Harvesting dispersed computational resources with Openstack

a Cloud infrastructure for the
Computational Science community

Mirko Mariotti
mirko.mariotti@unipg.it
ISGC 2018 - Academia Sinica - Taipei

Agenda

- General overview of our problem.
- Some words on our OpenStack Installation.
- How we extend our system to remote resources.
- Use cases.

General overview

Harvesting dispersed computational resources is an important topic for nowaday, in particular for a small center.

The main goal of the present work is to illustrate a real example on how to build a geographically distributed cloud to share and manage computing resources, owned by heterogeneous cooperating entities.

Openstack @ Perugia (Italy)

- Small OpenStack installation (~600 cores)
- Computational resources for local researcher, students, labs, events.
- Not only services, base for our R&D on cloud technologies

INDIGO - DataCloud

EOSC-hub

eXtreme DataCloud

Openstack @ Perugia (Italy)

- AA federated with INFN-AAI and Unipg IDM
- Network virtualization via neutron and VLAN backend
- Storage: cinder, ceph
- Two installation, one production (Mitaka), one development (latest available)
- OpenStack core machine also virtualized (outside OS)

Dispersed resources

Some of our researchers have access to other computational centers geographically distributed.

The centers are not cloud-based:

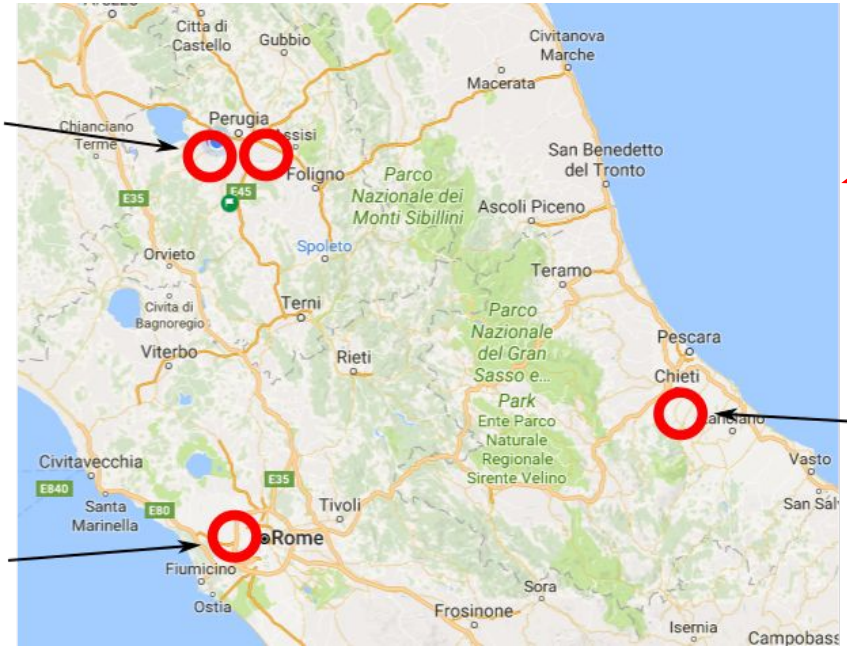
- Lack of local manpower.
- Not big enough to install a complete cloud system.

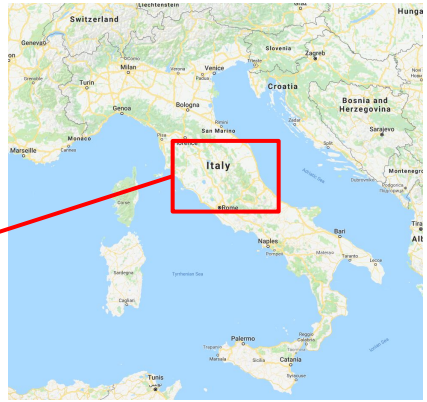
Dispersed resources

Locations

Dept of Physics and Geology/ INFN Perugia

Dept of Chemistry





Dept of Pharmacy

ASI-SSDC Space Science data center at the Italian Space Agency

The technical objectives

- To include remote resources into our local OpenStack installation.

- To make sure that the included satellite resources are used efficiently by the cloud framework.

- To give back to the owning research group in the form of cloud resources (instances, storage, and recipes)

The Pillars

- A single OpenStack installation.
- Resource organized in different zones logically correspondent to different geographical locations.
- SDN (software-defined networking) solution to connect the different zones.
- All build with standard servers and Linux systems.

A single OpenStack Installation

A central OS installation control all the sites but …

the sites have to be as much autonomous as possible especially regarding:

- Storage
- Outbound connectivity

Cross-site operations have to be possible (knowing the risks).

Ideally the traffic among sites would be only the OpenStack management one.

SDN (Software-Defined Networking)

Software-Defined Networking is a way to overlay multiple networks to a single physical fabric and to control them via software.

Openvswitch is an open source project for SDN

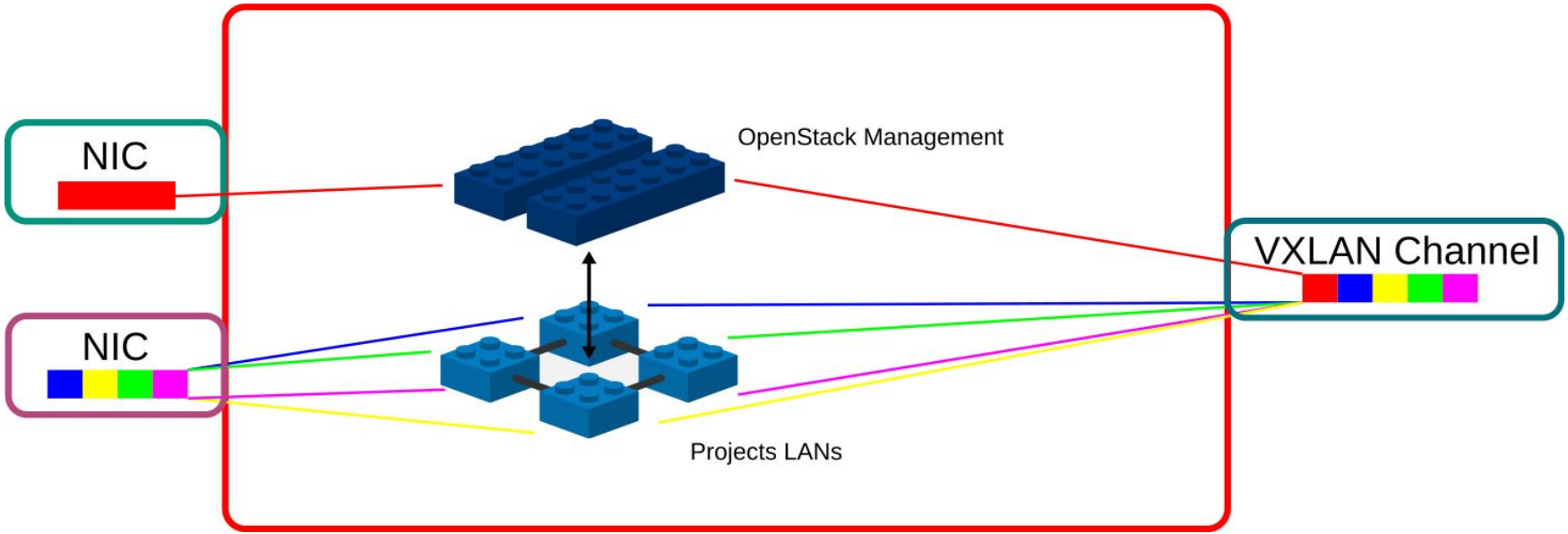
Used for network virtualization in many cloud framework

We are using this approach and Openvswitch also to the physical infrastructure.

SDN

How we use it

We use a Linux box for each site to “virtualize” the openstack LAN (Both management and projects) and transport it to other sites.



Harvesting the resources

A single node

- Ubuntu 16.04 LTS server (with a 4.8 kernel)
- Openvswitch 2.5.2

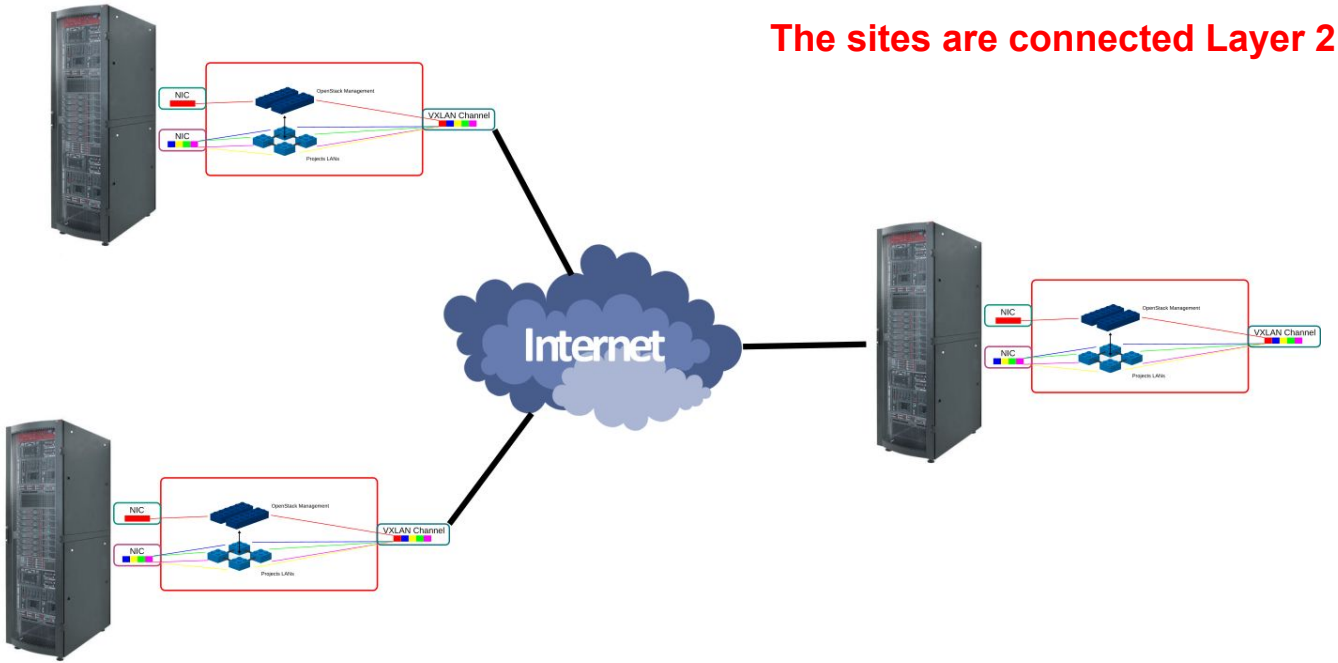
```
[asigw1] /root #  ovs-vsctl show
14f2df09-7723-41df-beb4-67d1ad48d064
    Bridge asitunnel
        Port asitunnel
            Interface asitunnel
                type: internal
        Port "enp4s3"
            trunks: [402, 1016, 1065]
            Interface "enp4s3"
        Port "vxlan0"
            tag: 0
            trunks: [402, 1016, 1065]
            Interface "vxlan0"
                type: vxlan
                options: {df_default="false", key=flow, remote_ip="10.199.190.4"}
        Port ams
            tag: 0
            Interface ams
                type: internal
    ovs_version: "2.5.2"
[asigw1] /root #
```

```
auto asitunnel
allow-ovs asitunnel
iface asitunnel inet manual
        ovs_type OVSBridge
        ovs_ports ams enp4s3 vxlan0

allow-asitunnel enp4s3
iface enp4s3 inet manual
        ovs_bridge asitunnel
        ovs_type OVSPort
        ovs_options vlan_mode=trunk trunks=402,1016,1065

allow-asitunnel vxlan0
iface vxlan0 inet manual
        ovs_bridge asitunnel
        ovs_type OVSTunnel
        ovs_tunnel_type vxlan
        ovs_options tag=0 vlan_mode=native-untagged
trunks=402,1016,1065
        ovs_tunnel_options options:remote_ip=10.199.190.4
options:key=flow options:df_default=false
```

Connected nodes



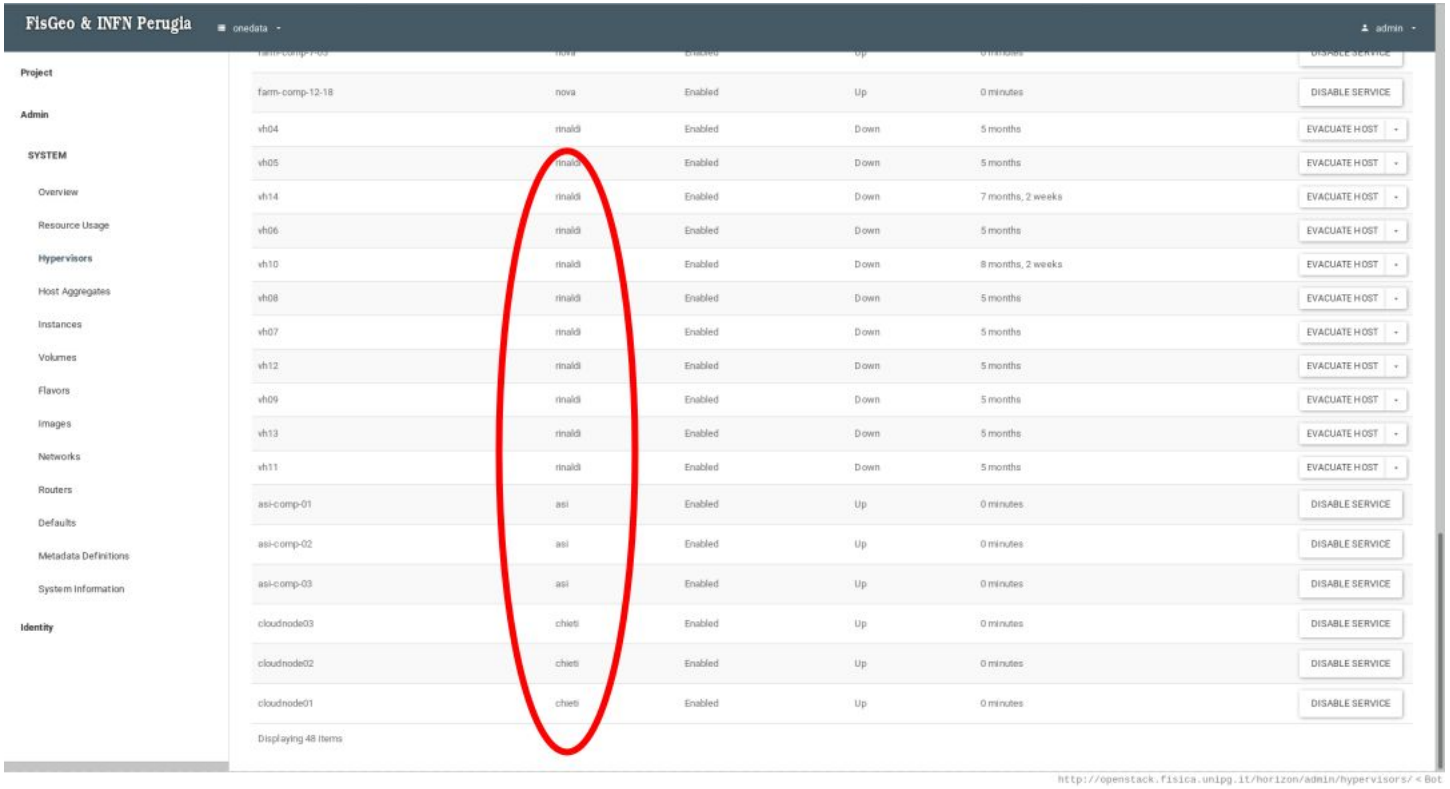
The sites are connected Layer 2

Network Security

The VXLAN tunnel has to be encrypted, we tried two solution:

- OpenVPN point to point
  - Routers friendly (standard UDP/TCP traffic)
  - Less performant
- IPSEC
  - Routers unfriendly
  - More fragmented traffic
  - More performant

Zones


FisGeo & INFN Perugia
onedata
admin
Project
Admin
SYSTEM
Overview
Resource Usage
Hypervisors
Host Aggregates
Instances
Volumes
Flavors
Images
Networks
Routers
Defaults
Metadata Definitions
System Information
Identity
farm-comp-12-18 nova Enabled Up 0 minutes DISABLE SERVICE
vh04 rinaldi Enabled Down 5 months EVACUATE HOST
vh05 rinaldi Enabled Down 5 months EVACUATE HOST
vh14 rinaldi Enabled Down 7 months, 2 weeks EVACUATE HOST
vh06 rinaldi Enabled Down 5 months EVACUATE HOST
vh10 rinaldi Enabled Down 8 months, 2 weeks EVACUATE HOST
vh08 rinaldi Enabled Down 5 months EVACUATE HOST
vh07 rinaldi Enabled Down 5 months EVACUATE HOST
vh12 rinaldi Enabled Down 5 months EVACUATE HOST
vh09 rinaldi Enabled Down 5 months EVACUATE HOST
vh13 rinaldi Enabled Down 5 months EVACUATE HOST
vh11 rinaldi Enabled Down 5 months EVACUATE HOST
asi-comp-01 asi Enabled Up 0 minutes DISABLE SERVICE
asi-comp-02 asi Enabled Up 0 minutes DISABLE SERVICE
asi-comp-03 asi Enabled Up 0 minutes DISABLE SERVICE
cloudnode03 chieti Enabled Up 0 minutes DISABLE SERVICE
cloudnode02 chieti Enabled Up 0 minutes DISABLE SERVICE
cloudnode01 chieti Enabled Up 0 minutes DISABLE SERVICE
Displaying 48 items
http://openstack.fisica.unipg.it/horizon/admin/hypervisors/

Per-zone customizations

In order to avoid cross-site interactions some consideration has to be taken into account:

- Storage:
  - VMs in each zone has to use storage backend from the same zone.
- Network:
  - It is a nonsense to allow outbound traffic from satellite sites to go back to the main site.
  - Custom gateways for projects network on those zones. (more on the use cases)

Possible Issues

Site security

All the sites are on the same Layer 2.

Errors, misconfigurations, problems can potentially impact on the whole system.

Sites has to be trusted

Possible Issues (cont.)

Poor performance

Encryption (OpenVPN/IPSEC) and encapsulation (VXLAN) are bandwidth consuming (especially on commodity hardware).

This could be a problem for cross-site operation, but not a real problem for OpenStack control traffic (~50 kBit/s each hardware node)

Some measures

```
Device eth0 [10.13.100.1]
==========================================
Incoming:                 Outgoing:
Curr: 1.40 MBit/s         Curr: 1.19 MBit/s
Avg: 2.28 MBit/s          Avg: 1.34 MBit/s
Min: 132.29 kBit/s        Min: 72.03 kBit/s
Max: 4.91 MBit/s          Max: 2.82 MBit/s
Ttl: 7746.23 GByte        Ttl: 4209.48 GByte
```

Traffic on the OS controller node

AVG in: ~ 50 kBit/s/hwnode
AVG out: ~ 25 kBit/s/hwnode

```
Device eth0 [10.13.0.4]
==========================================
Incoming:                 Outgoing:
Curr: 1.13 MBit/s         Curr: 2.41 MBit/s
Avg: 868.66 kBit/s        Avg: 1.84 MBit/s
Min: 172.24 kBit/s        Min: 355.59 kBit/s
Max: 1.81 MBit/s          Max: 3.88 MBit/s
Ttl: 2594.84 GByte        Ttl: 6880.46 GByte
```

Traffic on a DB/rabbitmq node

AVG in: ~ 17 kBit/s/hwnode
AVG out: ~35 kbit/s/hwnode

Still possible issues

Network problems

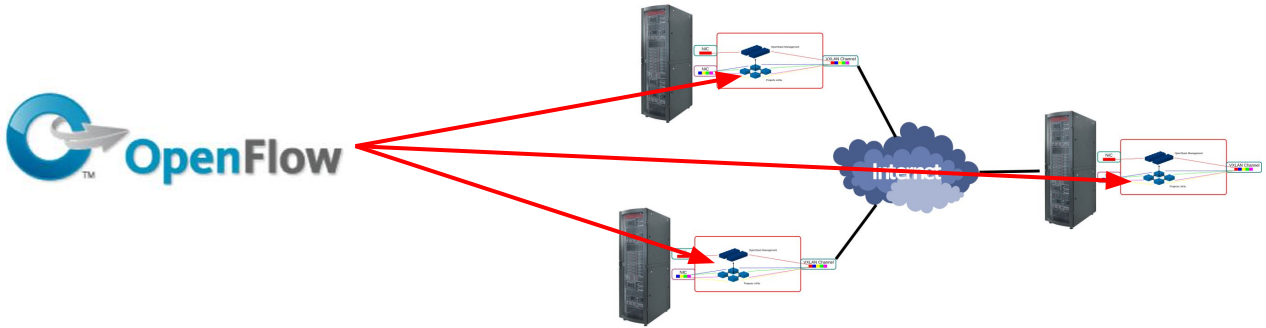
For any reason the network connection to a site is severed what happen to VMs ?

OpenStack is resilient to this situation, it cannot contact anymore the resources but VMs continue to work correctly (provided their storage is not cross-site).

Other sites are not affected.

Automation

The sites are L2 connected, every automatic installation/configuration mechanism available on the master site work out of the box on the remote sites (preseed, puppet etc).



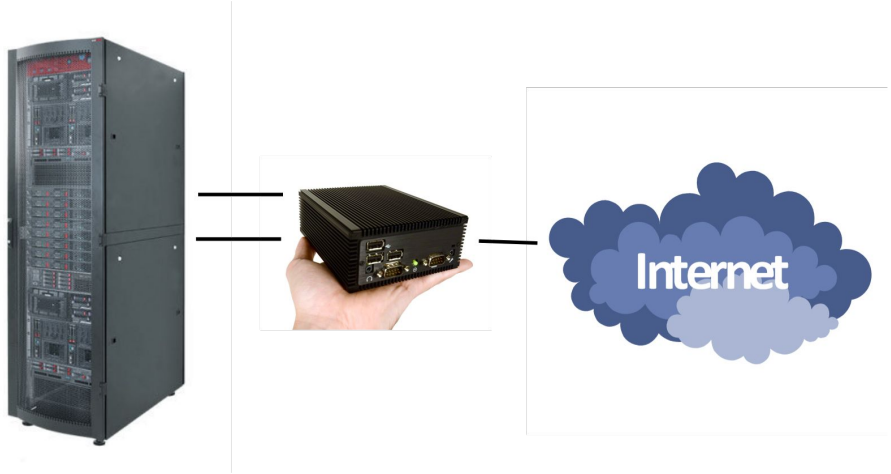
No problem for Openvswitch but constraints on the switching hardware

Next step

A pre-configured system that with the prerequisites of:

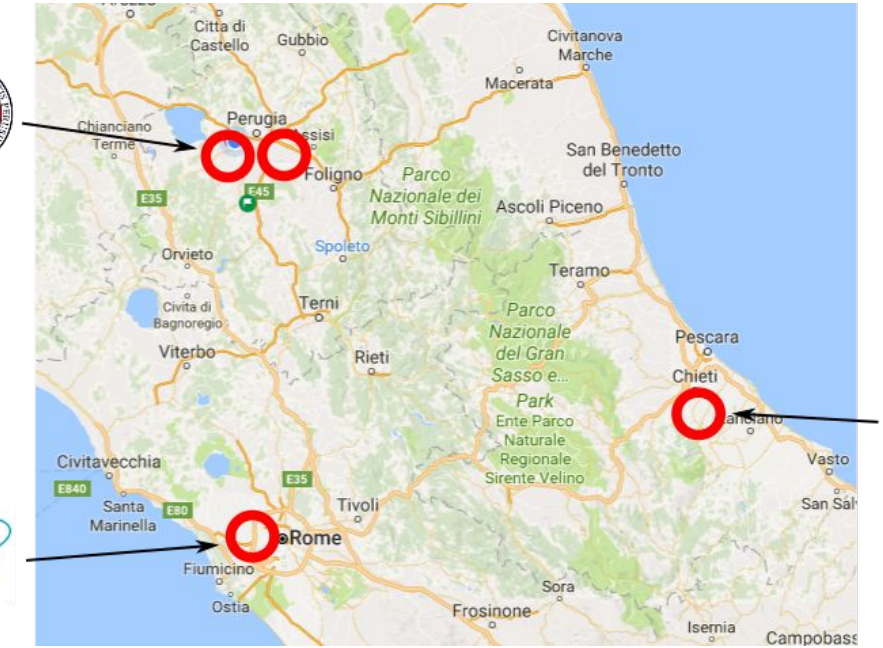
- Openflow compliant switches.
- A standard way of cabling a rack.
- A public IP.

Deploy and configure that rack as an extension to our OS installation.



Use cases

Use case 1:
AMS analysis with DODAS



Use case 2:
Computational Chemistry

Use case 1: DODAS

Slide from the talk of D.Spiga (Thursday 20/3/2018)

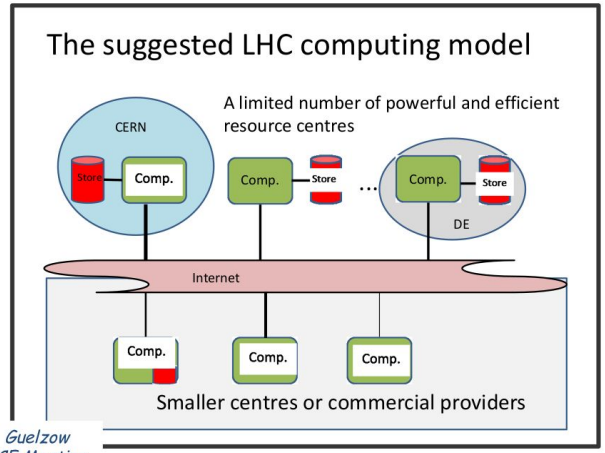
EOSC-hub A Service for HEP data processing

In this context, DODAS has been developed to allow

- ***Opportunistic resources integration***
  - Targeting both private and pubblic cloud providers
- ***Extension of already existing facility***
  - Both to absorb peaks of usage and to generate mission specific facilities
- ***User-friendly computing infrastructure generation and management***
  - of computing cluster to analyse research data

Look to the future:

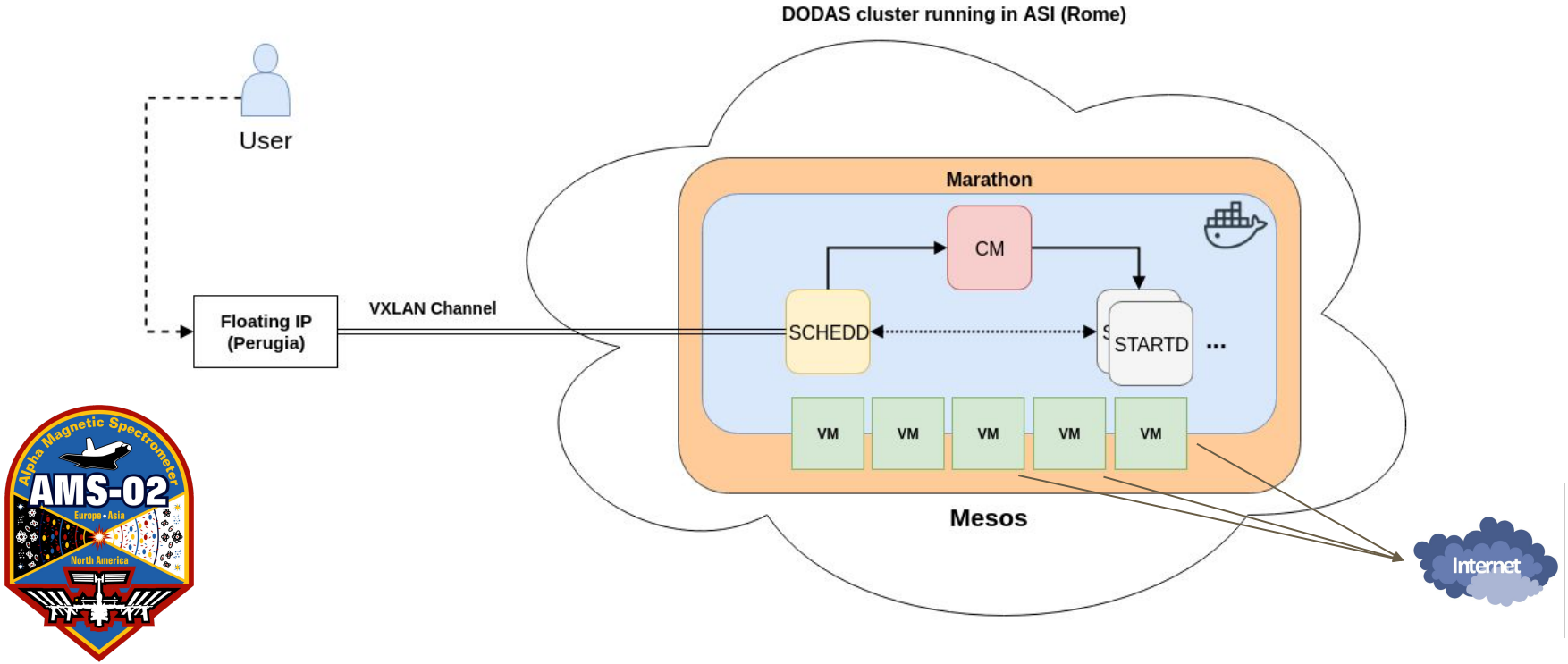
- ***DODAS design compliant with future paradigms***



V. Guelzow SCF Meeting 12/5/2017

Daniele Spiga ISGC 2018, 16-23 March Academia Sinica, Taipei 5

Use case 1: DODAS



Use case 2: Computational Chem

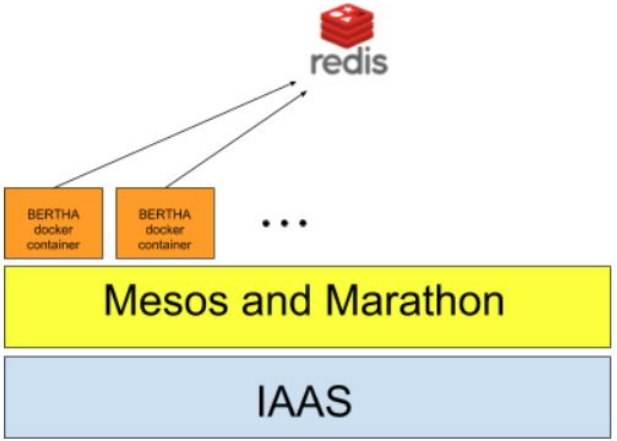
We may name several scenarios that can be easily adapted to a Cloud architecture as the one deployed:

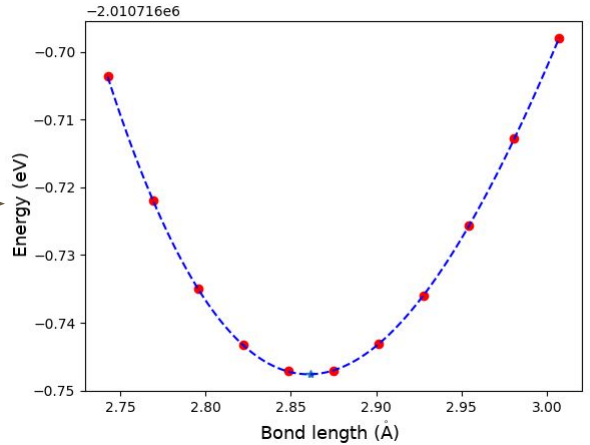
- **Complex workflows :** e.g. calculation of the ab-initio values of the potential energy surface (PES), fitting of the points, integration of the nuclei dynamics equations and the final statistical analysis and visualization of the results
- **Drug Design**: need to build computational protocols made of many different steps, e.g. Virtual Screening run an entire sequence of jobs to screen a large collection of ligands against one or multiple targets.

L. Storchi. F. Tarantelli, A. Laganà,. "Computing molecular energy surfaces on a Grid." LNCS 3980, 675 (2006)
F. Milletti, L. Storchi, G. Sforna, S. Cross, G, Cruciani, "Tautomer Enumeration and Stability Prediction for Virtual Screening on Large Chemical Databases" , Journal of Chemical Information and Modeling, 49 (1), 68 (2009).

Use case 2: Computational Chem

Quantum Chemistry : e.g. we deployed an approach to perform a geometry optimization using the Dirac-Kohn-Sham module of BERTHA, a full 4-component DKS calculation (bond length of the $AuOg^+$ molecular system).





L. Storchi , S. Rampino , L. Belpassi , F. Tarantelli , H. M. Quiney,"Efficient parallel all-electron four-component Dirac-Kohn-Sham program using a distributed matrix approach.II" JCTC, **2013**, *9* (12), pp 5356–5364

Conclusions

We successfully built a cross-site OpenStack centrally managed, including resources otherwise poorly used

The whole system is based on open source standard solutions and commodity hardware.

Resource are isolated in zones for efficiency and to avoid cross-site interactions.

We used successfully this infrastructure in real cases.

Thanks

People that contributed to this project:

VITILLARO, Giuseppe (Istituto di Scienze e Tecnologie Molecolari, Consiglio Nazionale delle Ricerche)
FORMATO, Valerio (INFN sezione di Perugia)
DURANTI, Matteo (INFN sezione di Perugia)
SPIGA, Daniele (INFN sezione di Perugia)
CIANGOTTINI, Manuel (INFN sezione di Perugia)
STORCHI, Loriano (Dipartimento di Farmacia, Università degli Studi "G. D'Annunzio", Chieti.)
MERGÉ, Matteo (INFN sezione Roma 2);
D'ANGELI, Paolo (Space Science Data Center at the Italian Space Agency.)
GUERRA, Antonio (Space Science Data Center at the Italian Space Agency.)
PRIMAVERA, Roberto (Space Science Data Center at the Italian Space Agency.)