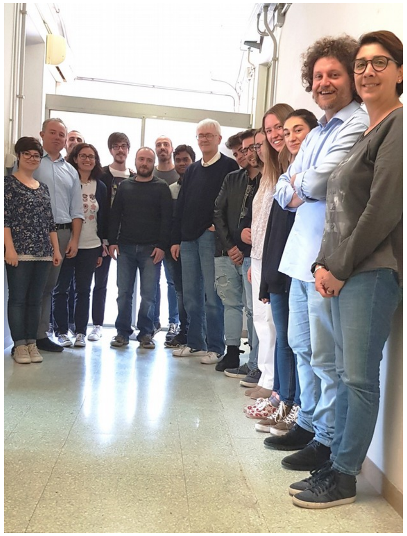
CLOUD Platform
A Virtual Cluster

CNR – SCITEC

Giuseppe Vitillaro Tecnologo

sysadm ***thch.unipg.it***

Consiglio Nazionale delle Ricerche
SCITEC
Istituto Scienze e Tecnologie Chimiche
UOS Perugia



SCITEC - Production Linux Clusters

2010

2013

2019

SC SimpleCore: 12 nodes - 84 cores - 456Gb Memory - 7Tb Storage - Gbit

MC MultiCore: 36 nodes - 528 cores - 3664Gb Memory - 60Tb Storage - IB

TC ThchCore : 24 nodes - 768 cores - 6144Gb Memory - 92Tb Storage - IB

SC

MC

TC

http://www.clhyo.org/about-us/resources.html

Clusters – DCBB/CNR History

1999 **BW** 8 nodes Intel Pentium Beowulf - 100Mbits cluster - RedHat (32bit)

2003 **RX** 6 nodes Intel Itanium Beowulf - Gbits cluster - RedHat 3 (64bit)

2005 **DC** 26 nodes Intel Pentium Cluster - Gbits cluster - RedHat 4 (64bit)

2010 **SC** 12 nodes Intel Xeon Cluster - Gbits cluster - RedHat 5 (64bit)

2013 **MC** 36 nodes Intel Xeon Cluster - 40Gbits (IB) cluster - CentOS 5 (64bit)

2018 **HS** 7 nodes Intel Xeon Clutster - Gbits Virtual cluster - SL 6 (64bit)

2019 **TC** 24 nodes Intel Xeon Clutster - 40Gbits (IB) cluster - SL 6 (64bit)

Clusters built for Molecular Sciences HPC

- From the very beginning, after IBM RISC/6000 and IBM SP (XX century technologies), our beowulfs and clusters were built to run for days, weeks, months, intensive, parallel, Molecular Sciences HPC (High Performance Computing) applications
- The main goal was always that of building a reliable cluster of nodes which users may see as a single computing system, where jobs may be submitted, monitored and managed from a single access node
- Keeping in mind the complexity of computing systems should hide from the user eyes, usually a scientist which is mainly interested in running applications

Clusters – Single System Image

- Single access node, control workstation (CW)
  - hostname after the cluster id name
  - **SC**cw, **MC**cw, **TC**cw

- Single System image
  - *user view* achieved using “old technologies”, simple, but still effective
  - Yellow Pages, aka NIS or NIS+ and NFS
  - It may easily scale up to hundred of nodes clusters
  - we are using this model, from the ninety, to build our clusters

- Node images built from the access node image
  - easy to build, to test and verify, after booting CW and a couple of nodes
  - generally small, ~2Gbytes -- node provisioning? a matter of minutes, at gigabit speed
  - images easy to be modified and replicated
  - as simple as doing a tar image
  - a process driven from simple shell scripts, easy to manage
  - nodes can be installed in parallel - an hundred nodes cluster in hours
  - flexible enough to allow migration to new distros or distro versions
  - done, up to now, manually, with redhat-like distros
  - Ubuntu is a possible option, MAAS/juju ***may*** simplify provisioning

Clusters – System Image HPC toolbox

- **TORQUE**: Terascale Open-source Resource and *QUEue Manager*
  - extended from the original PBS (used in old beowulfs)
  - handle user job submission and resource allocation
  - easy to use, after some training, and well documented online
- **MAUI** Cluster Scheduler
  - effective *Job Scheduler* for High Performance clusters and supercomputing
  - capable of supporting multiple scheduling policies, dynamic priorities, reservations, and fairshare capabilitie

Clusters: User View

TORQUE: qstat

```
Job id                    Name             User            Time Use S Queue
------------------------- ---------------- --------------- -------- - -----
57228.sccw                 sadlosshc.g09    marzio                0 Q q1
57229.sccw                 sadlosscn.g09    marzio                0 Q q1
57232.sccw                 symmblyp         gluca                 0 Q q1
```

lmod modules

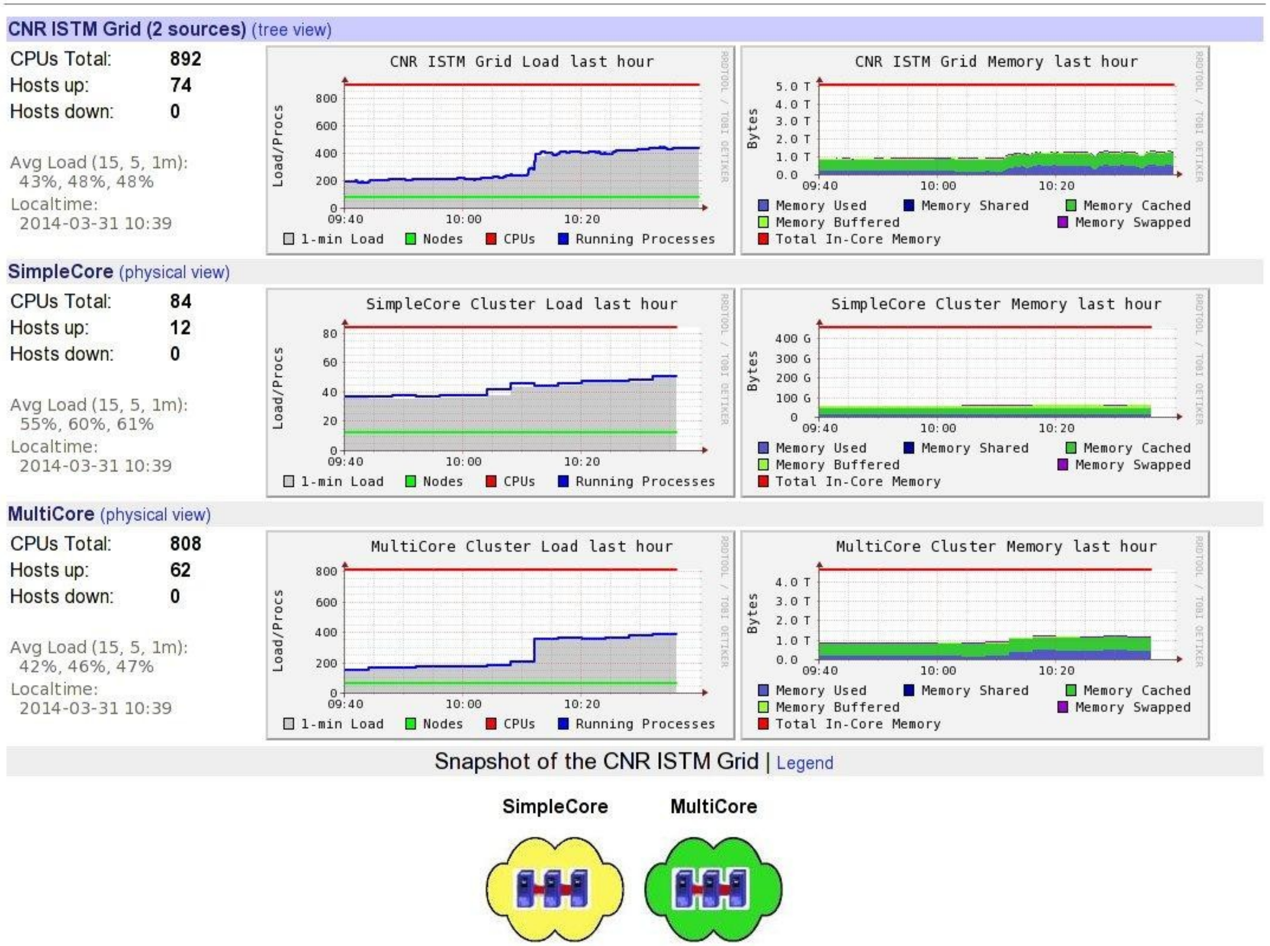
```
[peppe@hscw ~]$ module list

Currently Loaded Modules:
  1) intel/14.0.2   2) mvapich2/1.9   3) StdEnv


[peppe@hscw ~]$ module avail

----------------- /usr/local/modulefiles/MPI/intel/14.0.2/mvapich2/1.9 -----------------
   espresso/39.62     gaussian/09-c01-omp    nwchem/6.5.26243
   gamess/050113R1    molpro/2010p-omp       siesta/3.2-pl-4

-------------------- /usr/local/modulefiles/Compiler/intel/14.0.2 --------------------
   gotoblas2/1.13-omp        mkl/11.1.2-omp (D)    mkl8/11.1.2-omp    mvapich2/1.9
   gotoblas2/1.13      (D)   mkl/11.1.2            mpich/3.1          scalapack/2.0.2

---------------------------- /usr/local/modulefiles/Core -----------------------------
   StdEnv    gcc/sys    intel/14.0.2    open64/5.0.0

------------------------ /usr/local/lmod/lmod/modulefiles/Core ------------------------
   lmod/5.4.2    settarg/5.4.2

  Where:
   (D):  Default Module

Use "module spider" to find all possible modules.
Use "module keyword key1 key2 ..." to search for all possible modules matching any of the
"keys".

[peppe@hscw ~]$
```

MAUI: showq

```
ACTIVE JOBS--------------------
JOBNAME            USERNAME      STATE  PROC   REMAINING            STARTTIME


     0 Active Jobs       0 of   44 Processors Active (0.00%)
                         0 of    7 Nodes Active      (0.00%)

IDLE JOBS----------------------
JOBNAME            USERNAME      STATE  PROC     WCLIMIT            QUEUETIME


0 Idle Jobs

BLOCKED JOBS----------------
JOBNAME            USERNAME      STATE  PROC     WCLIMIT            QUEUETIME

57228                marzio  BatchHold     1 99:23:59:59  Fri Aug  9 11:27:49
57229                marzio  BatchHold     1 99:23:59:59  Fri Aug  9 12:05:19
57232                 gluca  BatchHold     1 20:08:00:00  Sun Aug 18 06:44:32

Total Jobs: 3   Active Jobs: 0   Idle Jobs: 0   Blocked Jobs: 3
```

The very same user view on any physical or virtual cluster

Clusters - Ganglia View



http://mccw.hpc.thch.unipg.it

Herla Project

2014

- Joint project beween CNR-SCITEC and DCBB (UniPG)
  - harvest experience from previous clusters, on obsolete hardware back at work
  - a verified, from scientists, Application Framework for Molecular Sciences HPC
  - state of the art, at the time it has been built, Linux distro (Scientific Linux 6.x)
  - **lmod modules** choosed to semplify user access to compilers, libraries and applications *https://lmod.readthedocs.io/en/latest*

- Clusters had to to be easily configured for:
  - Ethernet, Gbits (GbE/10GbE), easy
  - Infiniband, IB (OFED), 40Gbits, complex

2018
2019

- Its images run today a couple of production clusters
  - HS - CNR/DCBB - virtual cluster hosted from GARR CLOUD
  - TC - CNR SCITEC - production plysical IB cluster

Herla Clusters

http://cgcw.herla.unipg.it

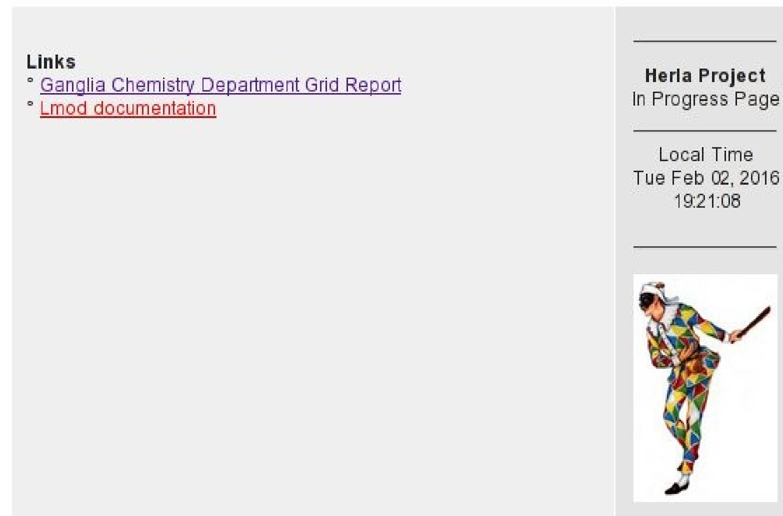
2014

- **Herla clusters**
  - **Chemgrid (CG)** for teaching
  - **FrontEnd (FE)** for HPC

  - **LMOD modules**

    Compilers and Libraries
  - Intel C/Fortran compilers (14.0.2)
  - Intel MKL (blas+lapack) (11.1.2)
  - mpich 3.1 – mvapich 1.9
  - ScaLAPACK 2.0.2

  Applications
  - Espresso (39.62)
  - Gaussian (09-c01)
  - Gamess-US (050113R1)
  - NwChem (6.5.26243)
  - Siesta (3.2-pl-4)
  - MolPro (2010p)



PBS-like Batch SubSystem

Torque Resource Manager (2.5.13)
Maui Scheduler (3.3.4)

FEcw First Virtual Access node (VMWare)
CG+FE – single NIS clusters – cgcw (master)

Herla Physical Clusters

built
2016

CG ChemGrid : 10 nodes - 20 cores - 22Gb Memory - 2Tb Storage - Gbits (32bit)
FE FrontEnd : 13 nodes - 52 cores - 52Gb Memory - 5Tb Storage - Gbits (64bit)

2017

Herla cluster virtual images

- A first step towards the creation of a *cloud image*
  - done under CMS²

    "Consortium for Computational Molecular and Materials Sciences"

    http://www.cms-2.org
  - hosted by Department of Physics and Geology (UniPG) and INFN-PG OpenStack infrastructure

    http://www.fisica.unipg.it http://www.pg.infn.it http://openstack.fisica.unipg.it
  - Herla application framework for Molecular Science HPC

- CW and node images
  - created from Herla physical clusters
  - QEMU images developed, and tested, on a single workstation
  - uing KVM and QEMU under Linux
  - almost complex, as usual, as deploying a tar image

- First succesfull application:
  - PhD School on Open Science Cloud (SOSC-2017)

    http://fisgeo.unipg.it/sosc17 Perugia, June 5-9, 2017

2017

Herla: First Virtual Cluster

- Built to run under OpenStack
  - basically a *testbed*
  - its main goals: verification and benchmarking
  - aim to become a production HPC system, *eventually*, one day
  - still a long road to go

- OpenStack Compute Nodes
  - again, obsolete hardware at work (**HP DL320 G5p)**
  - 15 new compute nodes, **vh01-15**, connected to the FisGEO/INFN running OpenStack
  - a bunch of old nodes, performance were not an issue at this stage
  - with a new dedicated virtual **HS**cw access node hscw.fisica.unipg.it
  - loaded with the Herla cluster QEMU virtual images, hosting the HPC Application Framework

- OpenStack Backend Storage
  - a small CEPH Storage Cluster, just four OSD/MON (Object Storage Daemon and Monitor) nodes
  - to verify CEPH scalability, reliability and performance
  - to harvest experience on this technology
  - it had been a nice trip ;-)

https://ceph.io/ceph-storage

CEPH storage cluster

2017

- CEPH choosed as a **scalable** OpenStack backend storage
  - installed in a very short time, using Ubuntu MAAS and juju
  - really effective for OpenStack admin
  - n.4 **HP DL320 G5p**, really obsolete nodes, as OSD/MON: **[ce01,ce02,ce03,ce04]**
  - four SATA1 OSD at work, 1.5Tb storage online (90Mb/sec SATA max hw bandwidth)
  - performance, as seen from ***each*** virtual node, around 50-70 Mb/sec, not too bad actually for such a small obsolete testbed configuration
  - still running unattended, ***without any maintenance***, after 2 years
  - even with a broken OSD/MON, **ce01** died some months ago ;-(, no data loss so far ;-)

```
    cluster 88cdfbae-10ba-11e7-83ab-c793df6d4379
     health HEALTH_WARN
            1 mons down, quorum 0,1,2 ce02,ce03,ce04
     monmap e2: 4 mons at {ce02=10.13.122.222:6789/0,ce03=10.13.122.223:6789/0,ce04=10.13.
122.224:6789/0,ce06=10.13.122.226:6789/0}
            election epoch 384, quorum 0,1,2 ce02,ce03,ce04
     osdmap e691: 4 osds: 3 up, 3 in
            flags sortbitwise,require_jewel_osds
      pgmap v3137275: 288 pgs, 5 pools, 204 GB data, 53328 objects
            442 GB used, 672 GB / 1114 GB avail
                 288 active+clean
```

Question

WHY virtual clusters?

- You may ask now: “Why bothering about virtual clusters?”
  - they are easy to build, if application images are already available
  - *indipendent,* in some limit, from the underlying bare hardare
  - images can be uploaded on the cloud in a very short time, everywhere virtualization is offered
  - ranging from a local infrastructure, to academic services, to commercial services already offered ***today*** from vendors like Amazon, Google, IBM or even Microsoft

- Not a free lunch, as always is not
  - good performance for Molecular Sciences HPC are not easy to be achieved
  - especially for parallel MPI applications
  - not such simple and cheap to build a local specialized and realiable infrastructure

- My personal 2 cents opinion?
  - virtualization technologies will move from software to hardware
  - in a near future users will probably *require* to upload and download virtual clusters in and from cloud platforms
  - in the meanwhile, better to gain experience on these technologies
  - may already be effective for some application (benchmarking, ITC research, teaching)

Answer

GARR Cloud

Consortium GARR

- GARR Cloud Platform https://cloud.garr.it
  - GARR Cloud Platform offers cloud services to the Italian academic and research community
    (**IaaS**) Infrastructure as a Service – (**DaaS**) Deployment as a Service
  - coordinates a federation of clouds
  - located in national datacenters owned by members of the GARR community,
  - the GARR Cloud allows creating and managing Virtual Machines
  - as deploying cloud applications, like a Virtual HPC Cluster

openstack.

- Virtual Machines
  - GARR Cloud delivers virtual machines
  - running in the data centers of the GARR Federated Cloud
    (eventually under OpenStack, over Ceph Storage clusters)
  - connected through the GARR high speed fiber network

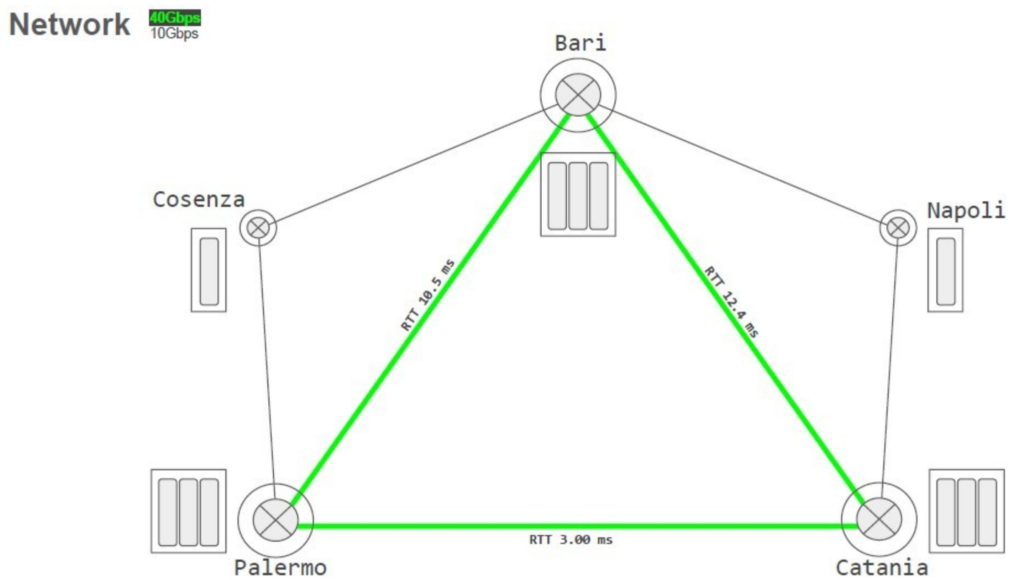
- Virtual Datacenters
  - a Virtual Datacenter consists of a set of virtual resources
  - a set of resources (vCPUs, memory, storage, networking) assigned to an administrator
  - admin can create users and enable them to use the resources assigned to a project

GARR Cloud Infrastructure

openstack.

8500 Cores

10 PetaBytes



Tutorial on OpenStack and the GARR Federated Cloud

A. Barchiesi, A. Colla, G. Marzulli

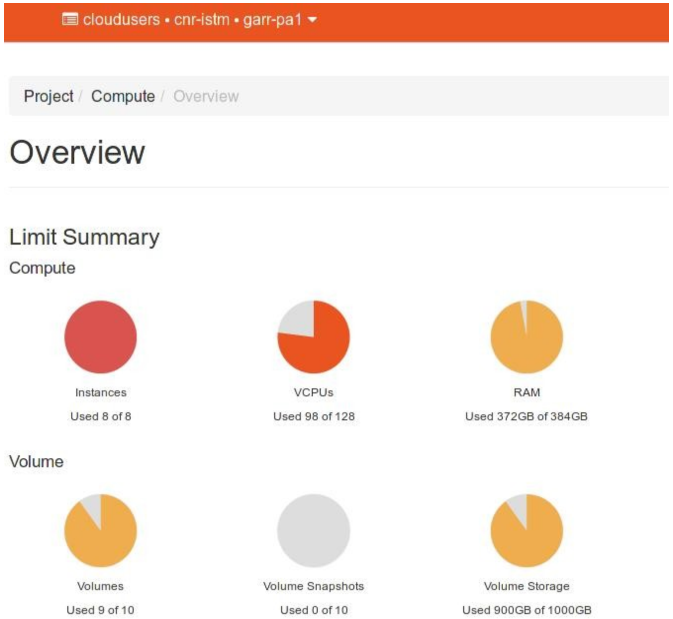
Workshop GARR – Rome 2017

https://www.eventi.garr.it/it/ws17/home/tutti-i-materiali/corso-openstack/132-presentazione-corso-openstack-a-barchiesi-g-marzulli-a-colla

June 2018

GARR Cloud: cnr-istm project

- GARR Cloud: project “cnr-istm” allocated for us
  - 8 instances, 128 vCPUs, 384Gb RAM, 1Tb virtual storage
  - bare hardware resources allocated by GARR-CLOUD in *Palermo*
  - GARR is also our Internet Service Provider, of course, connecting us to the Italian academic network
  - so, we are nicely connected with the cnr-istm project resources



July 2018

Cnr-istm (now SCITEC) project

- A new Virtual Cluster
  - Herla 2017 virtual images ready for the GARR-CLOUD environment
  - again, configured using KVM/QEMU, in a couple of days
  - and uploaded to the GARR Cloud Academic National Platform
  - very few changes, some minor glitch corrected

☐	Instance Name	Image Name	IP Address	Flavor	Key Pair	Status
☐	hs07	-	192.168.100.115	m1.xlarge	peppe	Active
☐	hs06	-	192.168.100.107	Not available	peppe	Active
☐	hs05	-	192.168.100.114	m1.xxl	peppe	Active
☐	hs04	-	192.168.100.110	m1.xxl	peppe	Active
☐	hs03	-	192.168.100.112	m1.xxl	peppe	Active
☐	hs02	-	192.168.100.103	m1.xxl	peppe	Active
☐	hscw	-	192.168.100.109 Floating IPs: 90.147.189.20	m1.medium	peppe	Active
☐	hs01	-	192.168.100.104	m1.xxl	peppe	Active

Virtual HERLA born

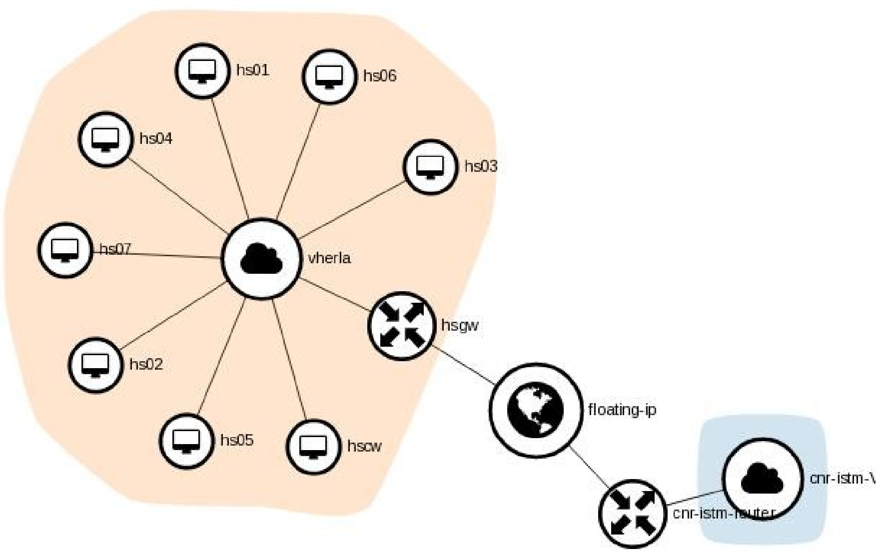
- HS Virtual Cluster: VHERLA
  - HScw: VHERLA cluster access node
    2 vCPUs, 4Gb RAM, 40Gb system volume + 100Gb user volume
  - n.7 compute nodes, 96 vCPUs (16x5+8x2), 368Gb (64x5+32+16) RAM, 512Gb scratch space

Torque Nodes

```
hs01 np=16  t100 x86_64  64Gb
hs02 np=16  t100 x86_64  64Gb
hs03 np=16  t100 x86_64  64Gb
hs04 np=16  t100 x86_64  64Gb
hs05 np=16  t100 x86_64  64Gb
hs06 np=8   t100 x86_64  32Gb
hs07 np=8   t100 x86_64  16Gb
```

Maui view

```
ACTIVE JOBS--------------------
JOBNAME            USERNAME      STATE  PROC   REMAINING

     0 Active Jobs        0 of   96 Processors Active (0.00%)
                          0 of    7 Nodes Active      (0.00%)
```



VHERLA, an HPC Virtual Cluster

- VHERLA: a virtualized Molecular Science HPC Cluster
  - now running in the Cloud, the Virtual Datacenter provided by GARR
  - on state of the art physical bare processors

    Intel Xeon E3-12xx v2, 2.6Ghz cores
  - over a well performing gigabit network
  - protected by a firewall
  - SSH access to control workstation HScw, world-wide: “ssh hscw.herla.unipg.it”
  - WWW http server, offer “ganglia” interface - http://hscw.herla.unipg.it

- Benchmarking
  - nice performances for *non parallel*, *production level*, HPC Applications
  - a cluster which may be easily replicated at any cloud service provider
  - which may be maintained remotely from virtual consoles
  - definitely a positive experience
  - it had been already succesfully used for real applications

VHERLA Molecular Science Applications

Aug 2018

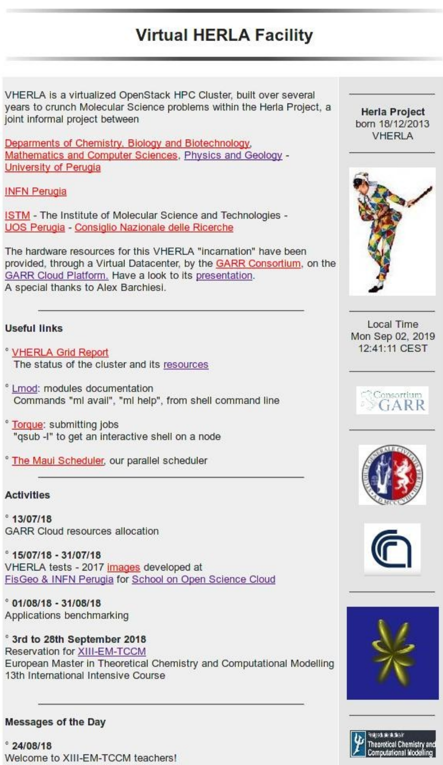
- VHERLA Molecular Sciences HPC Application Framework passed base tests and had been benchmarked:
  - QUANTUM ESPRESSO/39.62
  - GAUSSIAN/09-c01-omp
  - NWCHEMS/6.5.26243
  - GAMESS-US/050113R1
  - MOLPRO/2010p-omp
  - SIESTA/3.2-pl-4

- Results?
  - applications has been *verified* by our research group scientists
  - performances, for *non parallel* MPI applications: quite good
  - the previous, first attempt, ***Herla First Virtual Cluster***, just a *toy* compared with this new incarnation of the very same images
  - *general purpose* gigabit networking is simply not enough for real parallel MPI, production level applications, compared to our Infiniband (40Gbits) production physical clusters
  - can't be expected from a general purpose Cloud Service Provider!

Sept 2018

VHERLA: applications

- XIII-EM-TCCM
  - European Master in Theoretical Chemistry and Computational Modelling
    13th International Intensive Course – Perugia, September 3-28, 2018
    http://www-old.chm.unipg.it/chimgen/mb/theo2/TCCM2018/EM-TCCM2018/EM-TCCM/Welcome.html



Virtual HERLA Facility

VHERLA is a virtualized OpenStack HPC Cluster, built over several years to crunch Molecular Science problems within the Herla Project, a joint informal project between

Deparments of Chemistry, Biology and Biotechnology, Mathematics and Computer Sciences, Physics and Geology - University of Perugia

INFN Perugia

ISTM - The Institute of Molecular Science and Technologies - UOS Perugia - Consiglio Nazionale delle Ricerche

The hardware resources for this VHERLA "incarnation" have been provided, through a Virtual Datacenter, by the GARR Consortium, on the GARR Cloud Platform. Have a look to its presentation.
A special thanks to Alex Barchiesi.

Herla Project
born 18/12/2013
VHERLA

Local Time
Mon Sep 02, 2019
12:41:11 CEST

Useful links

- VHERLA Grid Report
  The status of the cluster and its resources
- Lmod: modules documentation
  Commands "ml avail", "ml help", from shell command line
- Torque: submitting jobs
  "qsub -I" to get an interactive shell on a node
- The Maui Scheduler, our parallel scheduler

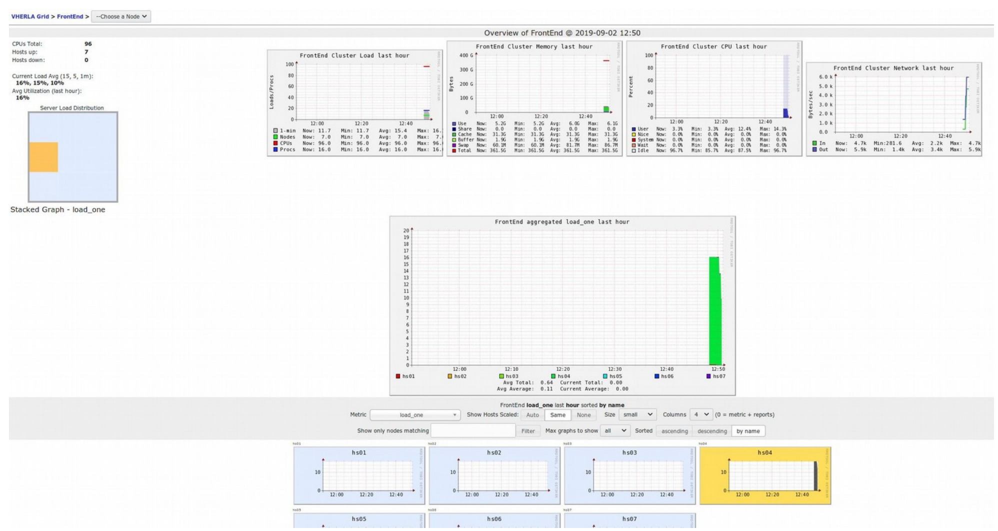
Activities

- **13/07/18**
  GARR Cloud resources allocation
- **15/07/18 - 31/07/18**
  VHERLA tests - 2017 images developed at FisGeo & INFN Perugia for School on Open Science Cloud
- **01/08/18 - 31/08/18**
  Applications benchmarking
- **3rd to 28th September 2018**
  Reservation for XIII-EM-TCCM
  European Master in Theoretical Chemistry and Computational Modelling
  13th International Intensive Course

Messages of the Day

- **24/08/18**
  Welcome to XIII-EM-TCCM teachers!
  Your accounts are ready.

http://hscw.herla.unipg.it



- Last, but not least, ***teaching***
  students seems to like playing with it

2018-2019

A local (UniPG) cloud platform

(CMS)² CONSORTIUM FOR COMPUTATIONAL MOLECULAR AND MATERIAL SCIENCES

- Quite satisfied of GARR-CLOUD VHERLA testbed
  - we began to think about a local OpenStack Cloud specialized infrastructure
  - wich may run other *future* virtual incarnations of VHERLA cluster
  - built around our requests and constraints, for our applications
  - designed to achieve the perfomance we need for HPC MPI parallel Molecular Science Applications, of the same order of our latest production level physical clusters
  - to be, possibly, in a far future, federated with the National GARR-CLOUD academic service or with other local Perugia cloud platforms

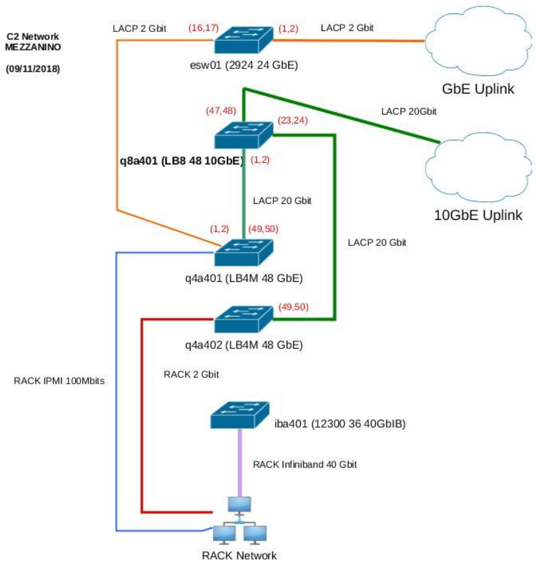
- GbE networking is not enough for us, latency too high, bandwith too narrow
  - we used, once again, hardware we already own, not such obsolete, to build a new Infiniband cluster
  - which may host our experiments on OpenStack and CEPH storage technologies
  - moving from a GbE copper backbone to a 10GbE fiber backbone, to be able to distribute cloud services, at least in our Department, at full bandwidth
  - to allow to interconnect local UniPG clusters, in a near future

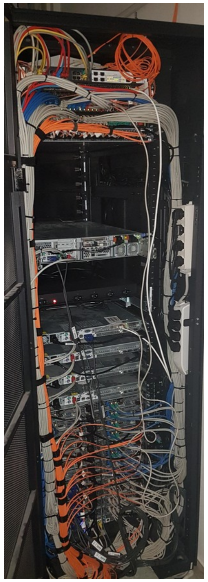
2018-2019

A glance to the future

RACK IBM C2 (MEZZ)

Switch 24 porte	GbE	esw01	1U
Switch 48 porte	10GbE	q8a401	1U
Switch 48 porte	GbE	q4a401	1U
Switch 48 porte	GbE	q4a402	1U
Switch 36 porte	40GbIB	iba401	1U
(tantalo)	IBM x346		2U
Console			1U
R1	(vh01)		1U
R2	(vh03)		1U
R3	(vh04)		1U
R4	(vh05)		1U
(mc15,mc16)	(→ D12)		1U
(mc23,mc24)	(→ D08)		1U
(mc25,mc26)	(→ D07)		1U
(mc29,mc30)	(→ D05)	IB broken	1U
(mc35,mc36)	(→ D02)		1U
(mc37,mc38)	(→ D01)		1U
(mc01,mc02)	(→ D19)		1U
(mc03,mc04)	(→ D18)	CEPH	1U
(mc05,mc06)	(→ D17)	CEPH	1U
(mc07,mc08)	(→ D16)	CEPH	1U
(mc11,mc12)	(→ D14)	CEPH	1U
(mc13,mc14)	(→ D13)	CEPH	1U
(titone)	HP ProLiant DL380 G3		2U





- August 2019: bare hardware installed
  - still to be tested and verified
  - connected to the *new* 10GbE fiber backbone
  - enough resources to host OpenStack systems, a real CEPH 80Tb Storage Cluster and a bunch of OpenStack Compute Nodes, connected through Infiniband