Implementation and use of a highly available and innovative IaaS solution: the Cloud Area Padovana

P.Andreetto, S.Bertocco, F.Costa, A.Crescente, A.Dorigo, F.Fanzago, E.Frizziero, M.Michelotto, M.Sgaravatto, S.Traldi, M.Verlato, L.Zangrando (INFN-Padova)

M.Biasotto, M.Gulmini, S.Fantinel, M.Venaruzzo (INFN-Legnaro)

S.Dal Pra (INFN-CNAF)

C.Aiftimiei (INFN-CNAF, IFIN-HH)

Why the Cloud?

The problem

The limited flexibility of grid computing environments, the no interactivity resource usage and the authorization based on X509 certificates are usually strong barriers for small experiments

Small experiments tend to buy independently their own clusters to satisfy their computing needs

- a lot of heterogeneous small sized clusters in the same site
- no constant usage: often underutilized while insufficient close to deadlines

↓

Low overall efficiency and high system administration cost

The possible solution: the Cloud model

The Cloud model can provide the elasticity required by small experiments at lower costs. The infrastructure is "as a service"

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 are enabled instead of X509 for user authorization.

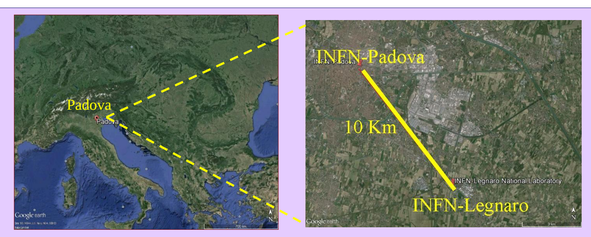
The experiment groups buy a quota of this shared computing facility instead of buying their own physical clusters

Even if a mechanism to address peak usage to resources temporary unused by other groups is still under development, a more efficient usage of resources is already obtained

The Cloud Area Padovana project

At the end of 2013 INFN-Padova and INFN-Legnaro started a project for a Cloud infrastructure spread between the two sites, providing computing and storage resources with the aim to address not only the need of already supported LHC experiments but also the needs of the numerous smaller sized physics experiments carried out by the local teams.

This was done also considering the experience with the distributed Padova-Legnaro Tier-2, which allowed the sharing of infrastructures, hardware and human resources between the two sites.



Cloud Area Padovana implementation

The network layout

An extended LAN, implemented via VLANs, was deployed between Padova and Legnaro.
VMs of the Cloud can be accessed from both Padova and Legnaro networks, even if they don't have public floating IPs

The framework

As **Cloud Management framework** one of the most popular open source solutions widely adopted in the scientific domain of INFN was chosen: **OpenStack**

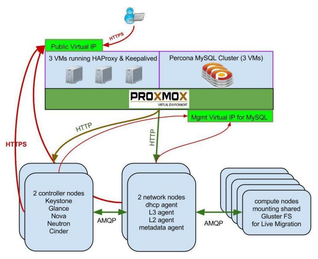
Infrastructure management

To reduce the manual effort and to avoid configuration problems, a combination of Foreman and Puppet is used to deploy the Cloud.
Some **customized Puppet modules** were implemented.

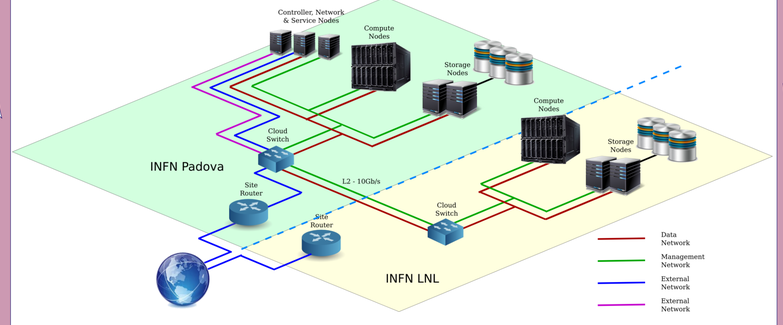
High Availability service

Database and OpenStack services are deployed in HA in Padova. HA is configured in Active/Active mode, stateless services run at the same time on all nodes of the redundant cluster, allowing easier maintenance without user service interruption.

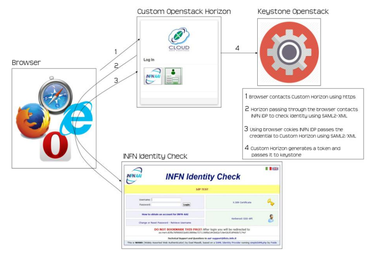
Thanks to the HAProxy's SSL termination capability, all the public services expose a **SSL-protected interface**, even if the services themselves are not configured to listen on SSL



Infrastructure layout



User authentication and authorization

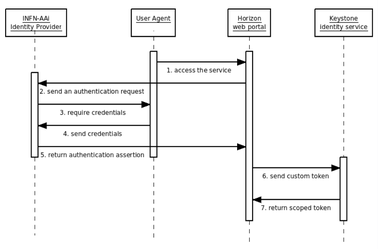


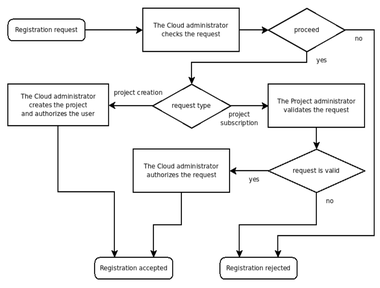
Authentication can be provided through user/password or through an external Identity Provider like **INFN-AAI**

The Single Sign On is based on **SAML v2**

The authorization phase relies on **Keystone protocol**

Portal service (Horizon) communicates with Keystone via **custom tokens**





A module to manage **registrations** of users and projects was implemented and integrated in Horizon.

Users must be registered to the portal service (Horizon).

Requests for registration must pass through the depicted **registration workflow**

Available resources

Computing system

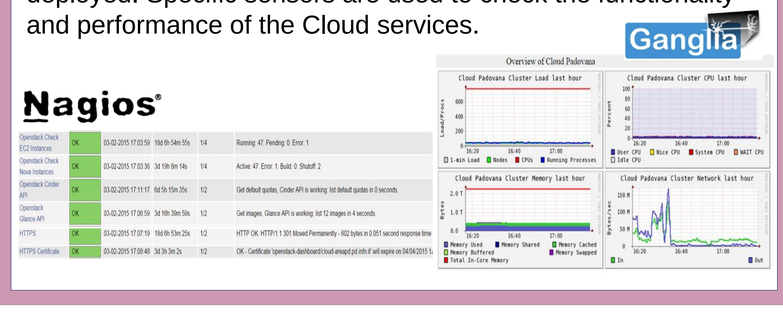
Location	# of servers	CPU Model	# of CPU	RAM (GB)	# of CPU cores	OpenStack role
PD	4	E5-2609	2	32	8	Controller, Network nodes
PD	5	E5-2650v2	2	96	40	COmpute nodes
PD	3	E5-2650v3	2	96	40	Compute nodes
LNL	6	E5-2650v2	2	96	32	Compute nodes
Total	18		32	1472	544	

Storage

Location	Storage system	# of disks	Disk size (GB)	Usage
PD	iSCSI DELL	23	900	for the cloud images (Glance service) ephemeral storage for VM (Nova service) persistent storage for VM (Cinder service)
LNL	Fiber channel	12	4000	general purpose user storage

Monitoring

A monitoring infrastructure based on **Ganglia** and **Nagios** were deployed. Specific sensors are used to check the functionality and performance of the Cloud services.



Infrastructure management

To reduce the manual effort and to avoid configuration problems, a combination of Foreman and Puppet is used to deploy the Cloud.
Some **customized Puppet modules** were implemented.

Current usage and work in progress

First experiments on Cloud Area Padovana

First experiments using this Cloud infrastructure were the following:

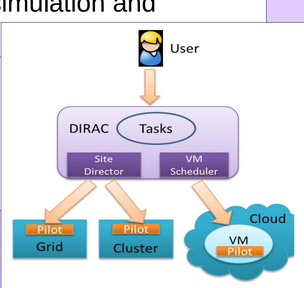
ALICE: needs an elastic cluster of VMs for interactive analysis (Virtual Analysis Facility) where worker nodes are instantiated on demand

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.

CMT: the Cosmic Muon Tomography, a not invasive imaging system to scan hidden high density materials, needs 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.

The project managers have to provide images in qcow format allowing users to instantiate the VMs ready to run required processes.

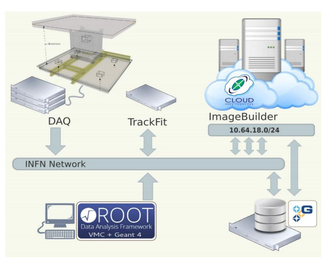


- The simulation of a calibration event needs more than 700 min
- JUNO Images include SL6 with CVMFS client installed.
- Jobs are submitted from the access point in Beijing
- Jobs reach the clouds through the DIRAC framework

More than 700 jobs executed on INFN Cloud Area Padovana with 100% success rate

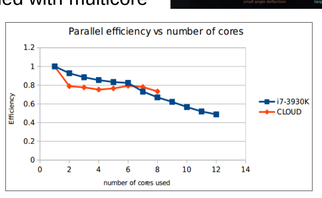
Some numbers …..

Currently 13 registered projects: ALICE, Belle II, CALET, CMS, CMT, CUORE, DIV.RICERCA-LNL, GAMMA, JUNO, LHCb, OCP, SPES, T2K.
About 60 users registered.



- Detector at INFN-LNL
- Input data at INFN-Padova storage server

- CMT image included reconstruction software and configured to mount the storage via NFS
- Good performance obtained with multicore VMs



Future service: the fair-share scheduler

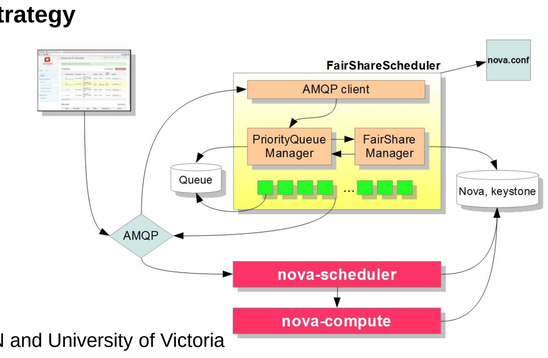
Currently OpenStack allows just the **static partitioning model**

- **low global efficiency and increased cost in terms of resource usage**
- **it doesn't allow continuous full utilization of all available resources**

In a scenario of full resource usage for a specific project, new requests are simply rejected. We started to address the problem by developing a new OpenStack service, the **FairShare Scheduler**, which allows a **dynamic resource allocation model.**

In particular it interacts with nova-scheduler and provides new scheduling capabilities as:

- **queuing mechanism for handling the user requests**
- **fair-share algorithm based on the SLURM Priority MultiFactor strategy**



Under testing at INFN and University of Victoria