MANAGING A TIER 2 COMPUTER CENTRE WITH A PRIVATE CLOUD INFRASTRUCTURE

Stefano Bagnasco, Riccardo Brunetti, Stefano Lusso (INFN-Torino), Dario Berzano (CERN)

ACAT2013 |Beijing, May 16-21, 2013

The amount of resources and the variety of applications is steadily increasing, manpower unfortunately is not

- It is becoming almost mandatory to consolidate such resources to achieve scalability and economies-of-scale
  - Separate application management from infrastructure management
  - Our Data Centres need to become providers of computing and storage resources, not (only) of high level services
- The Cloud approach (IaaS) might help to better provision resources to the different scientific computing applications
  - Grid Sites, small or medium computing farms, single users,...
- Several cloud computing projects are starting at national and European level
  - From definition of best practices and reference configurations to deployment of large-scale distributed infrastructures
  - A local working cloud infrastructure will also allow to take immediately part in such activities

PHILOSOPHY

- Ensure **interoperability**
- Favour **manageability** and **flexibility** over performance
- Provide a **production service** to applications

Keep it simple

Stay mainstream

Be user-oriented

PHILOSOPHY

- Don't use too many tools
- Develop as few pieces as possible
- Introduce features only when needed by applications
- Use few simple images plus contextualization

Keep it simple

Stay mainstream

Be user-oriented

PHILOSOPHY

Choose stable and widely used tools and components:

- OpenNebula cloud stack
  - Common interfaces: OCCI, EC2, OCA
- GlusterFS filesystem
- OpenWRT for network management

Keep it simple

Stay mainstream

Be user-oriented

PHILOSOPHY

- Adopt an **agile development** cycle
- Give resources to users as soon as possible
- Add functionalities as they become needed

Keep it simple

Stay mainstream

Be user-oriented

TWO CLUSTERS

VMs providing **critical services**:

- in- & out-bound connectivity
- public & private IP
- live migration
- no special I/O requirements

VMs providing **computing workforce**:

- example: Grid WNs
- private IP only
- high storage I/O performance

TWO CLUSTERS

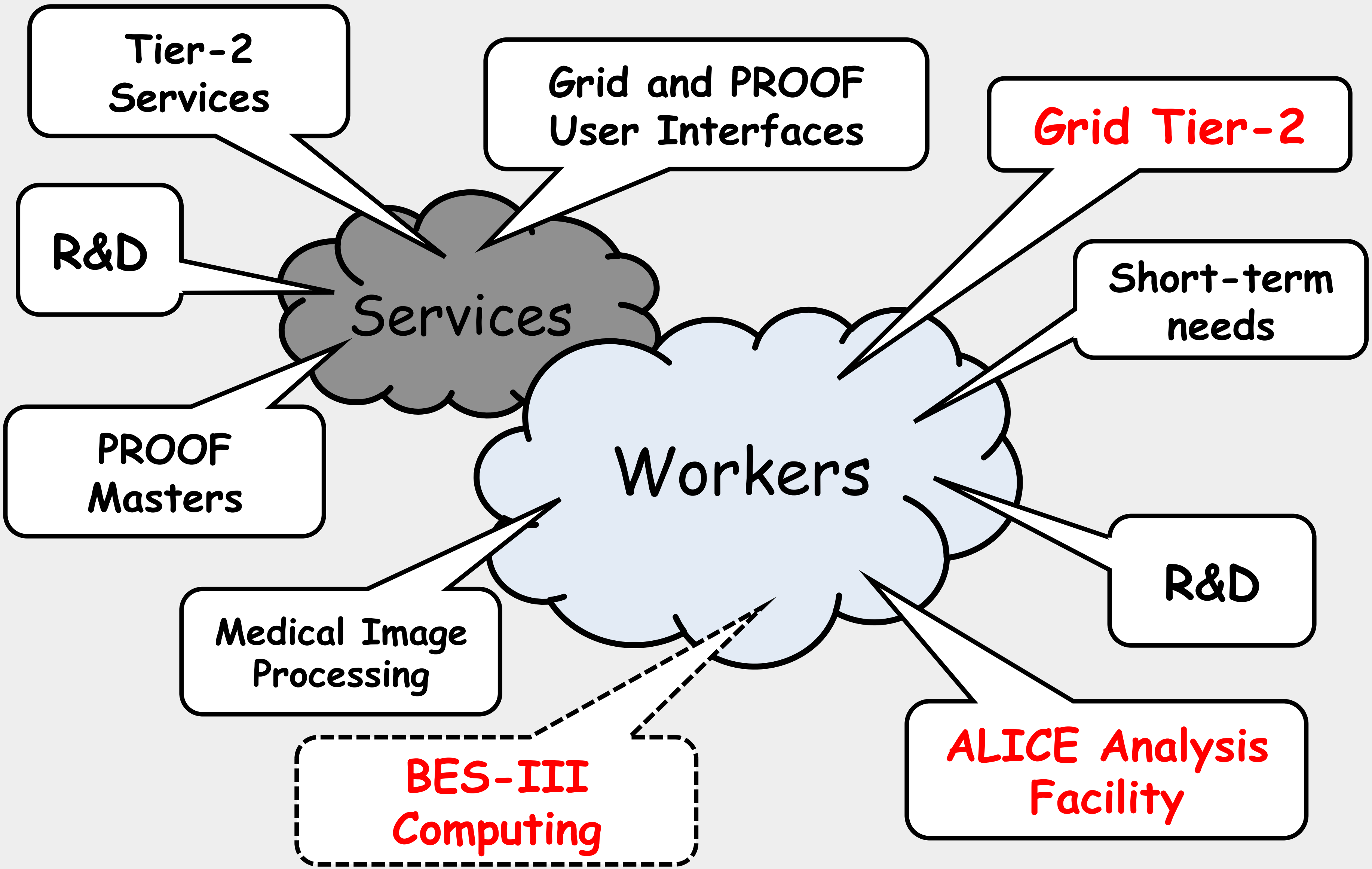
- Server-class hardware
- Shared image repository
- Resiliency-optimized FS for shared system disks
- Currently 4 hosts

Services

- Working-class hardware ☺
- Cached image repository
- Access to perfomance-optimized FS for data needs
- Currently 35 hosts

Workers

STAKEHOLDERS



- Cloud management Toolkit: **OpenNebula**
  - Open Source stack with a wide user community
  - Modular architecture
  - Already provides most of the required functionalities
  - Uses "standard" interfaces (EC2, OCCI, OCA)
  - Easy to customize (mostly shell and ruby scripts)
  - OpenStack, now widely adopted in new projects, was too embrionic when we started
  - ...and arguably* OpenNebula is better suited at Data Center Nebulization
  - Currently using version 3.6, will migrate to 3.8 soon (or 4.0, available since last week)
  - We use templates based on few very simple images plus full contextualization via context scripts and puppet (looking into CloudInit)

* See e.g. blog.opennebula.org/?p=4042

THE INGREDIENTS

- Backend storage: **GlusterFS**
  - Easy to setup in a basic configuration
  - Flexible enough to cater to different needs with a single tool (see next slides)
  - Proven robustness and scalability

- VM network management: **OpenWRT**
  - Light-weight Linux distribution for embedded systems
  - Provides tools for network configuration and management
  - We deploy "VRouters" for virtual clusters
  - Again, OpenVSwitch was not integrated in OpenNebula when we started

MULTIPURPOSE STORAGE: GLUSTERFS

GlusterFS mimics RAID functionalities at filesystem level by aggregating “bricks” on different machines:

- distributed
- replicated
- striped (can be combined)

- Horizontal scalability:
  - no master host, all synchronizations are peer-to peer
  - clients access data directly from the node hosting it
- Easy management:
  - On-line addition, removal, replacement of bricks

MULTIPURPOSE STORAGE: GLUSTERFS

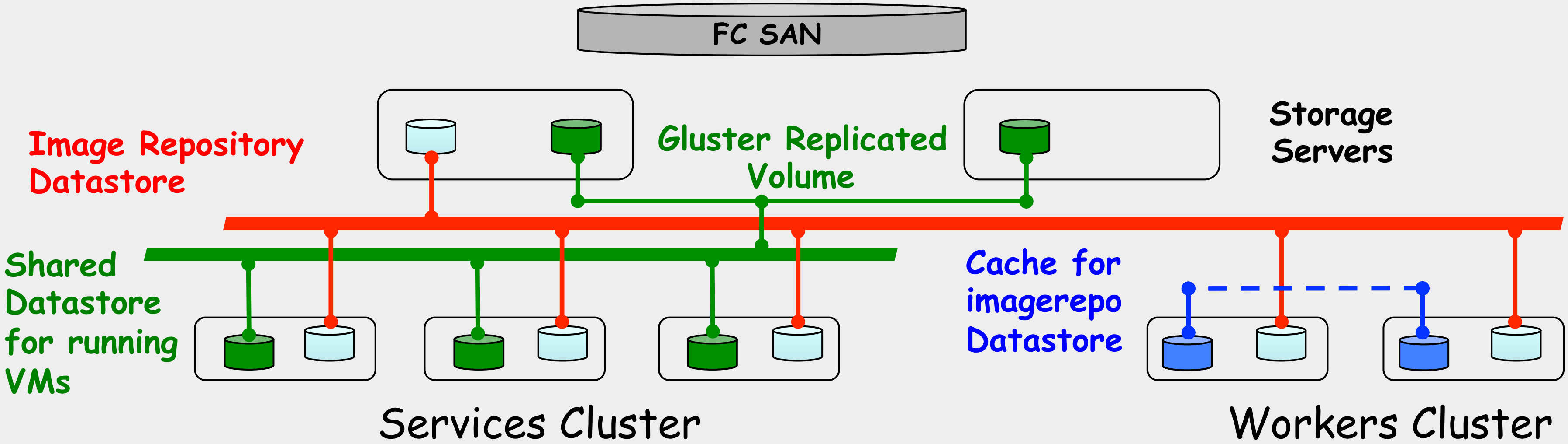
Our use cases:

- VM image repository:
  - one brick exported
- System datastore for service-class hosts:
  - replicated on two servers for redundancy.
  - Replica is synchronous, self-healing enabled.
  - Continuous r/w occurs
- Experiment data
  - pool of aggregated disks (currently ~50 TB).
  - Very high throughput towards many concurrent clients

MULTIPURPOSE STORAGE: GLUSTERFS

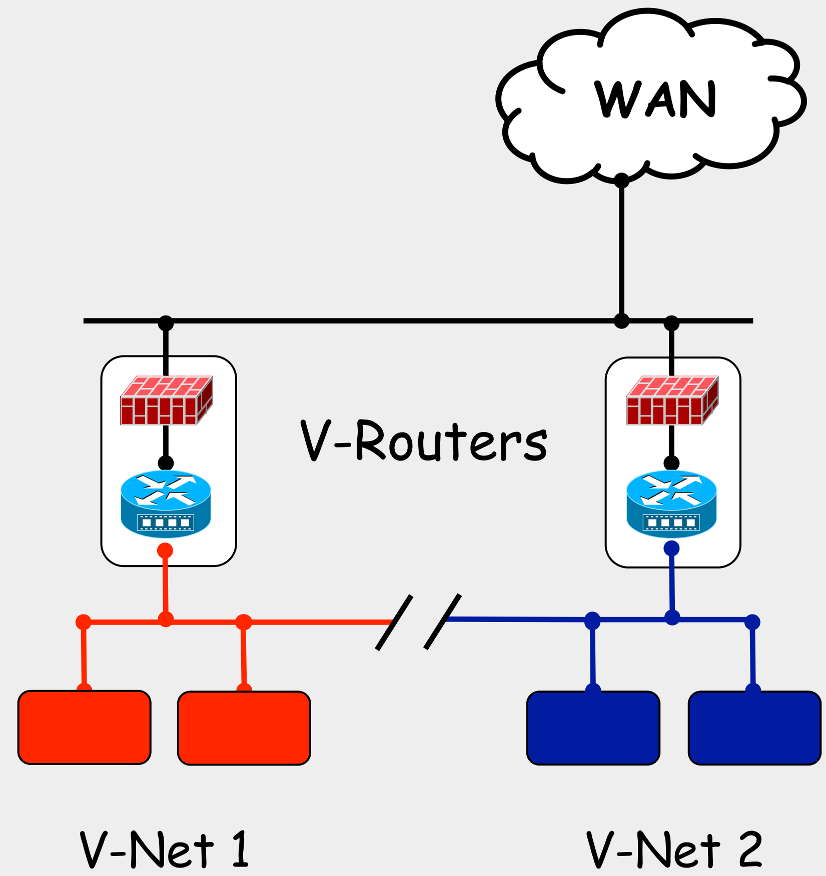
Two storage servers with 10Gbps interface provide some of the LUNs through GlusterFS

- All the virtual machines run on RAW or QCOW file images
- Services System Datastore is **shared** to allow live migration
- Workers System Datastore is **local** to the hypervisors to increase I/O capacity. Images repository is locally **cached** on each hypervisor to reduce startup time.
    - An ad-hoc script synchronizes the local copies using a custom “torrent-like” tool (scpWave + rsync) when new versions of the images are saved

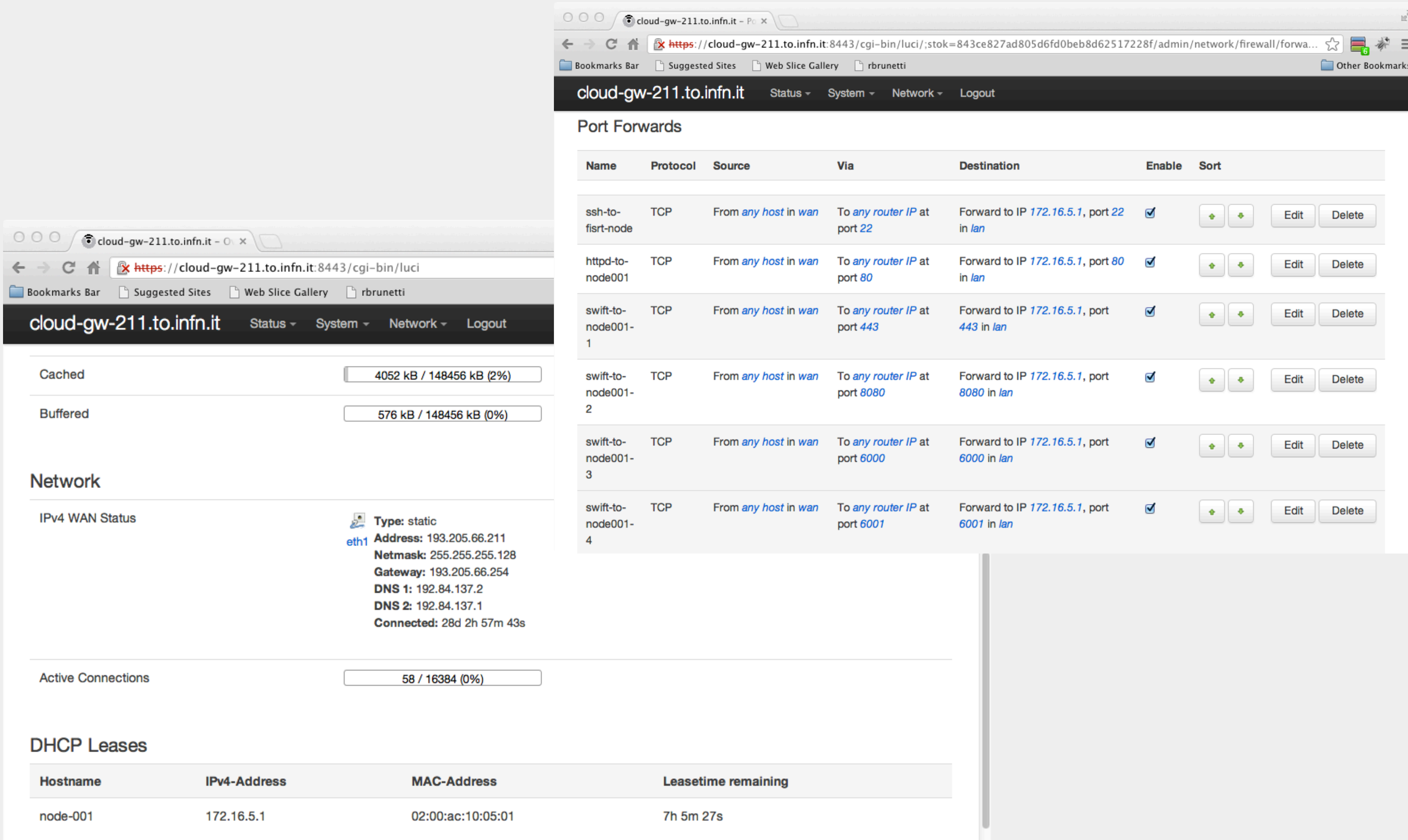


NETWORKING

- 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)
  - Lightweight VM image (1 CPU, 150 MB Ram) with a Linux distribution designed for embedded systems
  - DHCP Server, DNS Server, NAT
  - Firewalling/Port Forwarding
- This provides the user with a **dedicated fully featured class-C network** while the connectivity remains under our control (the user has no access to the V-Router)

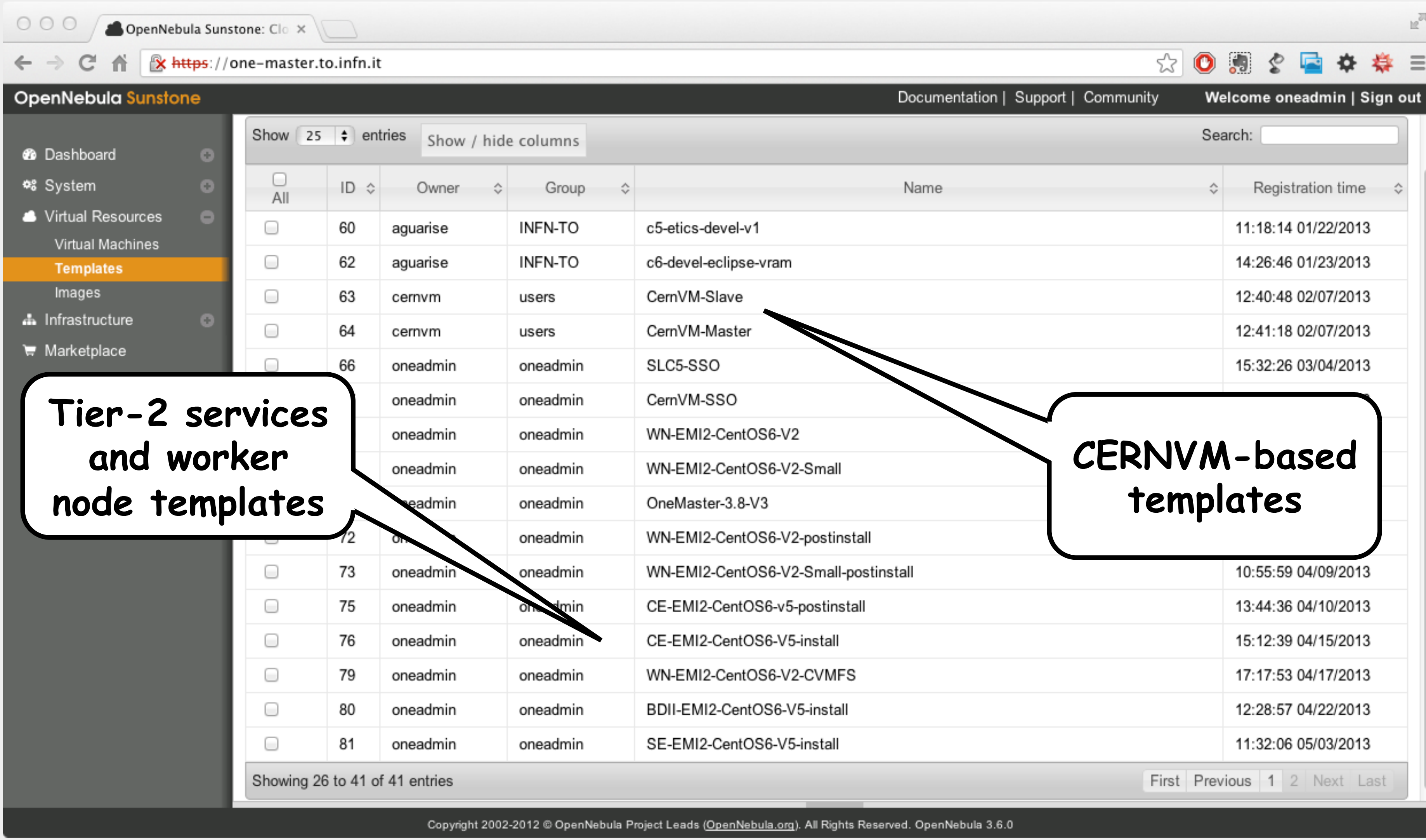


VROUTER GUI



cloud-gw-211.to.infn.it
https://cloud-gw-211.to.infn.it:8443/cgi-bin/luci/;stok=843ce827ad805d6fd0beb8d62517228f/admin/network/firewall/forwa…
Bookmarks Bar
Suggested Sites
Web Slice Gallery
rbrunetti
Other Bookmarks
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
Edit
Delete
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
https://cloud-gw-211.to.infn.it:8443/cgi-bin/luci
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

SUNSTONE DASHBOARD

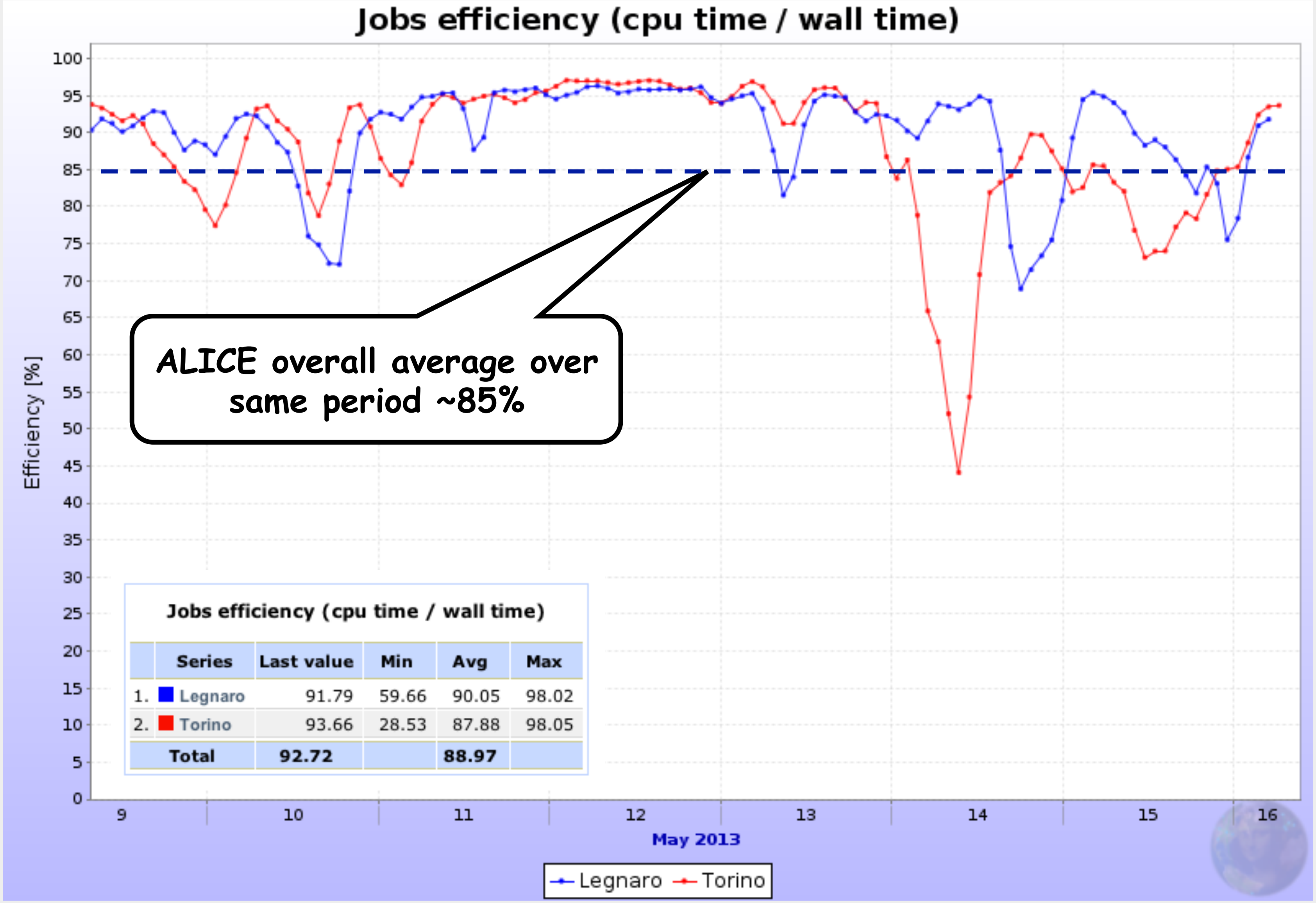


All	ID	Owner	Group	Name	Registration time
	60	aguarise	INFN-TO	c5-etics-devel-v1	11:18:14 01/22/2013
	62	aguarise	INFN-TO	c6-devel-eclipse-vram	14:26:46 01/23/2013
	63	cernvm	users	CernVM-Slave	12:40:48 02/07/2013
	64	cernvm	users	CernVM-Master	12:41:18 02/07/2013
	66	oneadmin	oneadmin	SLC5-SSO	15:32:26 03/04/2013
		oneadmin	oneadmin	CernVM-SSO	
		oneadmin	oneadmin	WN-EMI2-CentOS6-V2	
		oneadmin	oneadmin	WN-EMI2-CentOS6-V2-Small	
		eadmin	oneadmin	OneMaster-3.8-V3	
	72	on	oneadmin	WN-EMI2-CentOS6-V2-postinstall	
	73	oneadmin	oneadmin	WN-EMI2-CentOS6-V2-Small-postinstall	10:55:59 04/09/2013
	75	oneadmin	on min	CE-EMI2-CentOS6-v5-postinstall	13:44:36 04/10/2013
	76	oneadmin	oneadmin	CE-EMI2-CentOS6-V5-install	15:12:39 04/15/2013
	79	oneadmin	oneadmin	WN-EMI2-CentOS6-V2-CVMFS	17:17:53 04/17/2013
	80	oneadmin	oneadmin	BDII-EMI2-CentOS6-V5-install	12:28:57 04/22/2013
	81	oneadmin	oneadmin	SE-EMI2-CentOS6-V5-install	11:32:06 05/03/2013

Tier-2 services and worker node templates

CERNVM-based templates

TIER 2 CPU EFFICIENCY



Jobs efficiency (cpu time / wall time)

	Series	Last value	Min	Avg	Max
1.	Legnaro	91.79	59.66	90.05	98.02
2.	Torino	93.66	28.53	87.88	98.05
	Total	**92.72**		**88.97**	

http://alimonitor.cern.ch

THE VIRTUAL ANALYSIS FACILITY

VAF components: overview

- User interacts with:
  - PoD to request and book workers
  - PROOF to execute jobs
- Under the hood:
  - worker requests are scheduled by HTCondor
  - CernVM virtual machines are part of the HTCondor cluster

PROOF

PoD

HTCondor

CernVM

↑

Services stack

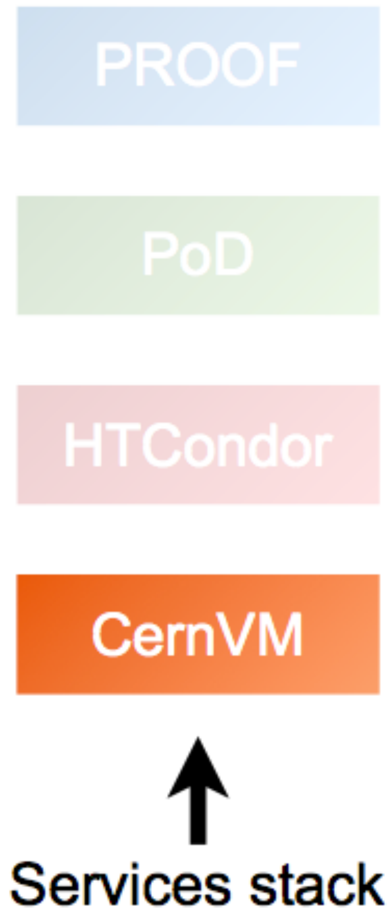
Dario.Berzano@cern.ch

http://goo.gl/CFnMM

THE VIRTUAL ANALYSIS FACILITY

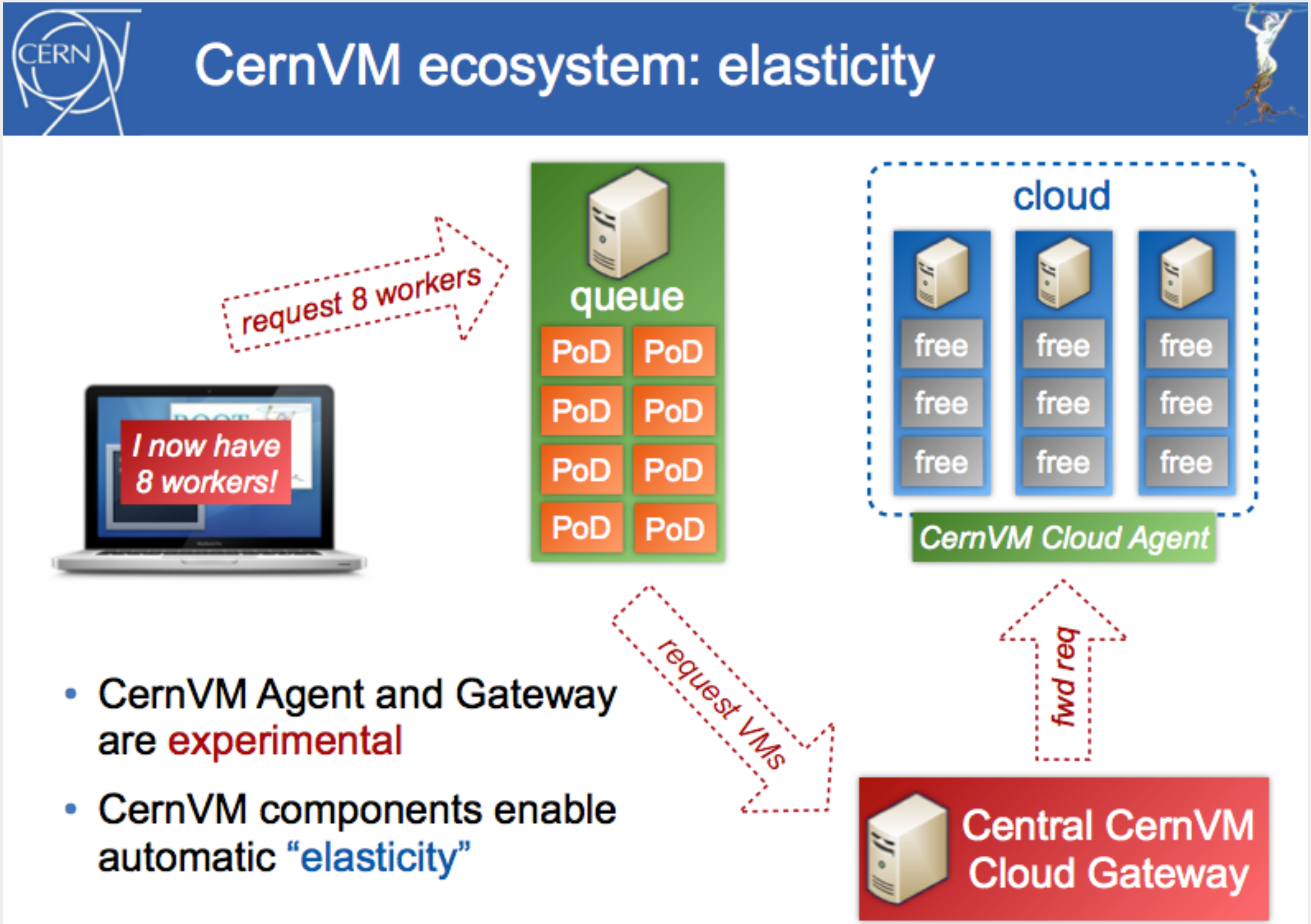
VAF components: CernVM

- CernVM is our reference platform:
  - uniform development environment
  - lightweightness: software downloaded on demand with cvmfs
  - online public context repository (sort of "marketplace")
- CernVM Cloud ecosystem: EXPERIMENTAL
  - Entire VAF cluster instantiated with one click using CernVM Gateway



Dario.Berzano@cern.ch
http://goo.gl/CFnMM

THE VIRTUAL ANALYSIS FACILITY



Dario.Berzano@cern.ch
http://goo.gl/CFnMM

FUTURE DEVELOPMENTS

- Understand the opportunities given by the CernVM "ecosystem"
- 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 users.
- Participate in upcoming projects aimed to develop a higher-level federated cloud infrastructure

CONCLUSIONS

- The infrastructure is in full production mode since more than one year
- The core software stack (OpenNebula + GlusterFS) proved itself stable and robust
- The management of the centre was actually simplified
  - Trivial example: rolling updates
- Lots of room for improvement and optimization
  - Example: there is no trivial method to optimize allocation of sets of identical machines on heterogeneous hypervisors (8, 12, 24 cores per host)
- Lots of room also for new features, extensions and integrations

- Questions?

Stefano.Bagnasco@to.infn.it