced
- ALICE physics programme includes also detailed studies of proton-lead and proton-proton collisions
- during the Run1, **7 PB** collected raw data (**16 PB** including reconstructions and Monte Carlo productions)
- in order to cope with storage and computing requests, a GRID environment organized in **more than 100 computing centres** on various Tier's is exploited



Virtualizion and Cloud Computing: A Virtual Analysis Facility for ALICE

- A **Virtual Analysis Facility** (**VAF**) for the ALICE experiment has been developed on Bari PRISMA [1] Openstack Infrastructure. The main aim is to investigate how to complement GRID Computing for the high-throughput tasks required by the High Energy Physics, in particular exploiting the **Cloud Computing** technology ;
- A **VAF** is a dedicated PROOF-based cluster of Virtual Machines (VMs);
- **PROOF** (Parallel ROOT Facility) [2] is an extension of the ROOT data analysis framework that allows interactive analysis on a local cluster parallelizing tasks;
- **VMs contextualization** and elastiq configuration are provided using the web interface of **CernVM Online** [3];
- The **VMs** are based on the light virtual machine image called **μCernVM**; they are provided and deleted elastically via the python tool **elastiq** [4]:
  - ✓ **SCALING UP**: if jobs are waiting, it requests new virtual machines;
  - ✓ **SCALING DOWN**: if some machines are idle, it turns them off;
- Grid computing jobs are replaced by **flexible VMs** ensuring **more interactivity** when running analysis tasks;
- The **VMs** are available for the end-user just **on demand** and **automatically created and terminated**, ensuring an **optimization of computing resource usage**;
- The **latest VAF configuration at Bari** provides up to **56 PROOF workers** with a total analysis job wall time reduction larger than 80 % going from 4 to 32 VMs.

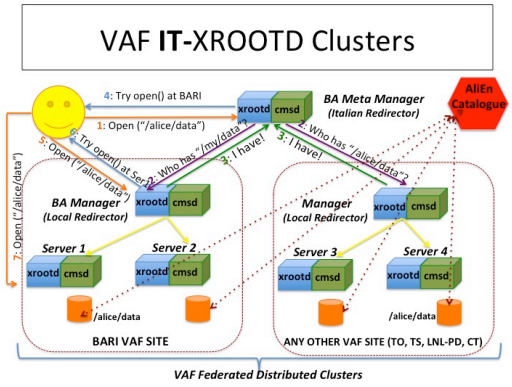
Distributed Storage and Data Federation using XRootD

- Most of the **Italian Virtual Analysis Facility sites** (BA, TO, TS, LNL-PD, CT, CA) are also **LHC Tier-2** hosting **powerful computing resources**;
- PROOF master and related workers are not real hosts:
  - ✓ The creation of a **dedicated storage system shared among all these sites in order to preserve datasamples** is highly recommended;
- Dataset staging from **AliEn Catalogue** is currently available on VAFs only locally:
  - ✓ The sharing of these resources creating a **Data Federation among VAFs** has been studied and implemented at Bari, exploiting the **XRootD** remote file access protocol via a unique national redirector;
- **Studies on XRootD** remote file access performances using an Italian redirector have been preliminarly carried out running both CPU intensive (≈83%) and I/O intensive (≈75%) analyses:
  - ✓ **The wall clock time of an I/O intensive analysis job task** accessing data **via XRootD** has been calculated to be **≈19% greater** than accessing files locally;
  - ✓ **The wall clock time** of a **CPU intensive analysis job task** accessing data via XRootD has been calculated to be **only ≈10% greater** than accessing files locally;

VAF Federated Distributed Clusters

- **Figure 1** schematically illustrates the **XRootD configuration for a VAF Data Federation** developed in Bari using the VMs provided by the PRISMA Openstack Infrastructure;
- The XRootD hierarchy tested includes a local redirector (**Manager**) for each VAF site - in which a number of servers, each carrying a block storage device, is provided – and a global Italian redirector (**Meta-Manager**) located in Bari;
- **Block storage devices** for each server currently range **from 10 to 100 TB**;
- The datasets for each analysis campaign are expected to be staged locally by system administrators from the AliEn Catalogue: authentication issues and waste of disk usage at user level can be therefore avoided.

Fig.1



Summary and Outlook

- A **VAF Cluster Federation using XRootD** remote file acces protocol has been implemented at **Bari**: it currently hosts the unique **Italian global redirector** for all the VAF sites in Italy;
- **Access**, **sharing** and **balancing** of all the **datasets** needed for any given ALICE analysis campaign can be optimized using the implemented XRootD configuration;
- Most of the infrastructure has been currently set-up and the related storage resources are ready to be populated:
  - ✓ more tests are needed to optimize the data staging and the global XrootD configuration among all the VAF sites.

References

- [1] http://recas.ba.infn.it/recas1/index.php/recas-prisma
- [2] http://root.cern.ch/drupal/content/proof
- [3] https://cernvm-online.cern.ch/
- [4] https://github.com/dberzano/elastiq

Contact Person: Domenico Elia domenico.elia@ba.infn.it | fabio.colamaria@ba.infn.it | domenico.colella@ba.infn.it | giacinto.donvito@ba.infn.it | antonio.franco@ba.infn.it | giorgio.maggi@ba.infn.it | giorgia.miniello@ba.infn.it | stefano.piano@ts.infn.it | svallero@to.infn.it | gioacchino.vino@ba.infn.it

CHEP2015 OKINAWA, japan