The “Cloud Area Padovana”:

a Distributed IaaS for the INFN User Community

C. Aiftimiei, P. Andreetto\textsuperscript{2}, S. Bertocco\textsuperscript{2}, M. Biasotto\textsuperscript{2}, F. Costa\textsuperscript{2}, A. Crescente\textsuperscript{2}, A. Dorigo\textsuperscript{2}, S. Fantinel\textsuperscript{1}, F. Fanzago\textsuperscript{2}, E. Frizziero\textsuperscript{2}, M. Gulmini\textsuperscript{3}, M. Michelotto\textsuperscript{2}, M. Sgaravatto\textsuperscript{2}, S. Traldi\textsuperscript{2}, M. Venaruzzo\textsuperscript{3}, M. Verlato\textsuperscript{2}, L. Zangrando\textsuperscript{2},

\textsuperscript{1} INFN, CNAF, Bologna, Italy, IFIN-HH, Bucarest, Romania. \textsuperscript{2} INFN, Sezione di Padova, Padova, Italy. \textsuperscript{3} INFN, Laboratori Nazionali di Legnaro, Legnaro (Padova), Italy.

\section*{INTRODUCTION}

The grid paradigm, based on geographically distributed computing, shows its power providing a solution for high CPU power and huge storage space requirements of special research's experiments, first of all the four High Energy Physics (HEP) experiments of LHC at CERN. The grid infrastructure ensures remote resources availability and assures remote data access to authorized users. It assumes to have resources assigned to the experiments (Virtual Organizations or VOs) and resources dedicated for grid services.

On the context of the grid scenario, INFN-Legnaro and INFN-Padova data centres are 10 km far away and are connected with multiple dedicated 10 Gbps optical link, providing a single Tier\textsuperscript{2} for ALICE and CMS experiments of the LHC, sharing infrastructure, hardware and human resources.

The limited flexibility of grid computing environments, the lack of interactivity in resource usage and the authorization based on X509 certificates, are strong barriers for small and medium size experiments.

Small experiments tend to buy independently their own clusters to satisfy their computing needs. This generates a lot of heterogeneous small sized clusters in the same site, each one often underutilized, but hardly sufficient when users are close to deadlines; all this described, produces a low overall efficiency resource exploitation and drives to a high system administrator cost.

\section*{THE POSSIBLE SOLUTION: THE CLOUD MODEL}

The cloud model can provide the \textit{elasticity} required by small experiments at lower costs. The infrastructure is \textquoteleft as a service\textquoteright (IaaS): no dedicated machines are reserved, resources and services are activated on demand by users and released when not utilized.

The configuration of resources as operating system, RAM, number of CPU and storage size is an user choice; federated identity management systems, as the INFN-AAI for example, are enabled instead of X509 personal certificates for user authorization.

The experiment groups buy a quota of this shared computing facility instead of buying their own physical clusters; even if a mechanism to address peak usage to resources temporary unused by other groups is still under development, a more overall efficient usage of resources is already obtainable.

\section*{THE CLOUD AREA PAODOVANA PROJECT}

At the end of 2013 INFN-Padova and INFN-Legnaro started a cloud-based project providing computing and storage resources with the aim to address the needs of the numerous smaller sized physics experiments carried out by the local teams that, in many cases, are spread on both sides; this new, joint project was started thanks to the long-term and well established collaboration experience, gained with the Tier\textsuperscript{2}; the new experience could improve, in a long term view, the future of the Tier\textsuperscript{2} resources in terms of management and integration with the technology evolution.

The first step of the collaboration was the selection of the primary component of the infrastructure: the Cloud Management Framework. One of the most popular open source solutions, widely adopted in the scientific domain of INFN and also at CERN, is OpenStack\textsuperscript{2} and this was chosen as the base of the Cloud Area Padovana infrastructure. The preliminary study on the implementation highlighted the need to concentrate on the deployment of high available services as to grant the business continuity to the users/experiments as required by a project of this level.

Database and OpenStack services are deployed now in High Availability (HA) in Padova, and configured in Active/Active mode: services run at the same time on all nodes of the redundant cluster, allowing for easier maintenance without user service interruption.

To reduce the manual effort and possible other problems, a combination of Foreman\textsuperscript{3} and Puppet\textsuperscript{4} is exploited; in-house customized Puppet modules are provided for the OpenStack Compute Node profile.

On the monitoring side, well known components as Nagios\textsuperscript{3} and Ganglia\textsuperscript{6} were adopted. Nagios is instrumented also with customized modules (plug-ins), to check OpenStack services and all the real hosts composing the infrastructure: it fires e-mail alarms when the programmed conditions will happen. With the graphs of Ganglia, it is possible to check the history of all the resources behavior, both physical and virtual, and drill down to inspect possible anomalies and/or real problems around certain time intervals.

Another important in-house development, is the enhancement of the authentication and authorization (AA)
module of OpenStack that provides basically only the classic username/password mechanism. The collaboration integrated the ability to manage the users AA through an external Identity Provider (IdP) service as, for example, the INFN-AAI; in addition, it was developed a new extra component to manage the registration workflow; for each project (tenant) a leader person is entitled to approve or reject a new user request for that project: the complex resulting workflow is represented in Figure 1.

Fig. 1. User registration workflow developed for the Cloud Area Padovana.

The computing and storage resources, with some details on the related technology and protocols, split by site, exploitable by the users, are summarized in Table 1; as mentioned earlier in the document, the resources are placed in two different rooms 10 km far away. This computing rooms are shared with the LNL-PD Tier2, but are linked with a different 10 Gbps network link.

<table>
<thead>
<tr>
<th>Computing Resources</th>
<th>Storage Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PD</strong></td>
<td><strong>PD</strong></td>
</tr>
<tr>
<td>8 servers; 320 cores (Intel E5-2670v2 / E5-2650v3), 768 GiB RAM</td>
<td>10 TiB via Cinder OpenStack service (based on iSCSI technology)</td>
</tr>
<tr>
<td><strong>LNL</strong></td>
<td><strong>LNL</strong></td>
</tr>
<tr>
<td>6 servers; 192 cores (Intel E5-2650v2), 576 GiB RAM</td>
<td>96 TiB via GlusterFS/NFS/Samba (based on Fiber Channel technology)</td>
</tr>
</tbody>
</table>

**USE CASES ON CLOUD AREA PADOVANA**

The experiments that evaluate in this initial phase the cloud infrastructure have different use-cases; from the most important:
- ALICE: needs an elastic cluster of VMs for interactive analysis (Virtual Analysis Facility); tested sending simple pilot jobs and accessing an external storage from VMs via xrootd protocol;
- CMS: needs a virtual facility to run interactive data processing, accessing from VMs the Tier-2 storage in Legnaro and an end-user storage (Lustre) in Padova; the users home are imported from Lustre;
- CMT: the Cosmic Muon Tomography, a not invasive imaging system to scan hidden high density materials, needs a big computing power to run the recursive algorithms for input data analysis and find the optimal density of target material;
- JUNO: a neutrino experiment under construction in China needs the cloud computing power for events simulation and analysis;
- SPES: beam dynamics errors study for the SPES project.

**FUTURE SERVICE: THE FAIR SHARE ALGORITHM**

Currently OpenStack allows only the static partitioning model of the available resources. One project cannot exceed its own quota even if there are unused resources allocated to other projects, so new requests are rejected. A better approach is needed to enable a more effective and flexible resource allocation and utilization in OpenStack, based on an advanced scheduling algorithm. The collaboration started to address the problem by developing a new OpenStack service named *FairShareScheduler*, which allows for a dynamic resource allocation; in particular it provides new scheduling capabilities as:
1. queuing mechanism for handling the user requests;
2. fair-share algorithm based on the Priority MultiFactor strategy of the SLURM7 Workload Manager.

**USE CASES ON CLOUD AREA PADOVANA**

The experiments that evaluate in this initial phase the cloud infrastructure have different use-cases; from the most important:
- ALICE: needs an elastic cluster of VMs for interactive analysis (Virtual Analysis Facility); tested sending simple pilot jobs and accessing an external storage from VMs via xrootd protocol;
- CMS: needs a virtual facility to run interactive data processing, accessing from VMs the Tier-2 storage in Legnaro and an end-user storage (Lustre) in Padova; the users home are imported from Lustre;
- CMT: the Cosmic Muon Tomography, a not invasive imaging system to scan hidden high density materials, needs a big computing power to run the recursive algorithms for input data analysis and find the optimal density of target material;
- JUNO: a neutrino experiment under construction in China needs the cloud computing power for events simulation and analysis;
- SPES: beam dynamics errors study for the SPES project.

**FUTURE SERVICE: THE FAIR SHARE ALGORITHM**

Currently OpenStack allows only the static partitioning model of the available resources. One project cannot exceed its own quota even if there are unused resources allocated to other projects, so new requests are rejected. A better approach is needed to enable a more effective and flexible resource allocation and utilization in OpenStack, based on an advanced scheduling algorithm. The collaboration started to address the problem by developing a new OpenStack service named *FairShareScheduler*, which allows for a dynamic resource allocation; in particular it provides new scheduling capabilities as:
1. queuing mechanism for handling the user requests;
2. fair-share algorithm based on the Priority MultiFactor strategy of the SLURM7 Workload Manager.