PROCEEDINGS OF SCIENCE

Elastic Computing from Grid sites to External Clouds

G. Codispoti[*]
INFN and Università di Bologna, Bologna, Italy
E-mail: Giuseppe.Codispoti@bo.infn.it

R. Di Maria
INFN and Università di Bologna, now at Imperial College, London, UK
E-mail: r.di-maria15@imperial.ac.uk

C. Aiftimiei
CNAF, Bologna, Italy and IFIN-HH, Magurele, Romania
E-mail: Cristina.Aiftimiei@cnaf.infn.it

D. Bonacorsi
INFN and Università di Bologna, Bologna, Italy
E-mail: Daniele.Bonacorsi@bo.infn.it

P. Calligola
INFN , Bologna, Italy
E-mail: Patrizia.Calligola@bo.infn.it

V. Ciaschini
CNAF, Bologna, Italy
E-mail: Vincenzo.Ciaschini@cnaf.infn.it

A. Costantini
CNAF, Bologna, Italy
E-mail: Alessandro.Costantini@cnaf.infn.it

D. DeGirolamo
CNAF, Bologna, Italy
E-mail: Donato.DeGirolamo@cnaf.infn.it

S. Dal Pra
CNAF, Bologna, Italy
Stefano.DalPra@cnaf.infn.it

C. Grandi
INFN , Bologna, Italy
E-mail: Claudio.Grandi@bo.infn.it

D. Michelotto
CNAF, Bologna, Italy
E-mail: Diego.Michelotto@cnaf.infn.it

M. Panella

PoS(ISGC 2016)023

CNAF, Bologna, Italy
E-mail: Matteo.Panella@cnaf.infn.it

G. Peco
INFN , Bologna, Italy
E-mail: Gianluca.Peco@bo.infn.it

V. Sapunenko
INFN , Bologna, Italy
E-mail: Vladimir.Sapunenko@cnaf.infn.it

M. Sgaravatto
INFN Padova, Padova, Italy
Massimo.Sgaravatto@bo.infn.it

S. Taneja
CNAF, Bologna, Italy
E-mail: Sonia.Taneja@cnaf.infn.it

G. Zizzi
CNAF, Bologna, Italy
E-mail: Giovanni.Zizzi@cnaf.infn.it

LHC experiments are now in Run-II data taking and approaching new challenges in the operation of the computing facilities in future Runs. Despite having demonstrated to be able to sustain operations at scale during Run-I, it has become evident that the computing infrastructure for Run-II already is dimensioned to cope at most with the average amount of data recorded, and not for peak usage. The latter are frequent and may create large backlogs and have a direct impact on data reconstruction completion times, hence to data availability for physics analysis. Among others, the CMS experiment is exploring (since the first Long Shutdown period after Run-I) the access and utilisation of Cloud resources provided by external partners or commercial providers. In this work we present proof of concepts of the elastic extension of a CMS Tier-3 site in Bologna (Italy), on an external OpenStack infrastructure. We start from presenting the experience on a first work on the "Cloud Bursting" of a CMS Grid site using a novel LSF configuration to dynamically register new worker nodes. Then, we move to an even more recent work on a "Cloud Site as-a-Service" prototype, based on a more direct access/integration of OpenStack resources into the CMS workload management system. Results with real CMS workflows and future plans are also presented and discussed.

International Symposium on Grids and Clouds 2016
13-18 March 2016
Academia Sinica, Taipei, Taiwan

[*]Speaker.

1. Introduction

Modern high-energy physics experiments (and not only) require massive amount of resources. The way such resources are used is frequently in burst mode. The experiment activities face periods of resource usage which are relatively flat and easy to predict, interleaved with (frequent and shorter) peaks of production needs, where resource usage would greatly increase with respect to periods of "normal" usage. Such peak needs go much beyond the resources pledged by Grid sites and available to the experiments (e.g. for LHC experiment this process is steered by WLCG [1], the Worldwide LHC Computing Grid collaboration), and they need to be absorbed somehow since they cause excessively long job queues at sites.

Traditional scientific (non-commercial) computing centres may find it difficult to size themselves, as: *i)* they cannot be sized for peak usage; *ii)* they cannot easily acquire extra resources on demand to serve the use-cases of all the supported communities at the same time; *iii)* they cannot absorb the peak usage of the experiments without generating excessively long job queues. As a results, the WLCG sites [2] and the institutions participating to the activity of the experiments are working to implement dynamic resources provisioning mechanisms, i.e. access to the cloud resources provided by external partners or commercial providers.

INFN, the Italian funding agency, is supporting the exploration of this approach to enforce high-energy physics experiments to access Cloud resources in order to cope with the request peaks. In particular, a very fertile ground for exploration is Bologna (Italy), that hosts the INFN-CNAF Tier-1 center - supporting CMS but also more than 20 other VOs [1] - as well as a small and agile Tier-3 centre for the CMS experiment [4] at the LHC [5]. The CMS computing team in Bologna is very active on CMS computing operations exploiting resources accessible via cloud interfaces, as well as in Cloud-related R&D activities. The CMS Bologna cloud team worked with experts from CNAF and the INFN-Bologna team on designing and implementing two prototypes:

- a "Cloud bursting" prototype, i.e. a dynamic resource provisioning mechanism to extend an existing Grid site's LAN-based batch system to other external resources (both inside the INFN-CNAF Tier1 domain, and towards external cloud resources using a CNAF Openstack [6] set-up);
- a "Cloud Site as-a-Service" prototype, i.e. the definition of a brand new independent site which is accessed via the standard workload management tools of the CMS experiment [7].

Both prototypes are presented and discussed in details in the following sections.

2. Cloud bursting: extending existing site queues

The main idea is to enable the dynamic extension of a batch system working inside a LAN (e.g. LSF [8]) to resources that would reside out of the LAN.

The main problem was that systems like LSF do not support dynamic extensions of the batch queues, hence most sites end up to be equipped with a batch system that is able to work just inside the LAN, and this case is not thought to allow any dynamic extension by default. In order to be

[1] Virtual Organization of WLCG, for more details see [3]

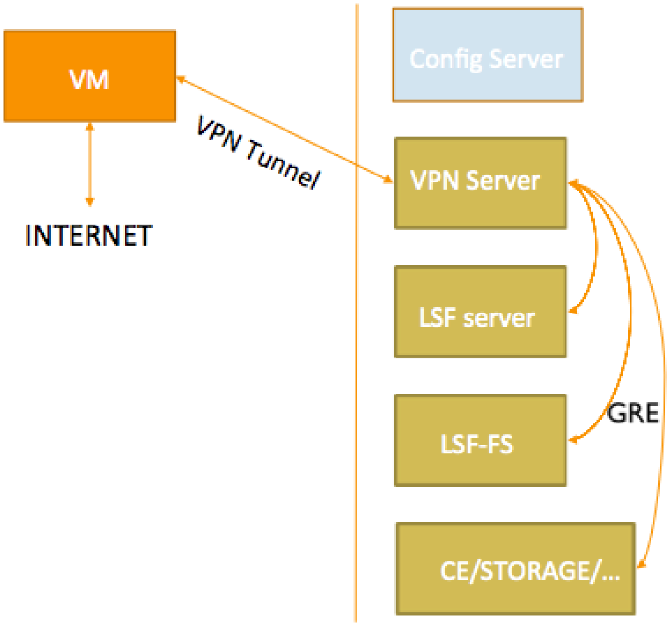


Figure 1: VPN configuration. See text for details.

able to extend an existing site into external resources one has to think of different mechanisms (e.g. LSF is bound to the host names, all small details that are lost when you enter a cloud environment where you have local address, local names, unique IDs, etc). The Bologna team created a system based on a VPN [9] where the new nodes can be added dynamically and hence be seen by the LSF master. In this way, there are no requests on the hypervisor, so the virtual machine (VM) can run everywhere, cloud providers included. The only request on the VM is the installation of just 2 additional RPMs for configurations. The VPN serves just the interaction between the nodes and the other part of the system, but the remaining traffic does not go through the VPN, hence the traffic is indeed reduced to the minimum.

The VPN configuration is shown in Figure 1. The basis is the configuration server. When a VM boots it contacts the configuration server and retrieves all the configurations and addresses of all the other services of the system. Basically, the first thing the VM does is to connect to a VPN server, be registered on the VPN. At this stage, a set of tools is enabled to allow the VM to talk to the LSF server and the LSF file systems where all the LSF configurations are stored (the CE, the SE, and any other element needed on the site). As soon as the configuration is complete, the LSF master sees the new nodes and starts sending the jobs transparently with no need for any additional actions. The full process requires 3 steps. The first step was the virtualization: a trivial step, for which custom lightweight images were used, relying wherever possible on remote service like CVMFS [10] and avoiding to install software locally (i.e. no EMI [2] Grid middleware). Wherever possible, the Tier-3 configurations were used, and access to the GPFS [11] Tier-3 storage was tested too. The second step was to extend the Bologna local farm, i.e. add static nodes to the

[2] European Middleware Initiative

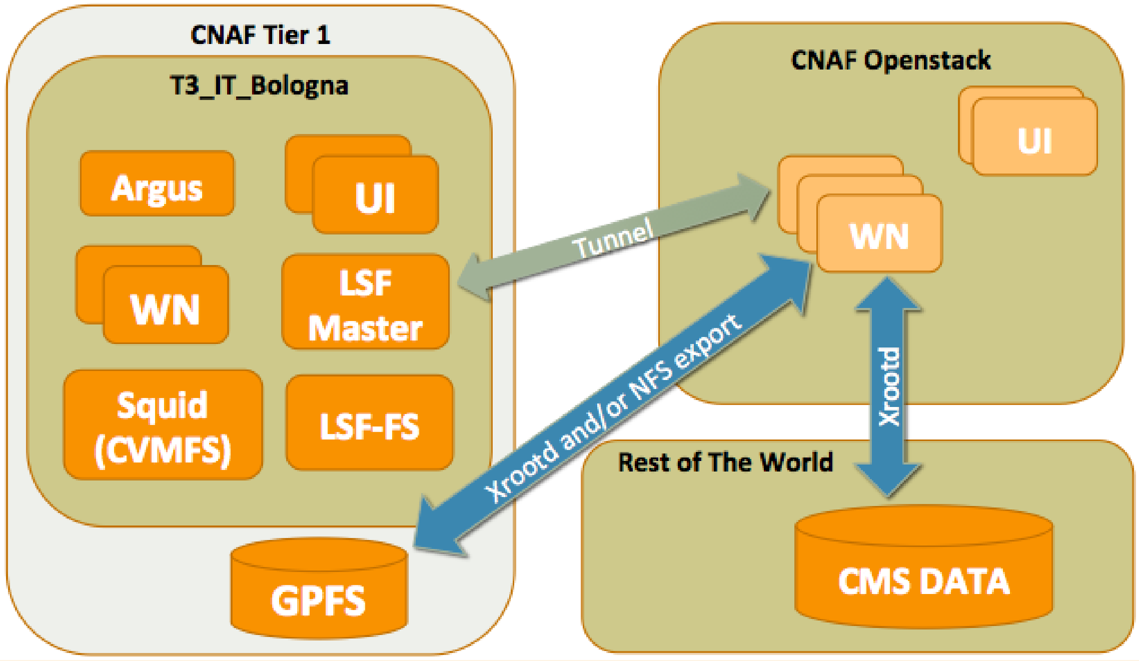
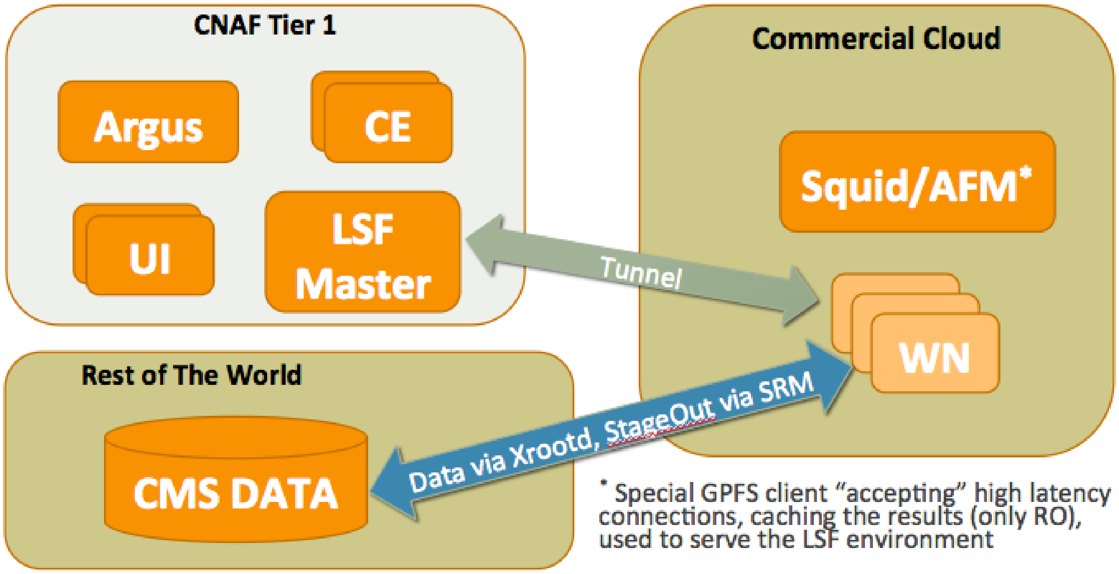


Figure 2: Extended Bologna Tier-3 configuration. See text for details.

farm and accessing them through a test LSF queue. This step allowed to test the LSF dynamic extension component of the prototype. The third step was to extend the full Tier-3 Grid site to CNAF Openstack, i.e. plug VM instantiated on Openstack into a Grid production queue. This step allowed to move to a Grid production queue and send jobs via Grid and see them running on a separate infrastructure, where the VMs were instantiated via the CNAF Openstack.

The extended Tier3 configuration is displayed in Figure 2. All the standard services needed by a Grid site are shown. This site is co-located to the Tier-1, i.e. it lives inside the domain of the Tier-1 itself, where you also have the GPFS storage system. In Openstack, extra VMs were instantiated to talk to the LSF master through a tunnel and can access the local data ("local" meaning local to CNAF) either through a GPFS export via NFS, or directly via xrootd [12]. The data access from the WAN is implemented just through xrootd.

The "Cloud bursting" prototype was successfully implemented and tested over the CNAF Openstack infrastructure (both Havana and Juno). The exploration of the access to the local storage (GPFS) through NFS export was the only part of the prototype that revealed issues, as this was not surprisingly far from ideal solution. The NFS export showed up as a low-performance bottleneck for the VMs and the GPFS system as a whole, and suggested to switch back to remote data access approach (xrootd, SRM). Apart from this, all was smooth: new nodes could be seen as "normal" Tier-3 nodes from Grid submission, they could be inserted into the official CMS production queues and the CMS workload management tools saw them transparently. A total of more than 3000 CMS jobs (real CMS workflows for Analysis Objects creation) were submitted, and the jobs spread smoothly between the normal (i.e. Grid) physical worker nodes and the newly plugged (i.e. Cloud) VMs. About 5% of the total number of submitted jobs reached the VMs, which was



* Special GPFS client "accepting" high latency connections, caching the results (only RO), used to serve the LSF environment

Figure 3: An example of extension of the INFN-CNAF Tier-1 over commercial resources, based on the prototype presented in this paper. More details in [13].

expected given the limited amount of resources that was actually instantiated in Openstack: hence, a good balancing of the jobs split among Grid and Cloud resources was observed. Additionally, no jobs failures were observed: the test had an admittedly more protected environment in Openstack (i.e. less concurrency with activities from other experiments), but still the results was positive in terms of functionality demonstration, and encouraging for the next steps in the program of work.

As a side note, in such a proof-of-concept we focussed on functionality demonstrations more than anything else, including detailed measurements or even scale tests application. But it is worth noting that this prototype already served much larger scale projects, also presented in this conference [13]. See e.g. Figure 3: a very similar configuration to the one described above can be (and was) used to try to extend a bigger site - like the INFN-CNAF Tier-1 - to commercial cloud providers. In order to do so, the main difference with respect to the original prototype developed and implemented by the CMS Bologna team and described here, is that we have to face a higher latency connection between the two sites (the Tier-1 and the Cloud), so some tools to cache the LSF file system in order to reduce the latencies is needed: this is addressed through a specific GPFS client which is called AFM. All the details on how the proof-of-concept of this paper has been exploited to address and solve a larger scale need is reported and discussed in Ref. [13].

3. Bologna Tier-3: a "Cloud site as-a-Service"

Apart from extending an existing site towards Cloud resources, the CMS Bologna team worked on a second prototype. The idea was to build a brand new site from scratch, completely decoupled from the existing Grid one. This site would be a fully-decoupled CMS site "as-a-service" in Openstack. It is eventually registered in the CMS workload management system (i.e. called "T3_IT_BolognaCloud" in CMS jargon), so it can be accessed via the standard CMS workload

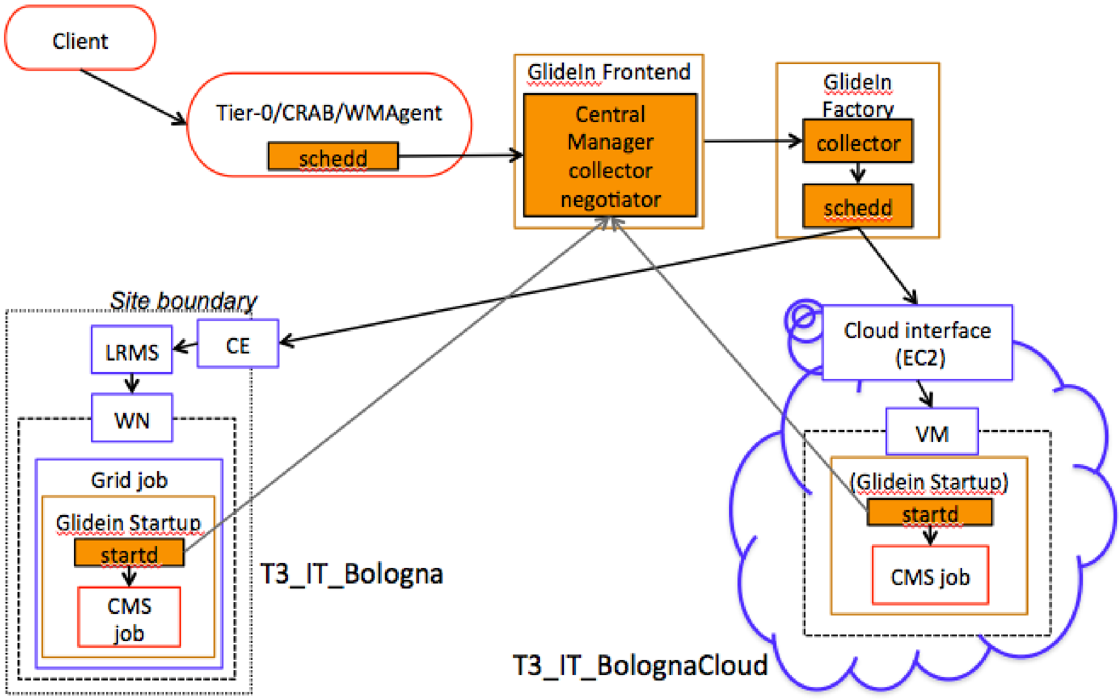


Figure 4: A pictorial view of the set-up of the CMS Bologna Cloud site “as-a-service” in Openstack. See text for details.

management tools (WMAgent [14] for scheduled production job submissions, CRAB [15] for unpredictable user analysis job submissions). But it is not registered as a WLCG resource, as it may have a very short life.

As a first step, we exploited a chain consisting of a former CRAB version (namely, CRAB v2) plus a custom GlideIn-WMS instance plus the standard Openstack Havana infrastructure. CRAB v2 was exploited here as it was a pure command line tool i.e. very easy to customise for our needs (e.g. tuned to work with the custom GlideIn-WMS instance). This first test allowed to get ready to moving to a standard production environment in a second test, i.e. a chain consisting of the currently latest CRAB version (namely, CRAB v3 [16]) plus the CERN Integration Testbed (ITB) plus the GlideIn-WMS infrastructure plus the standard CNAF Openstack Juno system. A peculiarity of this test was that we moved as much as possible towards the tools used in the real production environment. CRAB v3 has evolved to be a client-server infrastructure which is bound to a production GlideIn-WMS service. The ITB GlideIn-WMS is the central infrastructure for pre-production tests in CMS: it is in no way different from the production one, and it was used to avoid interference effects with any other users during the test. This exercise was also one of the first main test for the newly Openstack Juno CNAF installation, and definitely the first test in CMS.

The set-up of the CMS Bologna Cloud site “as-a-service” in Openstack is depicted in Figure 4. A client - that can be e.g. a CMS user submitting analysis jobs to the CRAB system - sends jobs to the GlideInWMS. This is made of two components at this point, the factory and the front-end. The factory submits the pilot jobs - in case of standard WLCG sites - to the Computing Element

(CE) of the site, then the pilots are queued in the batch system. When they start running, they initialise a condor [17] daemon that starts fetching jobs from the GlideInWMS front-end. The main difference with such a cloud approach is that the GlideInWMS does not submit the pilots any more, but it interacts with the EC2 [18] interface and asks for the creation of VMs. At boot, the VMs already start with condor_startd and start fetching jobs. All is needed is to properly configure the GlideInWMS (and have the proper packages on the VM, of course). The same GlideInWMS can deal with both WLCG sites and Cloud sites.

In summary, the Bologna Tier-3 set-up was used to instantiate a new CMS site "as a service" which actually is CNAF Openstack resources. Custom, "lightweight" images were used also in this case (basically, the only difference with the previous case was that one needs to eventually remove what was not needed and add the few packages needed for starting the condor queues). There is no need to actually create a new site in WLCG: only GlideIn-WMS need to be aware, which give a high flexibility in instantiating any needed sites on-demand. The CMS Bologna cloud team managed to use the full CMS workload management production infrastructure, and uses it with standard CMS workflows: final Analysis Objects creation was chosen, and in particular more CPU-intensive tasks used by CMS for the simulation of the upgrdaded detector have been selected (at the cost of just a little tuning of the flavour of the VMs, from Quadcore 8GB RAM VMs towards Quadcore 12GB RAM VMs). As the infrastructure was shared with other customers, some limitations had to be taken into account in the maximum number of jobs submitted; nevertheless, 4 tasks of 200 jobs each were submitted, no failures were observed, and a very good job CPU efficiency (defined as CPT/WCT) was observed, peaking at about 98% for most of the test time.

4. Possible evolutions of the project

Plans to build and evolve of the prototypes described in the previous sections are solid, and some have already become reality. A list, and a few remarks on each, in the following.

A first step would be to get rid of custom images, and move to something more standard, like the μCERNVM [19] images, already used by other LHC experiments and recently adapted also by CMS (generic ISO image 12-MB sized with OS entirely on CVMFS, faster to instantiate, easier to keep up-to-date).

A current limitation is that we tested the dynamic extension for the CMS-only case so far. The Bologna Tier-3 is a multi-VO environment, and expanding this work to other experiments would allow a general benefit out of the project described in this paper. As a starting target, we are currently working to extend the usability to ATLAS as well.

Using μCERNVM plus Parrot under Docker is another goal in front of us. It would allow to further reduce the requests on the host system, with no kernel privileged access.

The Bologna Tier-3 is a local users facility, and as such it can be costly in terms of maintenance and manpower. We plan to profit of this exercise to turn the Grid site into a purely Cloud site - as soon as other customer VOs are able to adapt to this technology. We can hence think of it as a pure cloud resource that lives inside the Tier-1, for instance, and can extend opportunistically onto the Tier-1 resources when they are not used. This will greatly benefit by reduce the experiment-specific personpower needs for the maintenance of the Tier-3.

5. Conclusions

The CMS Bologna cloud team, together with CNAF personnel and experts from the INFN-Bologna team, realised two prototypes for the extension of a Grid site into Cloud resources. The first prototype is a dynamic extension of a CMS site over external resources with production tools (the Bologna Tier-3 over INFN-CNAF Openstack), and the second prototype is the usability of a purely Cloud-instantiated CMS Site with CMS standard production and analysis tools. The functionalities of both prototypes have been demonstrated, and - despite at a limited scale to avoid interference with production activities on the same resources - quite some load for official CMS jobs was applied to the system to test it in real case scenarios, with very satisfactory results. We are looking forward to next challenges, among which the opportunity of defining the Tier-3 (at least for CMS) as a pure Cloud-instantiated site inside the INFN-CNAF Tier-1 Openstack infrastructure to reduce mantainance and operation costs.

References

[1] J. D. Shiers, *The Worldwide LHC Computing Grid*, Comp. Phys. Comm. 177 (2007) 219-223.

[2] D. Bonacorsi, *WLCG Service Challenges and Tiered architecture in the LHC era*, IFAE, Pavia, April 2006.

[3] http://wlcg.web.cern.ch/getting-started/VO

[4] The CMS Collaboration, *The CMS experiment at the CERN LHC*, JINST 3 (2008) S08004.

[5] Evans, Lyndon et al., *LHC Machine*, JINST 3 (2008) S08001.

[6] OpenStack, http://www.openstack.org/

[7] M. Cinquilli, D. Evans, S. Foulkes, D. Hufnagel, M. Mascheroni, M. Norman, Z. Maxa, A. Melo, S. Metson, H. Riahi, S. Ryu, D. Spiga, E. Vaandering, S. Wakefield, R. Wilkinson, *The CMS workload management system*, J. Phys.: Conf. Ser. 396 (2012) 032113.

[8] LSF, http://www-03.ibm.com/systems/services/platformcomputing/lsf.html

[9] OpenVPN software, http://openvpn.net/

[10] P. Buncic, C. Aguado Sanchez, J. Blomer, L. Franco, A. Harutyunian, P. Mato, Y. Yao, *CernVM - a virtual software appliance for LHC applications*, J. Phys.: Conf. Ser. 219 (2010) 042003.

[11] F. Schmuck, Frank, R. Haskin, *GPFS: A Shared-Disk File System for Large Computing Clusters*, Proceedings of the FAST'02 Conference on File and Storage Technologies. Monterey, California, USA: USENIX. pp. 231âĂŞ244. ISBN 1-880446-03-0.

[12] J. Andreeva et al., *Monitoring of large-scale federated data storage: XRootD and beyond*, J. Phys.: Conf. Ser. 513 (2014) 032004.

[13] S. Dal Pra et al., *Elastic CNAF DataCenter extension via opportunistic resources*, this Conference.

[14] E. Fajardo et al., *A New Era for Central Processing and Production in CMS CMS Collaboration*, J.Phys.Conf.Ser. 396 (2012) 042018, DOI: 10.1088/1742-6596/396/4/042018.

[15] Giuseppe Codispoti et al., *CRAB: A CMS Application for Distributed Analysis*, IEEE Trans. Nucl. Sci. 56 (2009) 2850-2858, DOI:10.1109/TNS.2009.2028076.

[16] M. Mascheroni et al., *CMS distributed data analysis with CRAB3*, J. Phys.: Conf. Ser. 664 (2015) 062038.

[17] HTCondor, http://research.cs.wisc.edu/htcondor/

[18] Amazon Elastic Compute Cloud, http://aws.amazon.com/ec2/

[19] J. Blomer et al., *Micro-CernVM: slashing the cost of building and deploying virtual machines*, J.Phys.Conf.Ser. 513 (2014) 032009, DOI: 10.1088/1742-6596/513/3/032009.