# GESTIONE DI APPLICAZIONI DI CALCOLO ETEROGENEE CON STRUMENTI DI CLOUD COMPUTING: ESPERIENZA DI INFN-TORINO

S.Bagnasco, D.Berzano, R.Brunetti, M.Concas, S.Lusso

# Motivations

- During the last years, the amount of “consumer level” resources devoted to the scientific computing has increased
  - Unfortunately the man power did not !
- It is becoming almost mandatory to consolidate these resources in order to achieve scalability and economy of scale
  - Our Data Centers must assume the role of real “providers” of computing and storage resources, not only of high level services
- The Cloud approach (IaaS) might help to better manage the resources provisioning to the different clients (ie. Grid Sites, small or medium farms, single users)
- Many cloud computing projects are starting both at national and European level
  - A local working cloud infrastructure will also allow to take part to these activities

# Our Philosophy

- Stay mainstream
- Keep it simple
- Interoperability
- Be user oriented
- Prefer flexibility and manageability over performances

# Software Tools

- Cloud management Toolkit (OpenNebula) (V 3.6)
  - Open Source stack with a wide user community
  - Modular architecture
  - Already satisfies most of the requirements in term of functionalities
  - Easy to customize (mostly shell and ruby scripts)
- Backend storage (GlusterFS) (V 3.3)
  - Easy to setup in a basic configuration
  - Robust
  - Scalable
- VM network management (OpenWRT)
  - Light-weighted Linux distribution for embedded systems with tools for the network configuration and management

# Multipurpose storage: GlusterFS

- GlusterFS mimics many RAID functionalities at filesystem level by aggregating single “bricks” on different machines:
  - Modes: distributed, replicated, striped (can be combined)
- Our use cases:
  - VM image repository: one brick exported
  - System datastore: replicated on two servers for redundancy. Replica is synchronous, self-healing enabled. Continuous r/w occurs
  - Experiment data: pool of disks aggregated (~50 TB). Very high throughput towards many concurrent clients
- Horizontal scalability:
  - no master host → all synchronizations are peer-to peer
- Easy management:
  - On-line addition, removal, replacement of bricks

# Clusters

Depending on which kind of host has to be instantiated, we identified two types of clusters

**Services**

- VMs providing critical services with in/outbound connectivity, public IP, live-migration etc.. but with limited disk I/O

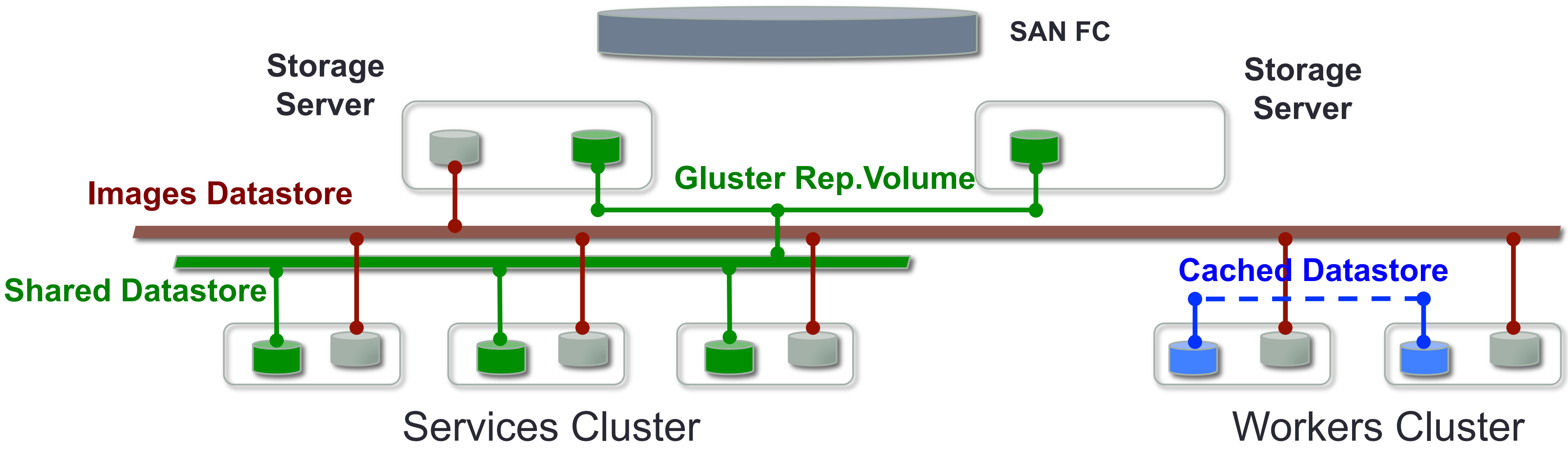
**Workers**

- VMs providing non critical computing services (like WNs) with private IP only but requiring high disk I/O capacity.

These two classes have been provided with different types of backend storage (“Datastores” in the OpenNebula terminology) in order to optimize the performances and satisfy the above requirements.

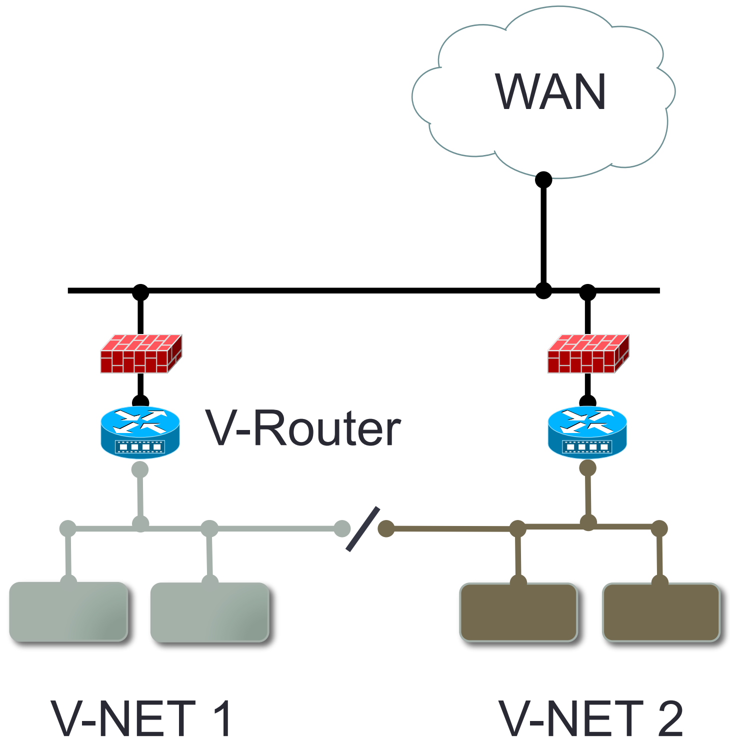
# Backend Storage

- **Two storage servers with 10Gbps interface provide some of the LUNs through GlusterFS**
  - Services System Datastore is shared to allow live migration of the machines.
  - Workers System Datastore is local to the hypervisors disks in order to increase I/O capacity. Images are cached locally to increase startup speed
  - An ad-hoc script synchronizes the local copies using a custom “torrent-like” tool when new versions of the images are saved
  - All the virtual machines run on RAW or QCOW file images.

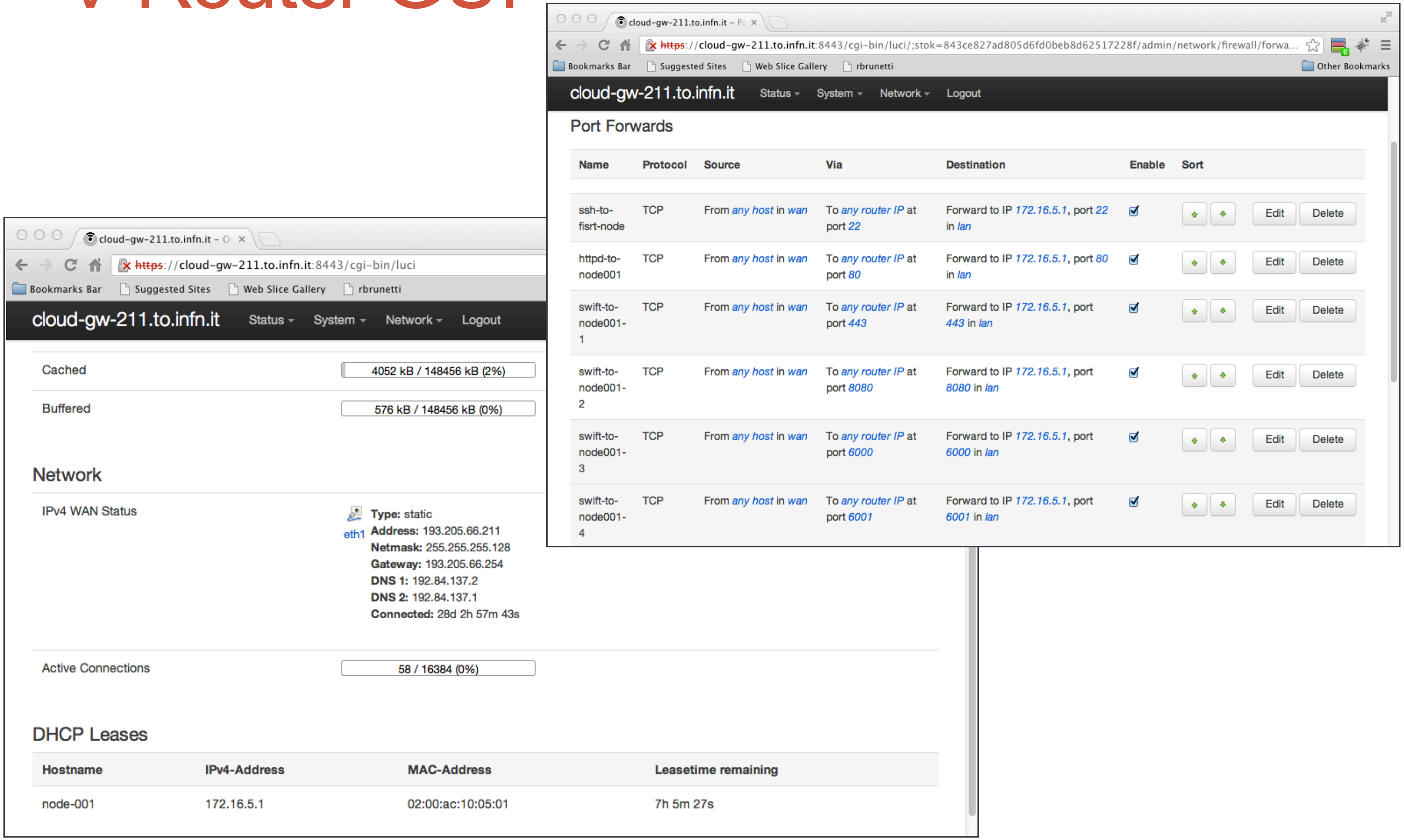


# Networking

- Network Isolation (Level 2)
  - Each user has its own Virtual Network, isolated using “ebtables” rules defined on the hypervisor bridge (OpenNebula V-net driver takes care of this).
- Virtual Routers (Level 3)
  - We prepared a light-weighted VM image (1 CPU/150 MB Ram) starting from a linux distro designed for embedded systems (OpenWRT).
    - DHCP Server, DNS Server, NAT
    - Firewalling/Port Forwarding
  - This provides the user with a dedicated fully featured class-C network whose connectivity remains under our control (the user has no access to the V-Router)



# V-Router GUI


cloud-gw-211.to.infn.it
Status
System
Network
Logout
Port Forwards
Name | Protocol | Source | Via | Destination | Enable | Sort
ssh-to-fisrt-node | TCP | From any host in wan | To any router IP at port 22 | Forward to IP 172.16.5.1, port 22 in lan
httpd-to-node001 | TCP | From any host in wan | To any router IP at port 80 | Forward to IP 172.16.5.1, port 80 in lan
swift-to-node001-1 | TCP | From any host in wan | To any router IP at port 443 | Forward to IP 172.16.5.1, port 443 in lan
swift-to-node001-2 | TCP | From any host in wan | To any router IP at port 8080 | Forward to IP 172.16.5.1, port 8080 in lan
swift-to-node001-3 | TCP | From any host in wan | To any router IP at port 6000 | Forward to IP 172.16.5.1, port 6000 in lan
swift-to-node001-4 | TCP | From any host in wan | To any router IP at port 6001 | Forward to IP 172.16.5.1, port 6001 in lan
Edit
Delete
Cached
4052 kB / 148456 kB (2%)
Buffered
576 kB / 148456 kB (0%)
Network
IPv4 WAN Status
eth1
Type: static
Address: 193.205.66.211
Netmask: 255.255.255.128
Gateway: 193.205.66.254
DNS 1: 192.84.137.2
DNS 2: 192.84.137.1
Connected: 28d 2h 57m 43s
Active Connections
58 / 16384 (0%)
DHCP Leases
Hostname | IPv4-Address | MAC-Address | Leasetime remaining
node-001 | 172.16.5.1 | 02:00:ac:10:05:01 | 7h 5m 27s

# Stakeholders

### Grid T2 Site

- All the Grid services (CE,SE,BDII ecc..) + xx job slot (WNs contextualize at boot time)

### ALICE Analysis Facility (PROOF)

- Shares the resources with the Grid T2, nodes are instantiated on-demand and contextualize at boot time

### M5L-CAD

- INFN & diXit spin-off for the automatic analysis of tomographic images
- Self provisioning of VM

### BES III Experiment

- Research group testing some ad-hoc WNs to understand how to use the Grid for their computing needs
- Plan to have a full BES-III Tier-2 next year

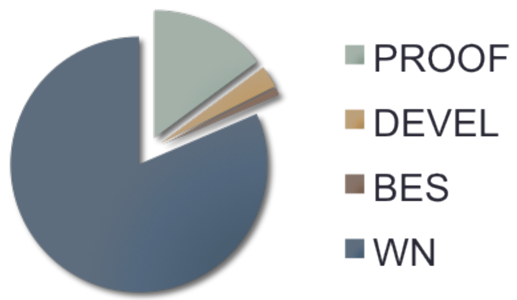
### Development Machines

- DGAS accounting development.
- Test of various services (ie. new versions of OpenNebula itself)

# Resources

## Physical Nodes:

- 31 host (KVM hypervisors)
  - 27 dedicated to "Workers" cluster
  - 4 dedicated to "Services" cluster
- 1 cloud controller (master)

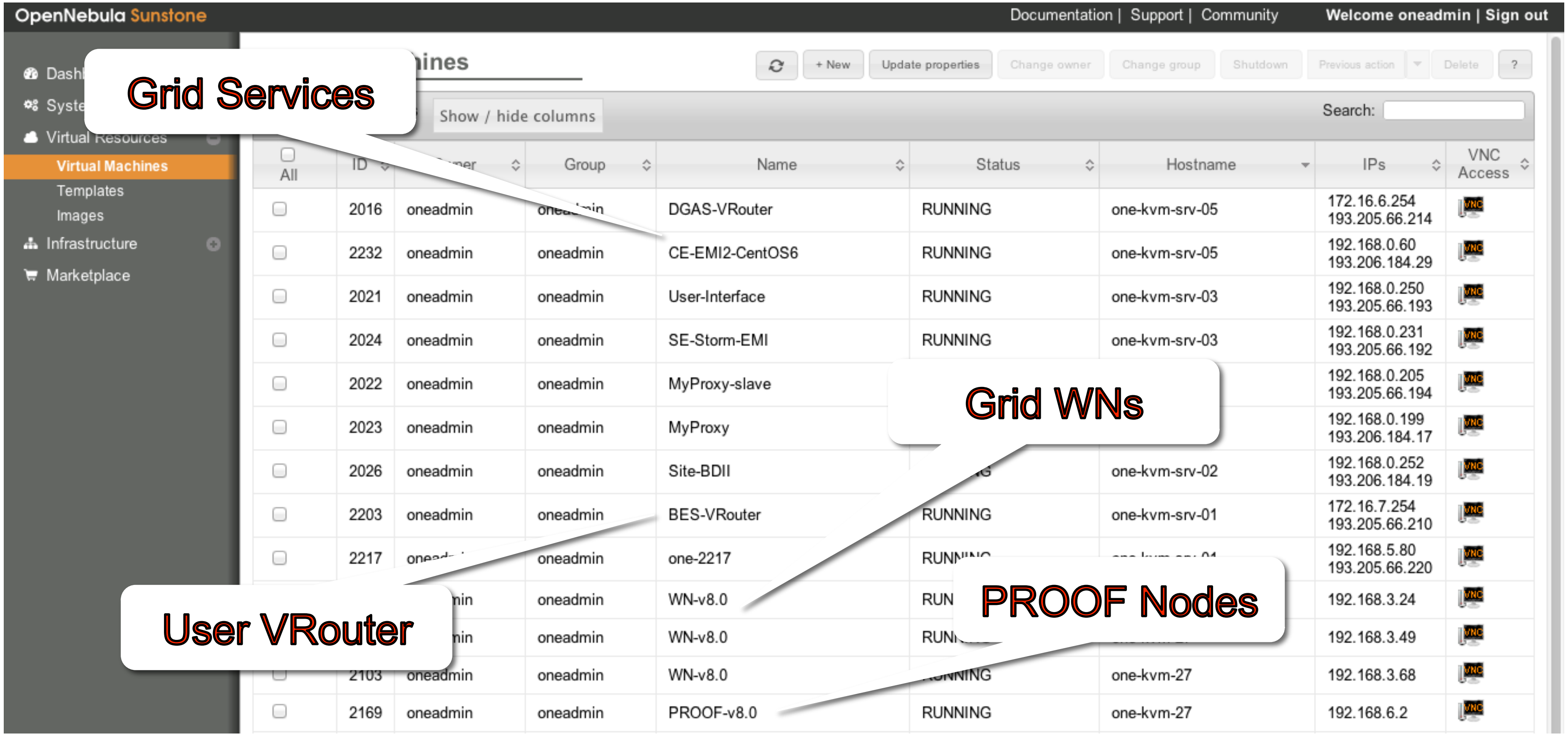



## Storage:

- 2 storage server with 10Gbps Ethernet interface (≈ 20TB SAN frontend)

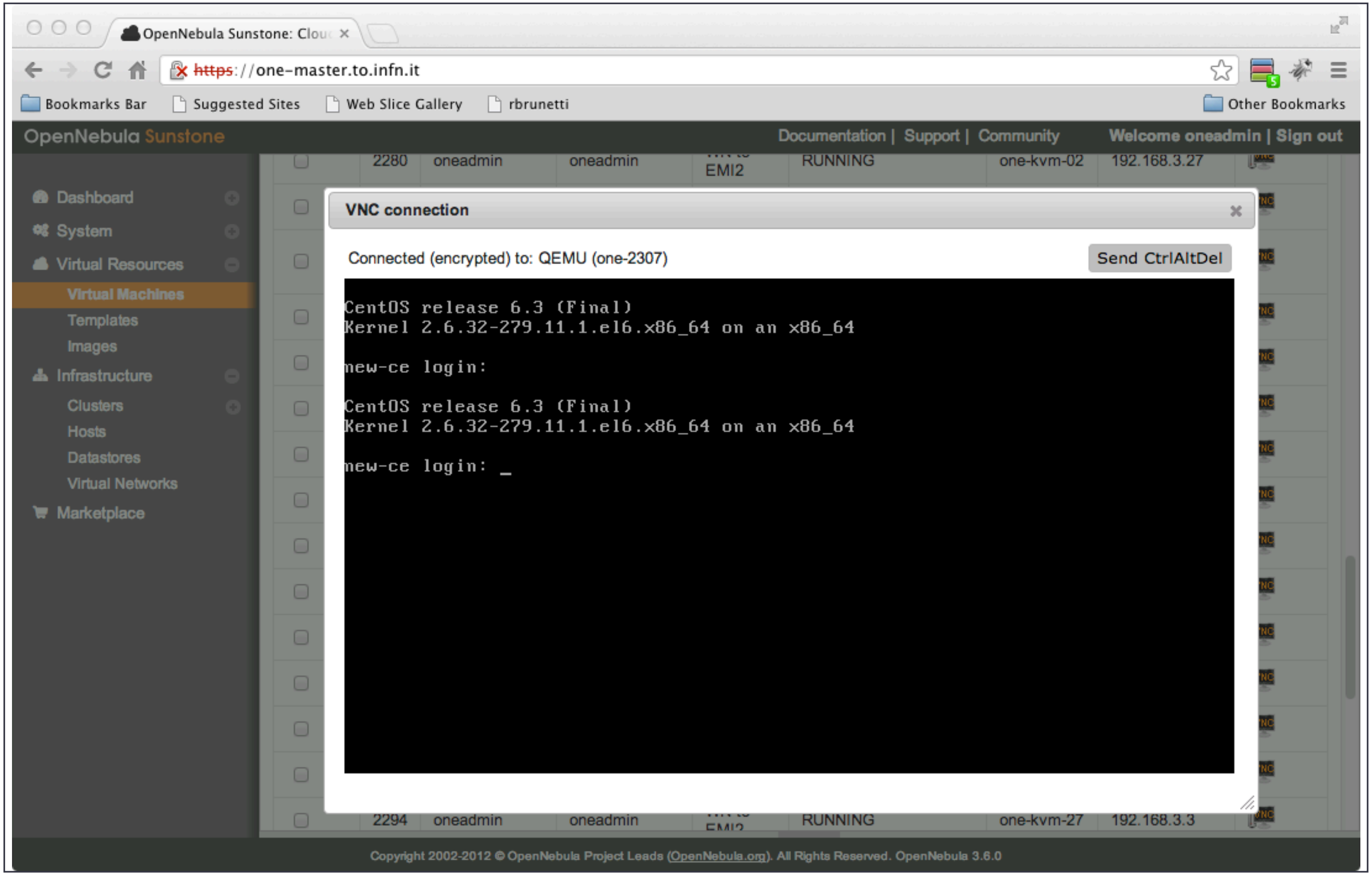
About 650 cores …. and growing

# Dashboard



OpenNebula Sunstone
Documentation | Support | Community
Welcome oneadmin | Sign out
Grid Services
Grid WNs
User VRouter
PROOF Nodes
Virtual Resources
Virtual Machines
Templates
Images
Infrastructure
Marketplace
+ New
Update properties
Change owner
Change group
Shutdown
Previous action
Delete
Show / hide columns
Search:
All
ID
Group
Name
Status
Hostname
IPs
VNC Access
2016 oneadmin DGAS-VRouter RUNNING one-kvm-srv-05 172.16.6.254 193.205.66.214
2232 oneadmin oneadmin CE-EMI2-CentOS6 RUNNING one-kvm-srv-05 192.168.0.60 193.206.184.29
2021 oneadmin oneadmin User-Interface RUNNING one-kvm-srv-03 192.168.0.250 193.205.66.193
2024 oneadmin oneadmin SE-Storm-EMI RUNNING one-kvm-srv-03 192.168.0.231 193.205.66.192
2022 oneadmin oneadmin MyProxy-slave 192.168.0.205 193.205.66.194
2023 oneadmin oneadmin MyProxy 192.168.0.199 193.206.184.17
2026 oneadmin oneadmin Site-BDII one-kvm-srv-02 192.168.0.252 193.206.184.19
2203 oneadmin oneadmin BES-VRouter RUNNING one-kvm-srv-01 172.16.7.254 193.205.66.210
2217 oneadmin one-2217 192.168.5.80 193.205.66.220
oneadmin WN-v8.0 192.168.3.24
oneadmin WN-v8.0 192.168.3.49
2103 oneadmin oneadmin WN-v8.0 one-kvm-27 192.168.3.68
2169 oneadmin oneadmin PROOF-v8.0 RUNNING one-kvm-27 192.168.6.2

# Remote VM Control



# Future Developments

- Test the possibility to turn to some form of hybrid/ public/distributed infrastructure using:
  - OCCI Interface (preferred choice due to interoperability with other cloud stacks)
  - OpenNebula “Zones”
- Study the integration of the OpenNebula Authn/Authz system in a VO context or using federated authentication mechanisms.
- Explore the GlusterFS *UFO* (Object Storage) to provide a “DropBox-like” storage to the users.
- Participate in the INFN projects aimed to develop a higher level distributed infrastructure

# Conclusions

- The experience of the setup of a private cloud in the context of the INFN-Torino computing center using OpenNebula + GlusterFS has been positive.
- The infrastructure is now running production services and the idea is to extend it and make it a “service” for all the local computing needs.
- We are very interested to collaborate in every activity aimed to develop and test the different cloud stack interoperability using:
  - Common images repositories (usage criteria, validation)
  - Standard interfaces (OCCI)
  - Federated Authn/Authz systems
- ..and to participate in the upcoming effort to explore clouds as a resource for scientific computing
  - And possibly to build a federated higher-level infrastructure

# THANK YOU

# User Management

## Users and Groups

- Now
  - Users have a simple Username/Password account on the cloud UI and are assigned to groups
  - User groups have access to selected VM images,V-net etc..
- Future
  - Moving to some stronger Authn/Authz method (Shibboleth, X509 etc..)

## Quotas

- The amount of resources that a given group can instantiate is limited by quotas on CPU number, RAM and storage.

# Customizations

- Most of the requirements of our use cases were already satisfied by the OpenNebula features
- Whenever we needed something more, we tried to stay “mainstream” adding (and not modifying) some of the administrative tools, but always using the OpenNebula APIs
  - Slightly modified the “onevm” command in order to provide some more information
  - Prepared some ad-hoc contextualization scripts to manage the automatic adding/removing of Grid WNs or PROOF nodes
  - Prepared the synchronization script “torrent-like” to manage the images in the Cached Datastore (Workers cluster)

# Datastores

- Represent the real “backbone” of the cloud infrastructure, providing the two main components:
  - Virtual machine images repository
    - **Images Datastore** : Simple GlusterFS Volume mounted on all hypervisors
  - Storage for the running virtual machine images
    - **System Datastore** : Replicated GlusterFS Volume available on all “Services” hypervisors
    - **Cached Datastore** : Local storage on all “Workers” hypervisors
      - VMs run on top of a qcow snapshot.
      - An ad-hoc script takes care to synchronize the local copies using rsync (scpwave) when new versions of the images are saved