PROCEEDINGS OF SCIENCE

Accessing Scientific Applications through the WNoDeS Cloud Virtualization Framework

Elisabetta Ronchieri
INFN CNAF
Viale Berti Pichat 6/2, I-40127 Bologna, Italy
E-mail: elisabetta.ronchieri@cnaf.infn.it

Marco Verlato[1]
INFN, Sezione di Padova
Via Marzolo 8, I-35131 Padova, Italy
E-mail: marco.verlato@pd.infn.it

Davide Salomoni
INFN CNAF
Viale Berti Pichat 6/2, I-40127 Bologna, Italy
E-mail: davide.salomoni@cnaf.infn.it

Gianni Dalla Torre
INFN CNAF
Viale Berti Pichat 6/2, I-40127 Bologna, Italy
E-mail: gianni.dallatorre@cnaf.infn.it

Alessandro Italiano
INFN, Sezione di Bari
Bari, Italy
E-mail: alessandro.italiano@ba.infn.it

Vincenzo Ciaschini
INFN CNAF
Viale Berti Pichat 6/2, I-40127 Bologna, Italy
E-mail: vincenzo.ciaschini@cnaf.infn.it

Daniele Andreotti
INFN CNAF
Viale Berti Pichat 6/2, I-40127 Bologna, Italy

[1] Speaker

© Copyright owned by the author(s) under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike Licence.

PoS(ISGC 2013)029

PROCEEDINGS OF SCIENCE

E-mail: daniele.andreotti@cnaf.infn.it

Stefano Dal Pra
INFN CNAF
Viale Berti Pichat 6/2, I-40127 Bologna, Italy
E-mail: stefano.dalpra@cnaf.infn.it

Wouter Geert Touw
Centre for Molecular and Biomolecular Informatics (CMBI), Nijmegen Centre for Molecular Life Sciences (NCMLS), Radboud University Nijmegen Medical Centre
6525 GA 26 Nijmegen, The Netherlands
E-mail: w.touw@cmbi.ru.nl

Gert Vriend
Centre for Molecular and Biomolecular Informatics (CMBI), Nijmegen Centre for Molecular Life Sciences (NCMLS), Radboud University Nijmegen Medical Centre
6525 GA 26 Nijmegen, The Netherlands
E-mail: vriend@cmbi.ru.nl

Geerten W. Vuister
Department of Biochemistry, University of Leicester, Henry Wellcome Building
Lancaster Road, Leicester LE1 9HN, UK
E-mail: gv29@leicester.ac.uk

Scientific applications are developed to run in specific computing environments and use several infrastructures that provide low-level resources, such as storage, networks, and fundamental computing resources, to the users who do not have knowledge of their exact location and hardware specifics. A framework is needed to scale-up or scale-out applications dynamically on provisioned resources. WNoDeS is a cloud and grid virtualization framework that introduces cloud-based services and supports virtual machines for local and grid users alike. This paper presents the access to scientific applications of different user communities, such as astro-particle physics with the Auger experiment, and life science with WeNMR virtual research community, through this framework. It details both use cases that had to be addressed for the described communities and further requirements that WNoDeS is able to achieve. The paper also shows how direct instantiation of virtual machines can be fulfilled by using both an OCCI-compliant CLI and a Web portal.

The International Symposium on Grids and Clouds (ISGC) 2013
March 17-22, 2013
Academia Sinica, Taipei, Taiwan

© Copyright owned by the author(s) under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike Licence.

PoS(ISGC 2013)029

1. Introduction

WNoDeS is a virtualization framework to provision cloud and grid resources by using standard scheduling mechanisms such as LSF and Torque/Maui and KVM virtualization technology [1]. Amongst its peculiarities [2] WNoDeS is able to dynamically instantiate Virtual Machines (VMs) on-demand; customize VMs in accordance with the site policy; interact with users' requests through traditional batch or grid jobs and also cloud interface to allocate compute and storage resources; and virtualize cluster resources in medium or large scale computing centers.

WNoDeS integrates grid and cloud provisioning through virtualization [3]. Resource brokering is done through a tight integration with an LRMS, allowing flexible policies to access resources; since these are allocated through an LRMS regardless of the requested interface, sites can enrich their offerings with IaaS provisioning exploiting current assets, know-how and tools. Also, static partitioning can be avoided, allowing optimization of resource utilization. Grid jobs can be run on VMs using a GLUE Schema attribute to specify virtual image name and characteristics such as RAM or cores. For VM instantiation an API (being made compliant to the OCCI 1.1 specification[2]) and a CLI are provided. Cloud provisioning is also integrated with a general purpose, web-based portal for the integration of grid and cloud resources, developed and maintained by the Italian Grid Initiative (IGI)[3]. A WNoDeS feature called mixed mode allows using physical resources both as traditional batch nodes and, at the same time, as hypervisors for VM instantiation. This lets sites to introduce features such as VM and cloud computing support on traditional resources without disrupting existing services and allows for efficiently deciding which workloads are to be virtualized and which should be run instead on non-virtual hardware.

The paper shows how WNoDeS has been used to support some real use cases: how a virtual image is managed, how a user can ask that jobs be executed on a VM running a given image, and how WNoDeS can run jobs on custom VMs for the WeNMR project. They have been implemented to support the Auger experiment through an adaptation of their experiment framework involving DBs accessed from and encapsulated in VMs; to run WeNMR CING (Common Interface for NMR structure Generation) software that has many external dependencies on a virtual set up. The paper describes how the Virtual CING machine is instantiated and how a selection mechanism allows WeNMR users to run jobs on the VM.

The paper is organized as follows. Section 2 introduces the WNoDeS framework, whilst Section 3 briefly presents mixed mode. Section 4 describes some examples of use of WNoDeS. Section 5 details future features of WNoDeS in the short and long terms, and finally Section 6 provides a conclusion.

[2] OCCI 1.1, http://occi-wg.org/2011/01/31/occi-1-1-document-series-in-public-comment/

[3] Welcome to IGI, Italian Grid Infrastructure, http://www.italiangrid.it/

PoS(ISGC 2013)029

2.WNoDeS Framework

The WNoDeS framework, as presented in Figure 1, can be conceptually divided into 5 main layers.

At the lowest level, the Virtual Resources level, there is the "site-specific" module, which is designed to be run on the VM itself to make it WNoDeS aware.

At a level over that there are the "bait" and "hypervisor" modules. The bait is WNoDeS interface to the batch system, and its duty is to translate requests for job execution into requests for virtual machine creation and then move the actual job to the newly created VM, and does this with the help of another instance of the site-specific module which redirects batch system requests to the bait itself. The hypervisor is the module that actually deals with the low-level detail of VM management (such as creation, configuration, and destruction) and works on top of an existing virtualization technology. This is what is called the "Business Logic" layer.

Above that there is the "Information Management" layer, which contains the "nameserver" and "manager" modules. The nameserver module keeps track of the set of virtual machines allocated to WNoDeS and assigns them to the hypervisor module when a new one is needed. It also keeps track of the configurations for the bait and hypervisor. The manager is a CLI component that is used by a system administrator to send commands to specific parts of the WNoDeS infrastructure for tuning.

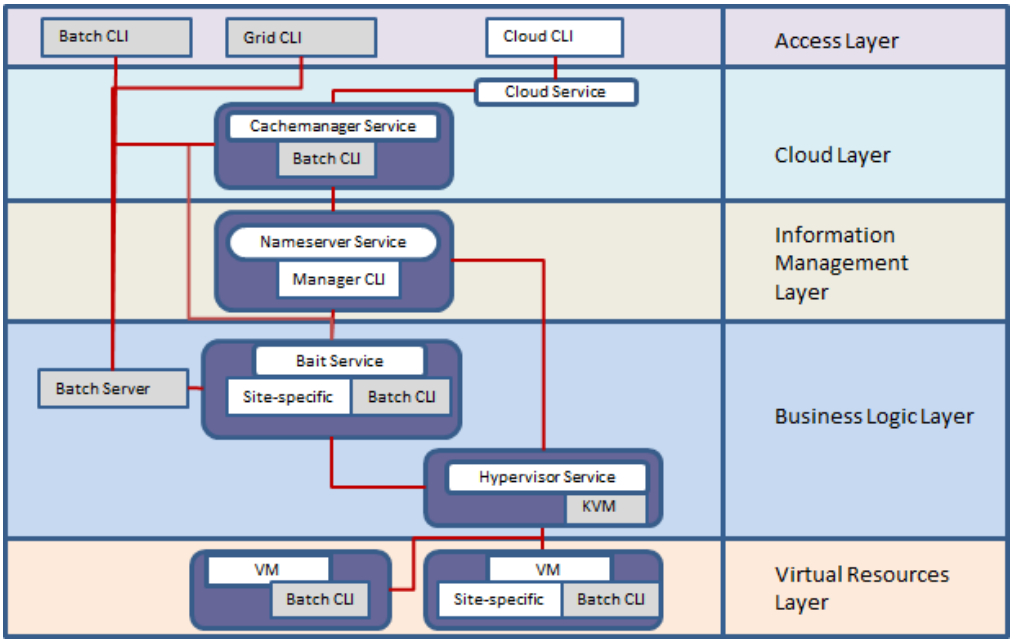

Figure 1: The WNoDeS core.

The fourth layer is the so-called "Cloud" Layer. At his level there are the interfaces used to request direct instantiation of a VM, which would be directly accessible to the user, rather than

PoS(ISGC 2013)029

being used to generate VMs to run batch system jobs. Two such interfaces are provided: "Cloud Service", which is a HTTP REST binding of OCCI 1.1, and a proprietary protocol implemented by "Cachemanager Service."

Finally, there is the "Cloud CLI" on a fifth "Access" Layer, which provides a CLI that can send commands to the cloud layer through the OCCI interface.

There are two common flows throughout this infrastructure: through jobs sent by the batch system to a WNoDeS-controlled farm or sub-farm or through user requests to allocate specific virtual machines for their use. The former flow is the following: when a new job arrives to a WNoDeS-controlled WN, the site-specific component first redirects the request to the bait service. The bait will then send a new request to the hypervisor to allocate a new virtual machine, and when the allocation is completed redirects the job to the new VM. After the job execution ends, the site-specific component handles the destruction of the VM after the job results have been sent back. In the latter case, the user sends a request for a VM creation, specifying parameters like memory, CPUs and disk and most importantly the user's SSH public key, to the OCCI interface or directly to the cachemanager service. Such a request then gets translated into a job for a VM with the parameters specified and a "sleep" command as payload and sent to the batch system; when this request reaches the bait component, the virtual machine allocation request is forwarded to the hypervisor. The cachemanager monitors this process, and as soon as the VM allocation is completed, copies the user's SSH public key into it and reports the name of the VM to the user, which can the login into it. The VM is then destroyed when the sleep job ends or when the user explicitly requests it destruction, whichever happens first.

3. Mixed Mode Feature

The Mixed Mode [3] is one of the main functionalities of WNoDeS. It was introduced to make it able to handle farm physical computing resources both as traditional member nodes of a batch system to execute jobs and as hypervisor node to instantiate VMs. This way a physical node can host at the same time both traditional jobs, directly running in the real machine, and also "virtual" ones, running in a virtual node instantiated there by the hypervisor.

This feature is optional and allows a computing centre to optimize the usage of its resources by integrating "real" and "virtual" resources. As in the WNoDeS deployment with mixed mode turned off, the instantiated VMs can be used to execute batch jobs, to perform analysis or to run applications. Through this functionality it is possible to execute jobs having special resources need, such as GPU, or higher I/O requirements directly on the physical node, thus avoiding the slowdown introduced by the virtualization overhead. Moreover, on the same physical node that makes available enough memory, CPUs and disk, the mixed mode can also offer virtualized services for those users requiring them. The mixed mode mechanism handles users' requests in collaboration with the batch system (see Figure 2). These can be traditional batch jobs requests, submitted locally or via grid: in this case by using the site-specific configuration file, the batch job can be executed either on the physical node or on the VM.

PoS(ISGC 2013)029

There can be also cloud provisioning requests, submitted via the Cloud CLI: in this case the provisioning will only be furnished by a VM. WNoDeS translates these users' requests into jobs that are handled by the batch system.

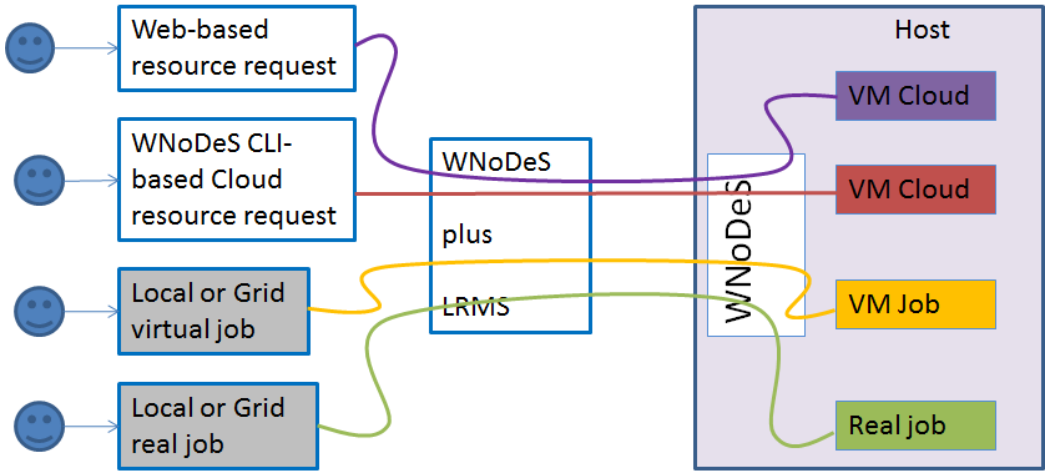

Figure 2: WNoDeS supported requests.

This Mixed mode has been available since WNoDeS 2.0.0-2 in the EMI-2 Matterhorn distribution [4] whilst the Cloud CLI that interacts with the WNoDeS cloud platform has been released since the version WNoDeS 3.0.0-1 in the EMI-3 Monte Bianco distribution [4].

4. Applications Examples

Scientists have learned to rely on the grid as a reliable and inexpensive way to process massive amounts of data. Starting from this common experience, which also involves existing computer centers, Auger and WeNMR over WNoDeS, are intended to provide grid and cloud services based on the proven reliability of the LRMS and overcome its limitations in terms of simplicity of access and configuration for the users, pre-packing the software and configuration over virtual machine images.

4.1 INFN Tier-1

The INFN Tier-1 computing centre (CNAF, Bologna) manages a WNoDeS production instance on top of the IBM/Platform LSF batch system and IBM GPFS storage disk since November 2009. It works in a fully integrated fashion with the overall number of 13 thousand cores of the farm, all belonging to a single LSF cluster. At the time of this writing, almost one thousand VMs can currently be instantiated on-demand out of this common farm to support cloud computing, or custom executing environments.

PoS(ISGC 2013)029

At INFN Tier-1 a certain number of Virtual Organizations (VOs)[4] have been configured to use WNoDeS during the computation of their jobs. Furthermore, the site adopts mixed mode in order to use physical resources both as traditional batch nodes and, at the same time, as HVs for VM instantiation. This lets sites to introduce features like the support of VM and cloud computing on traditional resources without disrupting existing services and allowing to efficiently deciding which workloads are to be virtualized and which should be run instead on top of non-virtual hardware.

With the mixed mode turned on, any node of a farm may act as a traditional, "real" node (that only runs traditional batch jobs), as a pure hypervisor (that only runs VMs), or as a node able of running, at the same time, both jobs on the physical hardware (without thus incurring penalties typical of virtual infrastructures), and jobs or cloud services on VMs (thus exploiting capabilities offered by virtual environments). Therefore, mixed mode greatly enhances the flexibility of the configuration of a farm and lets system administrators progressively integrated WNoDeS into existing data centres.

4.2 Astro-Particle Physics Community

The Pierre Auger Cosmic Ray Observatory [5] is a cosmic ray observatory located in Argentina that studies ultra-high energy cosmic rays, the most energetic and rarest of particles in the universe. Scientists involved in the Pierre Auger Observatory deal with so far unknown sources of cosmic rays that impact our planet. When these particles strike the earth's atmosphere, they produce extensive air showers made of billions of secondary particles. While much progress has been made in nearly a century of research in understanding cosmic rays with low to moderate energies, those with extremely high energies remain mysterious. These rays have energy many times greater than what we can achieve in our largest particle accelerators. The scientists measure showers of particles caused by cosmic rays using detectors deployed across three thousand square kilometers of the Pampas of Argentina: they require massive computational power for simulation in order to compare readings with theoretical models.

The Auger's computational model requires concurrent read only access to a condition data base (MySQL) from hundreds of computing nodes in the farm. The data base resides in a networked filesystem (IBM/GPFS) and may be concurrently accessed by one or more MySQL engines. This of course is only possible in the case of read only access. Being a single database engine ineffective at serving hundreds of clients, an opposite solution was initially preferred by Auger: a database engine would have been available at every node, each one independently accessing the filesystem with its data base. This approach however lead again to poor performances because of the high rate of non-serialized and concurrent disk accesses that causes overload on the GPFS side.

[4] VOs supported at INFN Tier-1 computing centre using WNoDeS are alice, ams, biomed, cms, lhcb, pamela and of course auger_db, enmr.eu. In the fedcloud context (https://wiki.egi.eu/wiki/Fedcloud-tf:FederatedCloudsTaskForce) the WNoDeS deployment also supports: dteam, fedcloud, gridit, ops, testers.eu-emi.

PoS(ISGC 2013)029

A sustainable solution has been successfully implemented at INFN Tier-1, by running the Auger's jobs on Virtual WNs managed by WNoDeS. Each one of these selects and retrieves its needed data from a MySQL engine installed on the Hypervisor where the virtual WN have been instantiated. In so doing, the number of needed database engines is reduced by a factor equal to the number of concurrent runnable jobs in a physical host, which in turn equals the number of cores of the hypervisor.

To further optimize this solution, a job packing configuration[5] has been recently investigated and tested, and is being applied to the LSF batch system of the INFN Tier-1. This is done to have the Auger's "virtual" jobs being concentrated on the smallest possible set of physical nodes, thus optimally reducing the number of running MySQL engines.

4.3 Life Science with WeNMR Virtual Research Community

WeNMR [6] is both a EU FP7 funded project and a Virtual Research Community (VRC) supporting the Nuclear Magnetic Resonance (NMR) and Small Angle X-ray Scattering (SAXS) structural biology user community. It currently operates the largest Virtual Organisation (VO) in life science domain of the European Grid Infrastructure (EGI), with more than 500 registered users worldwide.

While mainly offering application portals providing "protocolized" access to EGI grid, one of their use-case better fits with the cloud model. The Common Interface for NMR Structure Generation (CING [7]) is a software suite designed for the residue-based validation of biomolecular three-dimensional structures determined by NMR. The CING framework integrates about 20 different external programs and additionally uses internal routines to validate NMR-derived structure ensembles against empirical data and measured chemical shifts, distance and dihedral restraints. Due to the high number of external dependencies CING cannot easily be installed on traditional grid computing nodes. In contrast, the collection of programs could easily be packaged into a virtual machine that we called "VirtualCING". Therefore, VirtualCING is very suited to be used in a cloud environment. For example, it has been and still is successfully applied to establish the NRG-CING database, an annotated and remediated repository of experimental, structural, and validation data [8]. The import of one NMR structure into the database and the creation of a CING validation report on average require 20 min on a single core, taking about 3,300 core hours to process the current set of entries.

The experimental and computational procedures involved in the determination of biomolecular structures by NMR are continuously being developed and improved, and have become more advanced over the past 25 years. Therefore, by applying today's improved technology to the original data, better structures can be calculated [9, 10]. Recently, we extended VirtualCING with routines for the recalculation of NMR ensembles using the experimental data in the NRG-

[5] S. Dal Pra, "Job Packing: optimized configuration for job scheduling", presentation at HEPIX 2013, 15-19 April 2013, Bologna, Italy, http://indico.cern.ch/getFile.py/access?contribId=9&sessionId=5&resId=0&materialId=slides&confId=220443

PoS(ISGC 2013)029

CING repository, allowing us to extend our previous recalculation projects and redetermine NMR structures in the Protein Databank (PDB) [11] in the cloud on a large scale. The NMR_REDO project aims to improve quality of NMR-derived protein structure ensembles in terms of fit with the original experimental data and geometric quality. To date, 3,400 structures have been recalculated and we expect this number to increase to 5,000 of the 10,000 structures deposited in the PDB when our protocol can handle all usable deposited experimental data. A full NMR_REDO iteration currently consumes about 25,000 core hours. Preliminary results indicate that the recomputed ensembles generally show a better fit to the original data and to independent validation criteria than the original ensembles.

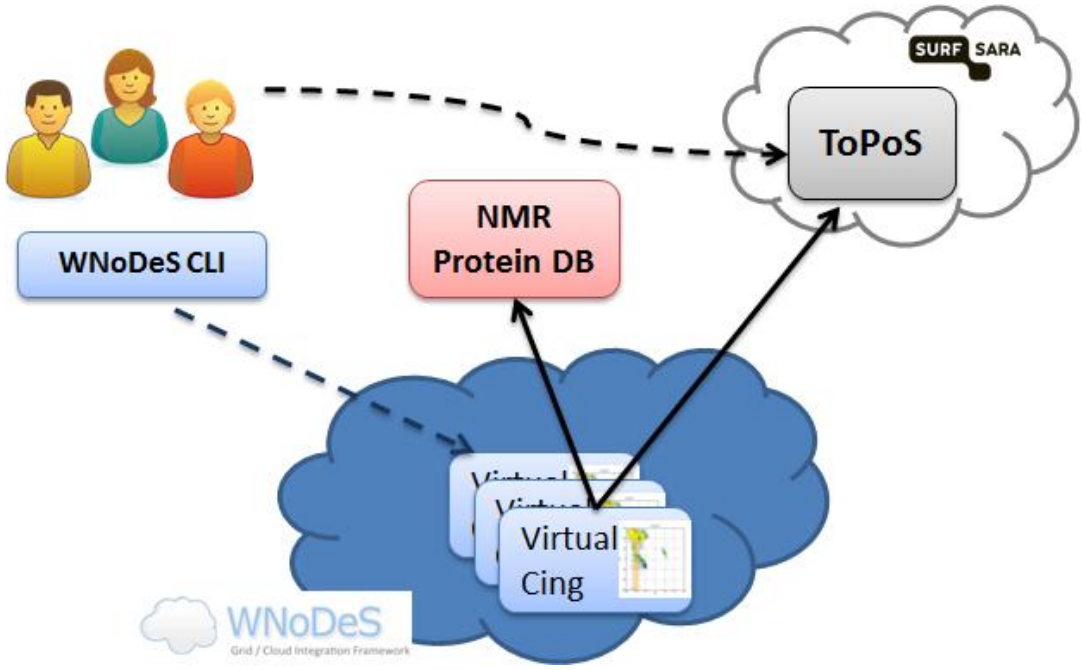

Figure 3: Architecture of the WNoDeS-based cloud infrastructure for the WeNMR/VirtualCING use case.

In order to verify how the WNoDeS cloud framework can enable scientists to perform NMR computations, a customized VirtualCING image has been deployed in the WNoDeS marketplace. The WeNMR user, after having created his own VOMS proxy certificate, instantiates a number of VirtualCING machines through the WNoDeS CLI. After booting, each VirtualCING automatically starts a process getting the job payload from the ToPoS token pool server[6], a pilot job framework offered through a HTTP server hosted by SURFsara organization, in The Netherlands. The tokens, previously uploaded on the ToPoS server by the user, contain the information about the proteins to be processed, the URL of the input data to be fetched from the NRG-CING Protein DB, the location of the NMR_REDO Protein DB where to upload the final output data (both DBs are represented by the pink box in Figure 3), and a set of parameters as arguments of the executable. At the end of the computation the token is deleted from the

[6] ToPoS, a Token Pool Server for Pilot Jobs, https://grid.sara.nl/wiki/index.php/Using_the_Grid/ToPoS

PoS(ISGC 2013)029

ToPoS server and the VirtualCING process asks for another unlocked token, until all tokens have been processed and nothing is left on the ToPoS server. The final results, for each recalculated protein, can be then visualized through the web interface of the NMR_REDO Protein DB[7].

5. Future Plan

Work on WNoDeS is still ongoing. Some specific improvements involve components of the Access and Cloud Layers that are already planned. In various stages of development they can be summarized in the following list:

1. Support for multiple images: in its current stage, the WNoDeS framework only supports one VM image. This is insufficient for a widespread usage of WNoDeS, and therefore support for the ability to choose the image to be instantiated from a set of pre-existing ones is under development.
2. Support for different cloud platform other than WNoDeS itself. The ability to integrate WNoDes as a backend for OpenStack and Opennebula, thus adding support for a real batch system to those system is interesting. A feasibility study is ongoing.
3. Integration in the IGI portal[8]: a fundamental feature to increase general usability of the system, which in its current incarnation can only be controlled from the command line, therefore putting very strong limits on ease of use and requiring moderately tech-save users. Integration in a Web portal will be a fundamental step in lowering the entry barrier to the system. This, too, is already under development.
4. Improvements to the HTML/REST rendering of the OCCI 1.1 protocol for a better compatibility with other implementations of OCCI.

The point 2 will consist of extending the "Cloud CLI" component in order to interact with several cloud platforms. Table 1 provides a short description of the coming commands.

Commands	**Description**
resource_providers_info	returns a list of providers registered in the BDII service.
size_images_info	returns for each site the images size of the adopted Cloud platform.
metadata_images_info	returns a list of the images metadata registered in the MarketPlace service.
create_instance	returns the instance location that satisfies the user's request in terms of the image size.
show_instance	shows information for the specified instance location such as the hostname, and the instance state.
list_instances	returns either the whole instance locations associated to the user who is running the command or the whole Cloud platform resources.
delete_instance	destroys the created instance.

Table 1: "Cloud CLI" commands.

[7] NMR Resources, https://nmr.le.ac.uk

[8] IGI Portal architecture and interaction with a CA Online, http://agenda.nikhef.nl/getFile.py/access?contribId=18&resId=1&materialId=paper&confId=1522

PoS(ISGC 2013)029

The point 3 will use the changed "Cloud CLI" component to provide their users with the access of cloud provisioning. Figure 4 show how a user can choose an image from a list of images pre-loaded in the MarketPlace service.

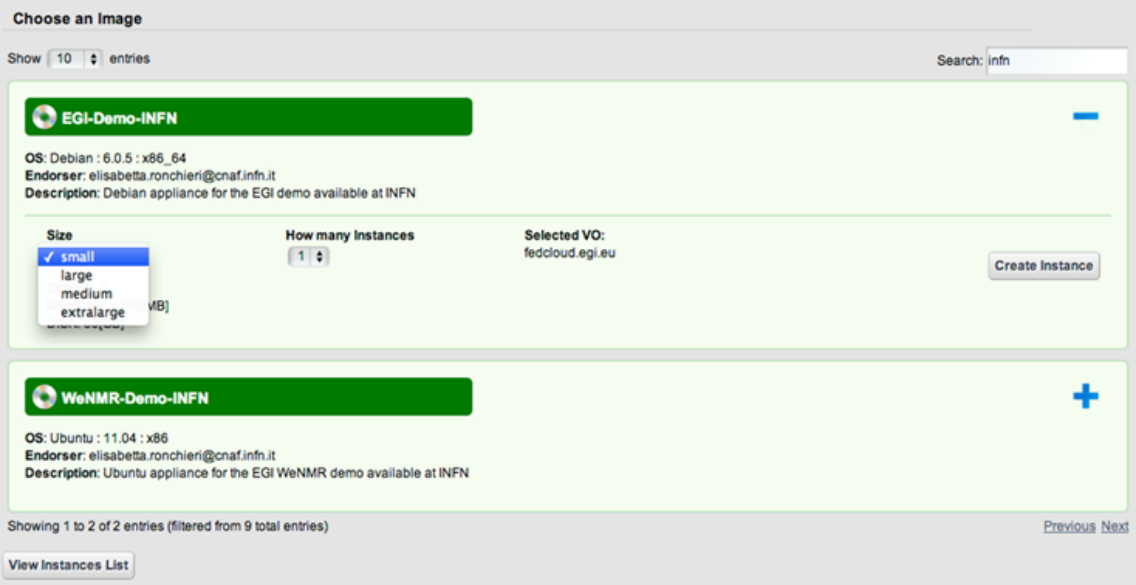

Figure 4: How to choose an image in the IGI portal.

6. Conclusions

During this five years at INFN Tier-1 the WNoDeS framework has proved itself to successfully handle thousands of VMs for the batch and grid requests in its deployment at INFN Tier-1. Recently, the WNoDeS cloud part has been released in the EMI-3 Monte Bianco distribution and deployed on an IGI testbed. This feature is then made available to scientific communities who require it. At different levels of the WNoDeS usage there are collaborations with experiments and working groups such as the Auger experiment of the astro-particle physics community, the WeNMR virtual research community in the structural biology domain, the federated cloud working group of the European Grid Infrastructure (EGI), and the IGI portal. A detailed future plan has been already defined in order to further improve the cloud feature.

PoS(ISGC 2013)029

References

[1] Davide Salomoni, Alessandro Italiano, Elisabetta Ronchieri, "*WNoDeS, a Tool for Integrated Grid and Cloud Access and Computing Farm Virtualization*", 2011 Journal of Physics: Conference Series Volume 331 Part 5: Computing Fabrics and Networking Technologies.

[2] Davide Salomoni, Elisabetta Ronchieri, "*Worker Nodes on Demands Service – Requirements for Virtualized Services*", http://web2.infn.it/wnodes/download/WNoDeS_requirements_v1.0.0.pdf

[3] Elisabetta Ronchieri, Giacinto Donvito, Paolo Veronesi, Davide Salomoni, Alessandro Italiano, Gianni Dalla Torre, Daniele Andreotti, Alessandro Paolini, "*Resource Provisioning through Cloud and Grid Interfaces by means of the Standard CREAM CE and the WNoDeS Cloud Solution*", PoS(EGICF12-EMITC2)124.

[4] Cristina Aiftimiei, Andrea Ceccanti, Danilo Dongiovanni, Alberto Di Meglio, Francesco Giacomini, "*Improving the quality of EMI Releases by leveraging the EMI Testing Infrastructure*", 2012 Journal of Physics: Conference Series Volume 396 Part 5.

[5] *Pierre Auger Observatory* project http://www.auger.org.

[6] Wassenaar *et al.* (2012), *"WeNMR: Structural Biology on the Grid"*, J. Grid. Comp., **10**:743-767

[7] Doreleijers, J. F. et al., *"CING: an integrated residue-based structure validation program suite"*, Journal of biomolecular NMR (2012). doi:10.1007/s10858-012-9669-7.

[8] Doreleijers, J. F. et al., *"NRG-CING: integrated validation reports of remediated experimental biomolecular NMR data and coordinates in wwPDB"*, Nucleic acids research 40, D519–24 (2012).

[9] Nabuurs, S. B. et al., *"DRESS: a database of refined solution NMR structures"*, Proteins 55, 483–6 (2004).

[10] Nederveen, A. J. et al., *"RECOORD: a recalculated coordinate database of 500+ proteins from the PDB using restraints from the BioMagResBank"*, Proteins 59, 662–72 (2005).

[11] Berman, H., Henrick, K., Nakamura, H. & Markley, J. L., *"The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data"*, Nucleic acids research 35, D301–3 (2007).

PoS(ISGC 2013)029