Interoperating Cloud-based Virtual Farms

2015 J. Phys.: Conf. Ser. 664 022033

(http://iopscience.iop.org/1742-6596/664/2/022033)

Journal of Physics: Conference Series **664** (2015) 022033 doi:10.1088/1742-6596/664/2/022033

Interoperating Cloud-based Virtual Farms

S Bagnasco[1], F Colamaria[2], D Colella[2], E Casula[3], D Elia[4], A Franco[4], S Lusso[1], G Luparello[5], M Masera[7], G Miniello[2], D Mura[3], S Piano[6], S Vallero[7], M Venaruzzo[8], G Vino[2]

[1] INFN Sezione di Torino

[2] Università degli Studi di Bari and INFN Sezione di Bari

[3] Università degli Studi di Cagliari and INFN Sezione di Cagliari

[4] INFN Sezione di Bari

[5] Università degli Studi di Trieste and INFN Sezione di Trieste

[6] INFN Sezione di Trieste

[7] Università degli Studi di Torino and INFN Sezione di Torino

[8] INFN Laboratori di Legnaro

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

Abstract. The present work aims at optimizing the use of computing resources available at the grid Italian Tier-2 sites of the ALICE experiment at CERN LHC by making them accessible to interactive distributed analysis, thanks to modern solutions based on cloud computing. The scalability and elasticity of the computing resources via dynamic ("on-demand") provisioning is essentially limited by the size of the computing site, reaching the theoretical optimum only in the asymptotic case of infinite resources. The main challenge of the project is to overcome this limitation by federating different sites through a distributed cloud facility. Storage capacities of the participating sites are seen as a single federated storage area, preventing the need of mirroring data across them: high data access efficiency is guaranteed by location-aware analysis software and storage interfaces, in a transparent way from an end-user perspective. Moreover, the interactive analysis on the federated cloud reduces the execution time with respect to grid batch jobs. The tests of the investigated solutions for both cloud computing and distributed storage on wide area network will be presented.

1. Introduction

The computing models of the LHC experiments reflect an organization originally proposed by the MONARC model, that foresees a hierarchically organized set of computing centers, all connected in a computing Grid and each with a size compatible with deployment at a single institution. Whereas in the original schema each tier was assigned a different role and purpose, in ALICE [1] such distinctions have become increasingly fuzzy. Since the very beginning, and with the evolution of the computing model, the sites have been assigned to Tier-1 or Tier-2 essentially according to their size and the availability of custodial storage for a second collective copy of the full RAW data sample. The computing resources of the LHC experiments are federated in the WLCG Project, in which the experiments and the computing sites participate with a formal MoU.

A implication of using the WLCG infrastructure for data analysis is the need to have a batch management for the data processing. This management may not be so efficient for some stages of the analysis, typically operated by small groups of physicists or by individual researchers, for which the latency time between analysis job submissions and their actual executions on the GRID is comparable with the execution time. Moreover, the analysis of large samples of ALICE experiment data (up to ~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), 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 Facilities: the first published works were fully based on analyses that were 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 Italian ALICE computing group, in collaboration with other Italian groups involved in the same project [2], 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 technologies. On the other hand, the project aims to federate the Italian ALICE analysis centres, taking full advantage of the good network connectivity provided by GARR-X, through a distributed cloud in order to be able to run cloud based applications over data existing in any of the federation members. Such a federated cloud extends the data available for interactive analysis to the whole set managed by Italian ALICE Tier-2 sites, reducing execution time with respect to grid batch job and the need of data duplication with respect to different “unfederated” analysis facilities.

In recent years many of the expected parameters of the computing model have been confirmed (e.g., the growth of CPU power); however, the network development has probably exceeded the expectations of the '90s. Today, all Italian Tier2 ALICE sites are connected at 10 Gbit/s through the network provided by Consortium GARR, the Italian NREN. 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 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. Such protocols, born in different contexts, are either open as HTTP, XROOTD [3], NFSv4.1 [4] and GlusterFS [5] or proprietary as Lustre [6] and GPFS [7]. 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.

2. Italian infrastructure for the Virtual Analysis Facility

The need for a structure allowing a fast and interactive access to the data is evident in several use cases such as optimization of algorithms, code debugging, fast data quality monitoring and, finally, data analysis, at least in its final stage that leads to results ready to be shown and published; in general, any application where the overhead and latency typical of running on the Grid makes its use impractical. The experience gained by the LHC experiments in these first years of data taking, allows to conclude that both data reconstruction and analysis up to the final results ready for publication, can be fully accomplished on a computational grid. However, the WLCG infrastructure, with its batch-based management of the data processing and access is not the optimal solution for the use cases mentioned above and for the final stages of the analysis process, starting from pre-processed datasets (AOD). Moreover, the analysis of large data samples (up to ~100 TB) implies a merging phase of partial results, dominated by I/O operations and file transfers, which depends on how the partial results are spread out on the grid.

The tool of choice for similar use cases is, in the context of ALICE, PROOF (Parallel ROOt Facility) [8], an extension of the ROOT framework. PROOF is a parallel computing framework distributed with ROOT in which several workers share the computing load exploiting the event-based parallelism that is intrinsic in most High-Energy Physics data processing. PROOF and ROOT are primarily designed for physics data processing, and in particular for HEP analysis, where input data can be logically divided into several independent physics events. The difference between PROOF and the grid is interactivity, the ability to send commands and execute them in parallel. The advantage of PROOF is that the users do not need to submit jobs to exploit computing resources: opening a PROOF session means acquiring the control over a distributed set of resources for some time, hence the user's interactivity.

The first visible advantage of such interaction is that different pieces of the outputs produced by several workers are collected automatically and presented to the user at once, directly on the client. On the contrary, batch models produce independent output data which have to be collected and merged manually.

PROOF features a scheduler [9] that dynamically assigns data to workers, as the analysis is progressing. The PROOF scheduler is called packetizer, as input data is divided into work units called packets. A packet is a set of physics events, or, by referring to a more general data structure, a set of ROOT tree entries. Among the different packetizers available for PROOF, the adaptive packetizer takes into account that not all events require the same time to be processed. When performing interactive analysis before collecting the results, waiting for the slowest worker to finish is a clear waste of resources. The adaptive packetizer purpose is the uniform completion time of workers by assigning the workload non-uniformly.

PROOF on Demand (PoD) makes PROOF run without any static deployment or administrative privileges. The advantage of PoD is that every user can use PROOF provided that the facility has some resources at its disposal: no intervention from a system administrator is needed.

Instead of configuring a virtual static farm, it is possible to have an application that expands and reduces automatically, depending on its load, by varying the number of running virtual machines. Here we call "elastic" applications which can automatically change the amount of cloud resources on demand.

An elastic application does not usually directly expose the Infrastructure as a Service (IaaS) interface to the end user: the user does not need to know that its application is running on top of virtual machines. Instead, it exposes the cloud using an additional layer, as the Platform as a Service (PaaS) approach that will be pursued in the present project. The users submit their own complex compiled programs using the provided platform, which is usually a batch system or Workload Management System. Such a system can exploit elasticity by requesting more virtual machines in cases it sees too many jobs waiting in the

queue, and turn off idle virtual machines in order. Elasticity becomes a very important factor on public clouds when considering their billing model: all the major cloud providers, including Amazon, charge with a per-use policy, meaning that an efficient model to expand the application only when needed ultimately saves money.

The present project plans to overcome the limitations of the existing solutions by devoting a dynamically determined fraction of the resources currently allocated for batch analysis to dynamic interactive analysis in a cloud computing environment. Furthermore, a model for efficient access to experimental data will be developed and implemented, which is likely to be based on heuristic placement and caching of stored datasets. This will leverage on the work of a currently ongoing project and on the experience gained in several years of data management for the ALICE experiment.

This project is part of the Italian government-funded STOA-LHC project, a common effort across the three largest collaborations aimed at improving the robustness and usability of the existing LHC Italian computing infrastructure.

3. Production activities on the Torino site

At the beginning of project, the ALICE Torino site has deployed a Virtual Analysis Facility (VAF) [12] to tackle the complexity of deploying a PROOF cluster within an Infrastructure as a Service (IaaS) approach [13]. The Virtual Analysis Facility is a cluster of zero-configuration virtual machines dedicated to PROOF analysis, even though the tools themselves can be disassembled and used independently of each other. The Virtual Analysis Facility is completely self-contained, meaning that it can run on every cloud, public or private, which allows the contextualization of virtual machines. Moreover, its configuration is not specific to either ALICE or other experiments, making the Virtual Analysis Facility a complete and generic “analysis cluster in a box”. The Virtual Analysis Facility works on top of an IaaS cluster of virtual machines. Its main components are:

ROOT and PROOF as an analysis framework and computing model; the CernVM ecosystem [14] for the cloud deployment; PROOF on Demand and HTCondor [15] for scheduling users and preventing the issues of a static PROOF deployment; the Elastiq daemon communicating [16] to an EC2 interface to provide automatic and transparent scalability of the VAF cluster; an interface for authentication and authorization allowing grid users to connect via SSH using only their grid credentials, called sshcertauth; a client providing the shell and workers environment based on specific experiments frameworks.

In production since Novembre 2013, the VAF in Torino provided up to about 100 worker nodes (WN) to a number of analysis activities, mostly from the local community; the range of possible activities is mostly limited by the available disk space on a dedicated storage, about 60 TB, effectively limiting the data that can be analyzed to ntuples or small ESD datasets.

The Torino group now plans to experiment applying automatic elasticity to different applications; for a report of this activity see [10].

Similar facilities are currently running using the Turin VAF model also in other Italian sites (Bari, Cagliari, Padova-Legnaro, Trieste) with the aim to create a federation of interoperating farms able to provide their computing resources for interactive distributed analysis. The present project has integrated these Italian sites (some of them are also ALICE Tier-2 sites) into a prototype for a federated infrastructure connecting cloud structures deployed. This is the first step to be able to form a national-wide cloud federation allowing to extend the VAF to all the Italian ALICE sites and making available for interactive analysis their data already stored into SE.

4. Benchmarking

To create a reliable benchmark for the VAF's, an ALICE analysis macro was "instrumented" in order to return the Wall and CPU Clock Time (WCT and CCT) of the different phases of the instance life and of the analysis steps. The code was originally implemented to analyse the secondary vertices coming from the multi strange particles decay and, coupled to a specific ALICE data set (96 GB), was used as a standard benchmark for system performance. The main result obtained in Torino site is summarized in the two next figures, where the deploy time of the VAF Worker Nodes (WN) and 3 steps of the PROOF analysis are measured increasing the number of the workers. The data are stored locally in a GlusterFS volume and exposed to the VAF by a XROOTD interface. If new virtual machines (VM) need to be instantiated, WN deploy time ranges from 2.5 min to 3.5 min and if VMs are already available, WN deploy time ranges from 16 s to 3 min. For this type of the analysis and for the number events considered the optimal performance is reached with ~30 workers, whereby the total analysis is performed in less than 7 min.

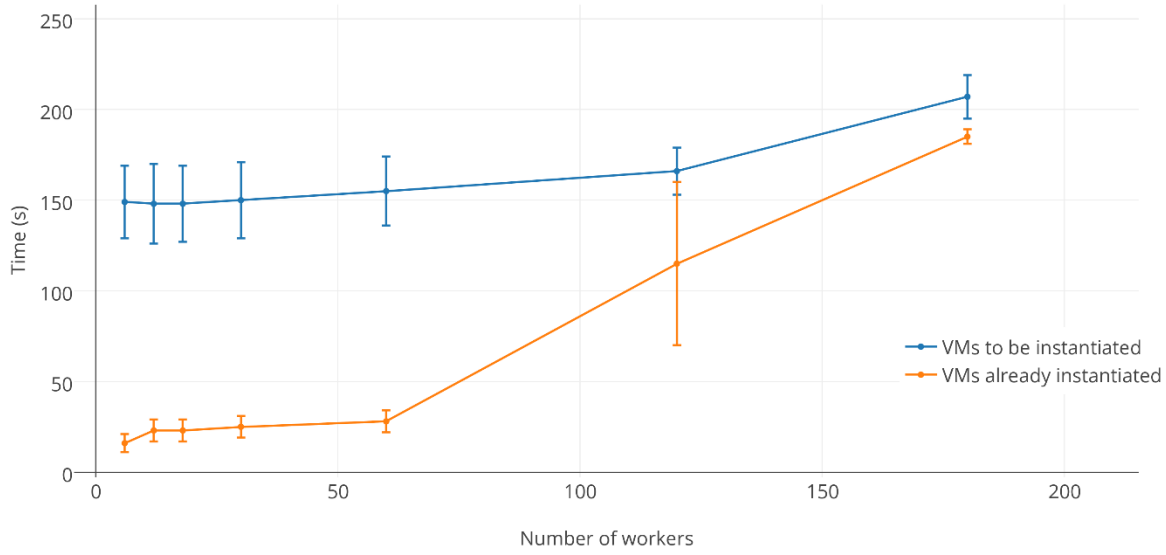

Fig. 1: Deploy time of VAF Worker Nodes vs number of workers in Torino site.

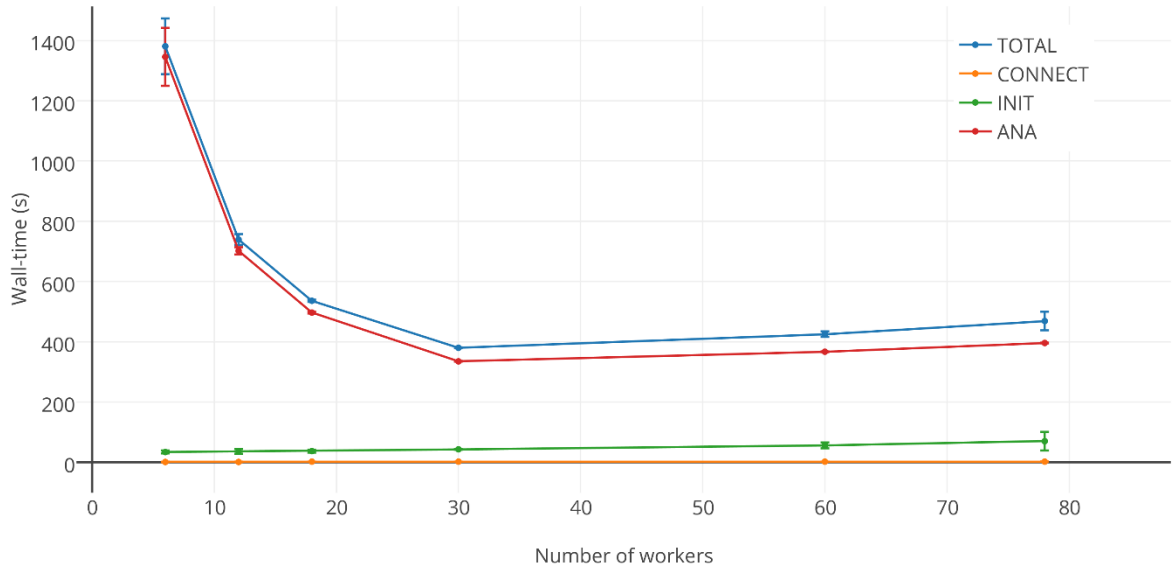

Fig. 2: PROOF analysis benchmark vs number of workers in Torino site.

The results shown in the above figures are used as reference to compare the performances of the VAF's built in different cloud infrastructures with different Cloud Management tools, i.e. OpenNebula and OpenStack and to test the different local and remote data access.

5. Monitoring

A side activity, mostly carried out at the Torino site, is the development of a monitoring, accounting and billing infrastructure able to consolidate data from all levels of the stack, from the IaaS up to the application. The system is based on the ElasticSearch ecosystem, composed by ElasticSearch (ES), Logstash and Kibana and generally referred to as the 'ELK stack'.

ES is a document-driven search and analytics engine built on top of the Apache Lucene information retrieval library. It allows for full-text search on unstructured data and can be interfaced with any RESTful API using JSON over http. Logstash is an open source tool used to collect and parse events and logs to a central service, that was integrated in the system by developing a simple plugin to retrieve input data from a MySQL database and a set of configuration files to customise data indexing (one for each application to be monitored). Finally, Kibana is a GUI that can be used to display and search ES data.

In order to monitor the Virtual Analysis Facility application, the system relies on the PROOF plugin TProofMonSenderSQL for collecting accounting data from the facility. At the end of each user query, data are gathered and sent to the database with a standard MySQL client/server protocol. In this case, the complex string processing capabilities of ES allows monitoring some additional observables such as e.g. the number of workers, the specific datasets analysed (LHC period, run, etc.) or the number of events processed.

Further details on this work can be found in [11].

6. Data access and storage federation

The most critical aspect of interactive analysis is data access: since each analysis process is typically short and repeated many times, the time needed to access data is likely to become of the order of magnitude of the analysis duration itself using current data access models. This is deemed unacceptable for an interactive analysis. Another issue is data distribution: data are geographically distributed among different computing sites in the world, thus their effective availability and integrity are not under control of the computing site effectively hosting the analysis resources. The data access strategy proposed in present project consists of two different data access levels used by interactive analysis. Data which are critical because analyzed more often will be completely replicated on a local storage system, shared with grid applications that access the same data.

The main technology proposed in the present project to aggregate data from different storage disks and serve them to the local computing nodes is a distributed filesystem: examples are GlusterFS [5], which is very popular in cloud computing, Lustre [6] or GPFS [7], which are popular choices for businesses with large data sites because their stability and scalability and in HPC clusters. The same data will be accessed from outside the computing site using a standard interface like XROOTD [3], which is independent on top of the underlying filesystem. The decision to expose a standard interface might turn out convenient for an expansion towards a federated national model for interactive analysis.

XROOTD is a storage solution capable to export and to aggregate filesystems from a distributed set of hosts in an efficient way. XROOTD is meant to be a scalable and reliable storage solution. Scalability is achieved via a redirector architecture: a typical XROOTD setup is constituted by a head node, called

the redirector, and many XROOTD servers exporting their filesystems. The set of filesystems all together form a disk pool, and both the redirector and the disk servers need to be network accessible from the client. The redirector of each Italian VAF site is connected to a global manager node (*global redirector*) that can serve all the requests of files belonging to the whole Italian Distributed Storage Cluster (DSC) [18], called VAF Data Federation (DF). The present design of the Italian VAF DF is schematically illustrated in Fig. 3: the picture includes all the expected request and response steps among the different parts involved in the data handling, access and usage.

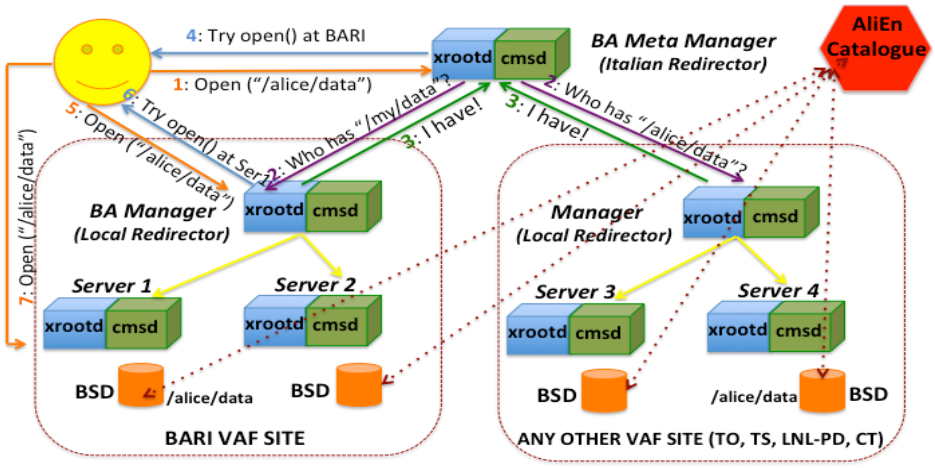

Fig. 3: Schematic picture of the design for the VAF Data Federation Cluster.

Such design has been preliminary implemented: in particular, to keep the architecture more flexible and reliable, all the nodes currently run on different VMs provided by the Bari PRISMA Openstack Infrastructure [17]. The first results obtained accessing data through DF indicate that it is possible to run the VAF benchmark analysis in less then 18 min, such results being limited by the saturation of the hardware resources currently available for the test of VAF DF (~1 Gbps bandwidth). The first comparison of the data access through VAF DF with the ALIEN data access [18] reflects the advantage of being connected by GARR-X network (at least 10 Gb/s) instead of accessing world wide distributed storage resources. A further gain could be explained by bypassing the latency of the AliEn Catalogue due to authentication and related queries to the DB.

7. Conclusions

The development of advanced resource management solutions, based on existing Cloud Management tools, has allowed the allocation of on-demand computing power for interactive use, optimizing the overall efficiency of the system. The parallelization of the computing activities has been accomplished by means of PROOF, that is an extension of the ROOT framework, widely used in the High Energy Physics community, explicitly conceived to this purpose. The proposed virtual infrastructure has been tested and validated with some use cases taken from the ALICE data analysis. The analysis facilities built in Italian ALICE Tier-2 sites have been federated with the purpose of reducing data duplication. The remote access to the data stored at the federated facilities is granted thanks to high throughput network connections provided by the Garr-X consortium by using and different tools like XROOTD and GlusterFS. The Cloud-based solutions can be easily extended to other research fields. Furthermore, the adoption of widely used software technologies and the exposure of industry-standard interfaces will allow the interoperability and the integration with future larger infrastructures.

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

References

[1] Aamodt K et al. (ALICE Collaboration) 2008 The ALICE experiment at the CERN LHC *JINST* **3** S08002
[2] Alunni Solestizi L et al. 2015 Improvements of LHC data analysis techniques at Italian WLCG sites. Case-study of the transfer of this technology to other research areas *these Proceedings*
[3] The eXtended Root Daemon (XROOTD) *http://xrootd.slac.stanford.edu/*
[4] Network File System (NFS) version 4 Protocol *http://www.ietf.org/rfc/rfc3530.txt*
[5] The GlusterFS *http://www.gluster.org/*
[6] The Lustre File System *http://lustre.opensfs.org/*
[7] IBM General Parallel File System (GPFS) *http://www03.ibm.com/systems/software/gpfs/*
[8] Ballintijn M, Brun R, Rademakers F, Roland G 2003 The PROOF distributed parallel analysis framework based on ROOT *arXiv preprint* arXiv:physics/0306110.
[9] Xu N, Guan W, Wu S L and Ganis G 2011 Data-oriented scheduling for PROOF *J. Phys.: Conf. Ser.* **331** 032009
[10] Bagnasco S 2015 Managing competing elastic Grid and Cloud scientific computing applications using OpenNebula *these Proceedings*
[11] Vallero S 2015 Integrated Monitoring-as-a-service for Scientific Computing Cloud applications using the ElasticSearch ecosystem *these Proceedings*
[12] Berzano D 2013 A ground-up approach to High-Throughput Cloud Computing in High-Energy Physics *PhD Thesis* (University of Torino)
[13] Bagnasco D et al. 2014 Managing a Tier-2 computer centre with a private cloud infrastructure *accepted for publication in J. Phys.: Conf. Ser.*
[14] Blomer J et al. 2014 Micro-CernVM: Slashing the Cost of Building and Deploying Virtual Machines *accepted for publication in J. Phys.: Conf. Ser.* arXiv:1311.2426 [cs.DC]
[15] The HTCondor distributed computing system *http://research.cs.wisc.edu/htcondor/*
[16] Berzano D et al. 2014 PROOF as a Service on the Cloud: a Virtual Analysis Facility based on the CernVM ecosystem *accepted for publication in J. Phys.: Conf. Ser.* arXiv:1402.4623 [cs.DC].
[17] *http://recas.ba.infn.it/recas1/index.php/recas-prisma*
[18] Colamaria F et al. 2015 Local storage federation through XRootD architecture for interactive distributed analysis *these Proceedings*