Workshop CCR INFN - May 25-29, 2015 - Frascati

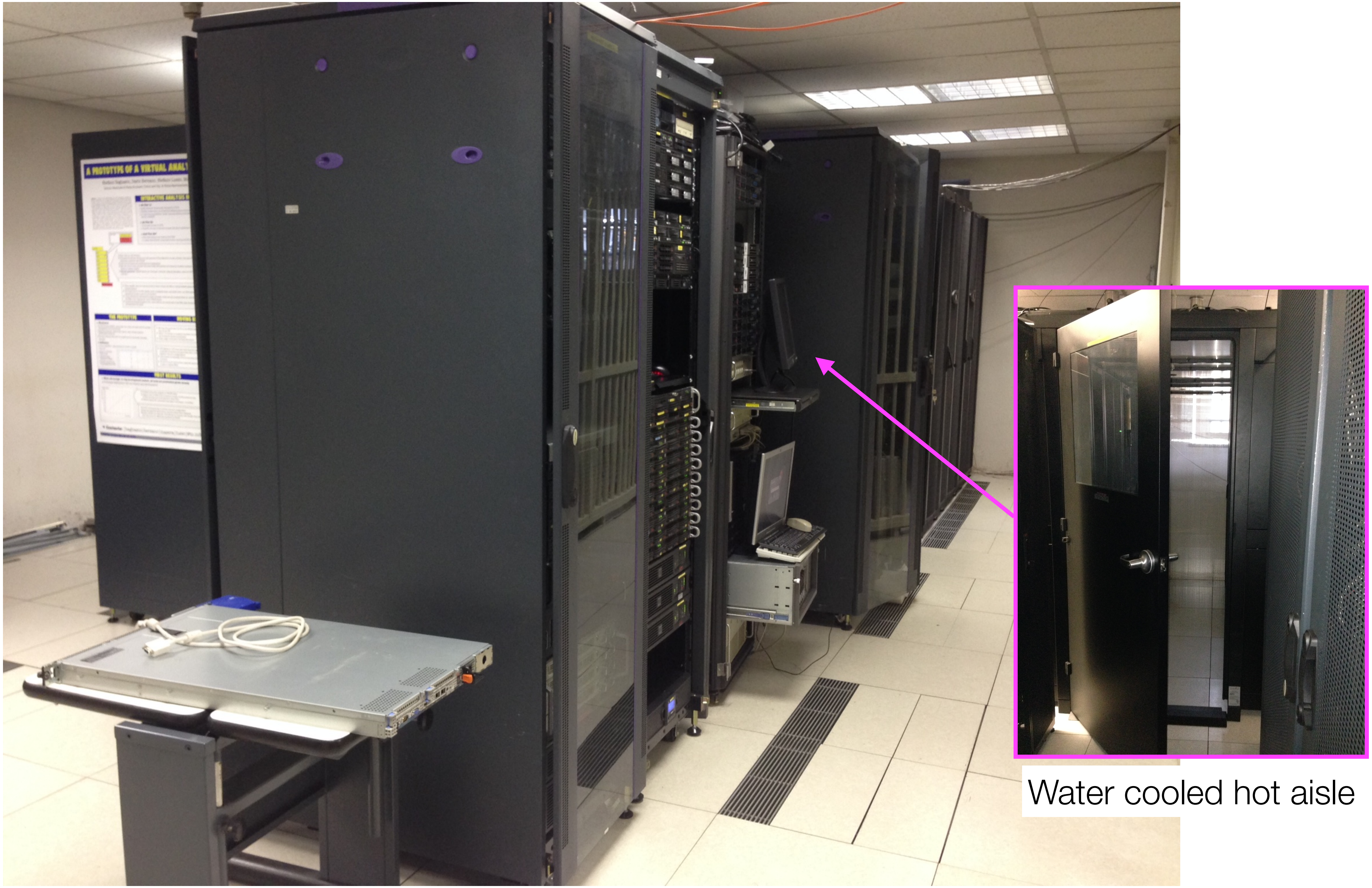
A multi-tenant Cloud at the Torino site

Speaker: Sara Vallero

on behalf of the INFN Torino Cloud Group

INFN
Istituto Nazionale
di Fisica Nucleare

The INFN Torino Computing Centre



Water cooled hot aisle

The INFN Torino Computing Centre

Servers

- **cloud controller**: HP DL360 (2011)
- 2 x disk server: HP DL360 G7 (2012)
- storage access servers for infrastructure:
  - 2 x HP DL360 G8 (2014)
  - HP DL360 G8 (2011)
- storage access servers for data:
  - XRootD
  - Storm

Hypervisors

- **73 hosts** (2011-2015):
  - AMD 6320 64 GB
  - AMD 6238 80 GB
  - AMD 6168 64 GB
  - Intel E5-2650v2 128 GB
  - Intel X5650 48 GB
- 8 to 24 cores per socket
- 1.9 to 2.8 GHz
- SATA from 500 GB 2.5” to 2TB 3.5”
- virtualization: **KVM**

OpenNebula
(version 4.8)

1.3k cores

1.8 PB (gross)

1-10 Gbps LAN
10 Gbps WAN

CentOS 6

Storage (for services and experiments)

- **1.8 PB**
- disk controllers for infrastructure:
  - HP P2000 G3 (2011)
  - Sun StorageTek 6140 (2007)
- disk controllers for data

Cloud storage

- 10 TB
- **iSCSI server**: HP DL360 G5 (2007)
- **iSCSI NAS**: Qnap (2010)

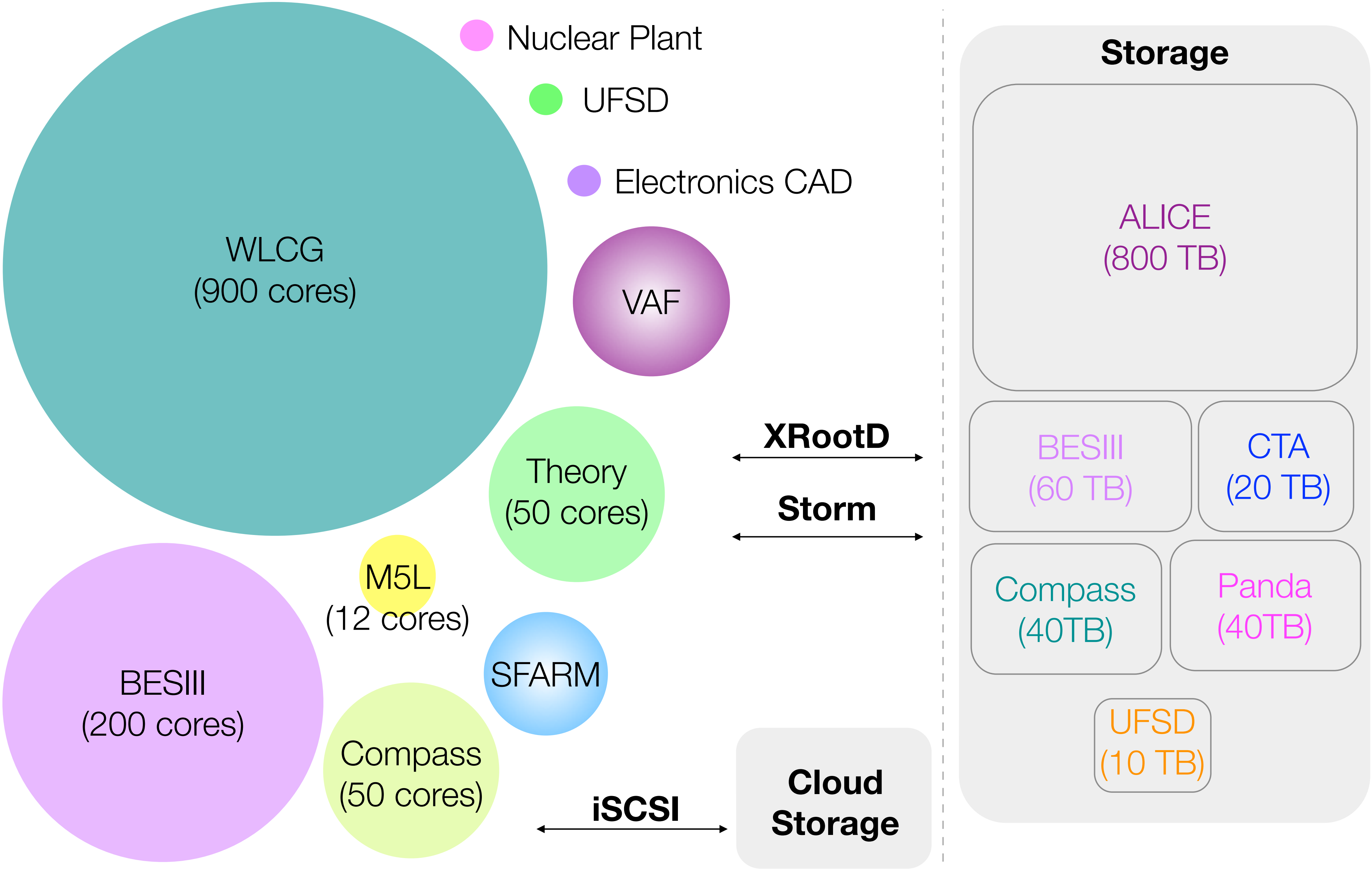
Networking

- **hypervisors**: 1Gbps
- **storage** servers: 10 Gbps
- storage controllers: 8/16 Gbps Fibre Channel
- 1 modular switch with 1/10 Gbps ports

Story of a private Cloud

- need to share resources between analysis facility and GRID site
- first virtualisation-based prototype presented at ACAT (2008)
- in 2011 switch to private IaaS Cloud paradigm:
  - ease site management (less manpower)
  - more flexibility
- basic OS images with complex contextualisation
- GRID worker-nodes and services as virtual machines
  - two classes of hypervisors (services and workers)
  - shared filesystem for live migration of services
- first version of the ALICE Virtual Analysis Facility (VAF)
- in 2013 current version of the *elastic* VAF
- since then more and more use-cases:
  - virtual batch farms
  - single instances
- in 2013-2014 BESIII Tier2

Multi-tenancy



● Nuclear Plant

● UFSD

● Electronics CAD

WLCG
(900 cores)

VAF

Theory
(50 cores)

M5L
(12 cores)

SFARM

BESIII
(200 cores)

Compass
(50 cores)

XRootD

Storm

iSCSI

Cloud Storage

Storage

ALICE
(800 TB)

BESIII
(60 TB)

CTA
(20 TB)

Compass
(40TB)

Panda
(40TB)

UFSD
(10 TB)

The ALICE case

Grid site

- other VOs are supported on the Tier-2 site besides LHC (CTA, BELLEII…)
- whole site on the Cloud:
  - first worker-nodes…
  - … then all services little by little

VAF

- elasticity
- Proof and PoD
- HTCondor
- micro-cernvm

SuperMario Farm

- replacing local batch farm
- same ingredients as VAF

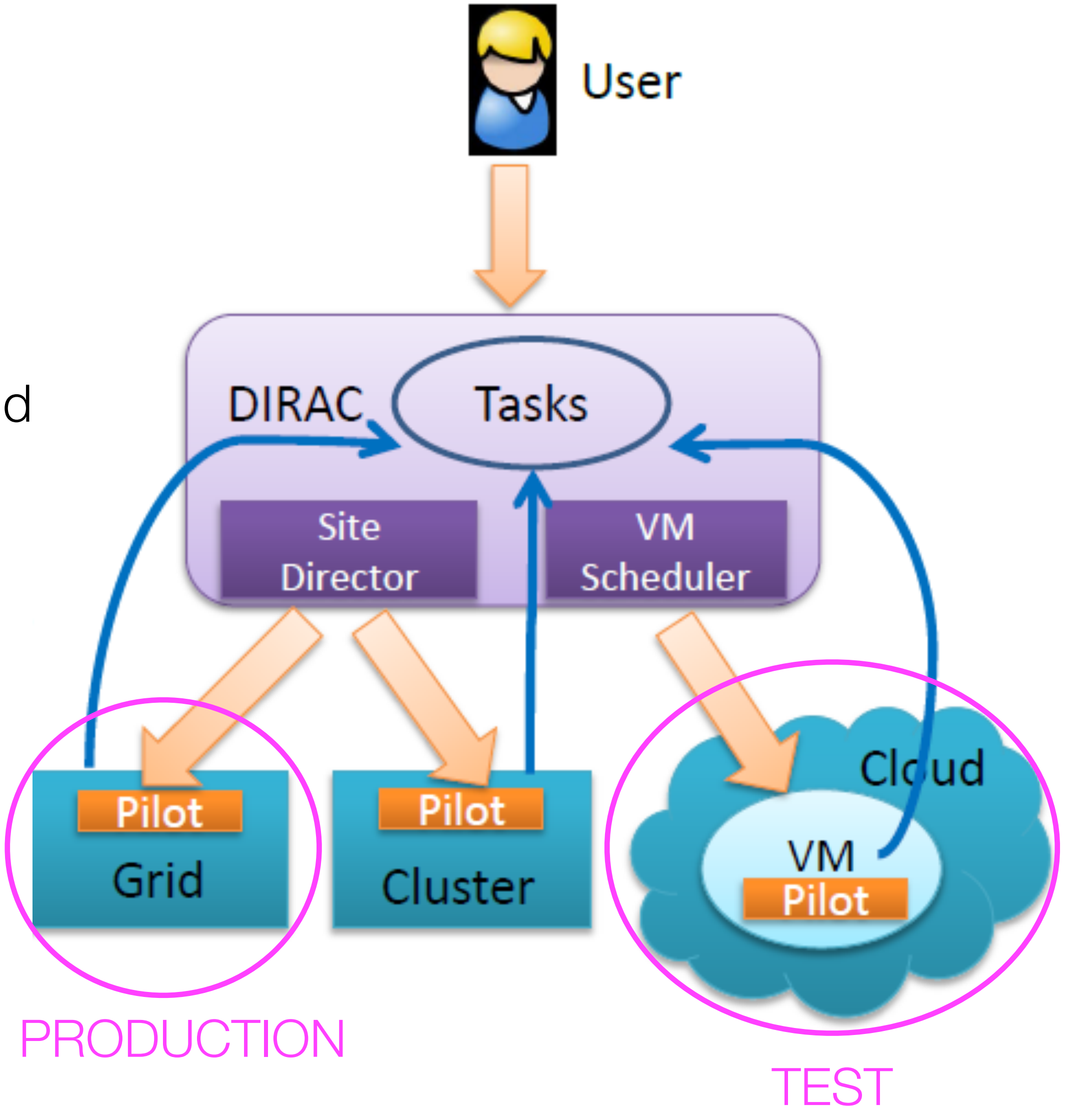
The BESIII case

How to integrate

- job scheduling scheme remains unchanged
- instead of Site Director for cluster and GRID
  → VM scheduler introduced to support Cloud

Workflow

- start new VM with 1 CPU core when there are waiting jobs
- 1 job scheduled on 1 VM at the same time
- delete the VM after no more jobs for a certain period of time



The BESIII case

Production activities

- CREAM Grid site (including services) running on the infrastructure
- 200 cores (~30 VMs) → working on elasticity
- stable running since end 2013

R&D activities

- separate test infrastructure
- direct collaboration with IHEP Computer Centre
- trying to consolidate collaboration
  (local BES group, Torino & IHEP Computing Centres)

The PANDA case

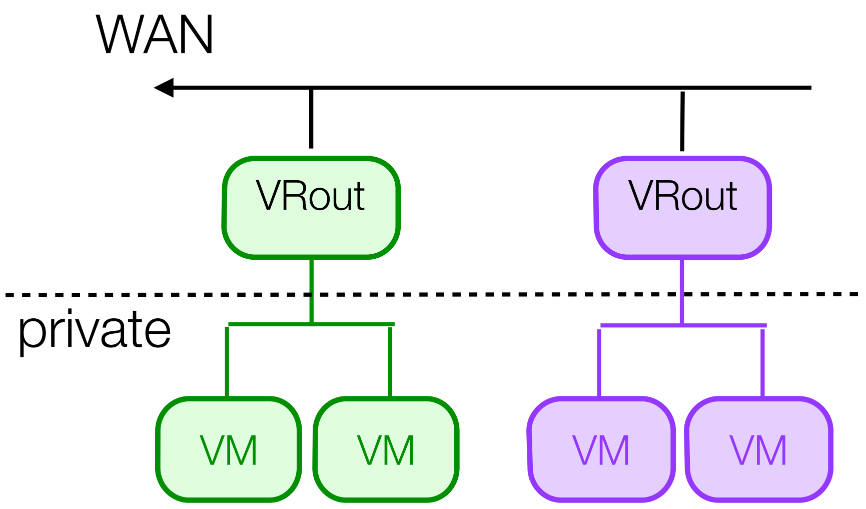
Disclaimer: in the following we pretend PANDA is not an endangered species…

- mostly uses the GRID site
  - LCG VO-box
  - external SE (for the time being)
- jobs running on ALICE nodes
- many tools shared with ALICE (AliEn2)
- running experiment-wide services
  - Monalisa repository
  - database replica

Virtual batch farm provisioning model

Network isolation (level 2):

- each user is assigned a Virtual Network
- each network is isolated with ebtables rules on the hypervisor bridge (OpenNebula V-net driver)



Connectivity remains under sysadmin control, user has no access to the VRouter!

Virtual Routers (level 3):

- private and public IP
- light-weight OpenWRT VM (1CPU, 150 MB)
- DHCP, DNS, NAT functionalities
- Firewalling / port-forwarding
- configuration possible via HTTPS or SSH

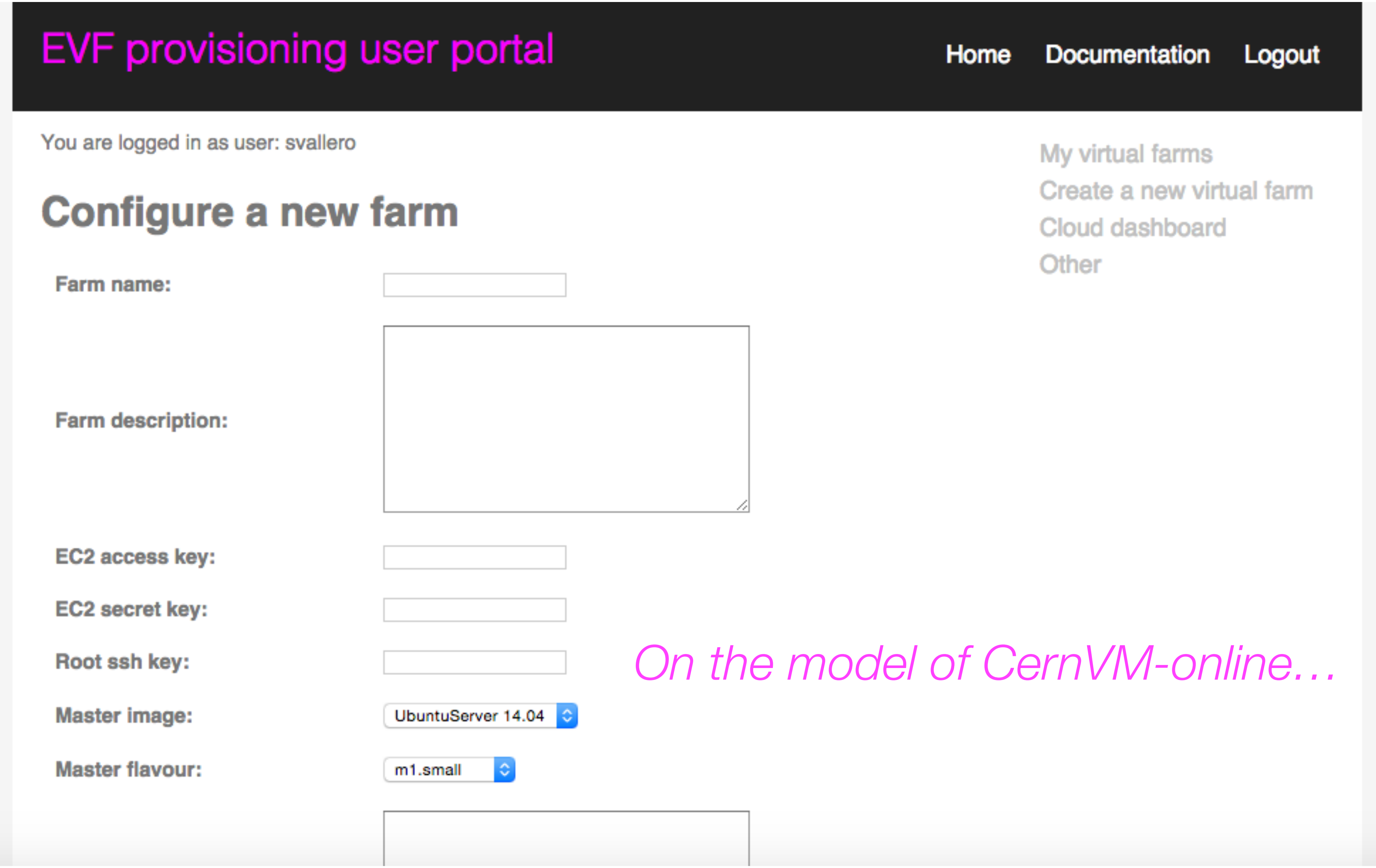
Elastic IPs

- bind dynamically a public IP to one of the private VM instances

Provisioning:

- configuration simplified through the definition of Amazon-like flavours
- VM instantiation via EC2 interface (euca-tools)

Self-service virtual farms (work in progress)



EVF provisioning user portal

You are logged in as user: svallero

Configure a new farm

Farm name:

Farm description:

EC2 access key:

EC2 secret key:

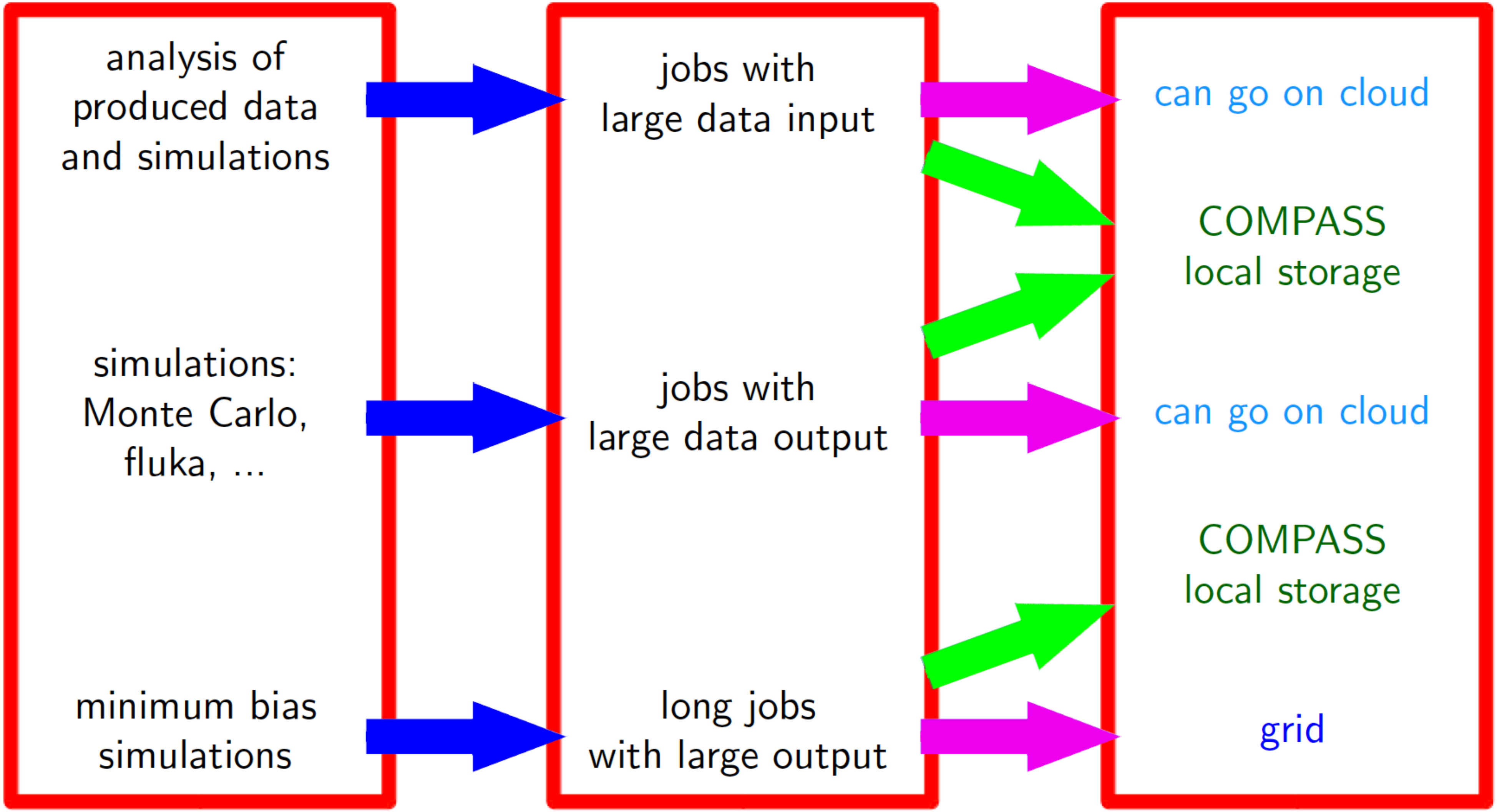
Root ssh key:

Master image: UbuntuServer 14.04

Master flavour: m1.small

On the model of CernVM-online…

The Compass case



analysis of produced data and simulations

simulations: Monte Carlo, fluka, ...

minimum bias simulations

jobs with large data input

jobs with large data output

long jobs with large output

can go on cloud

COMPASS local storage

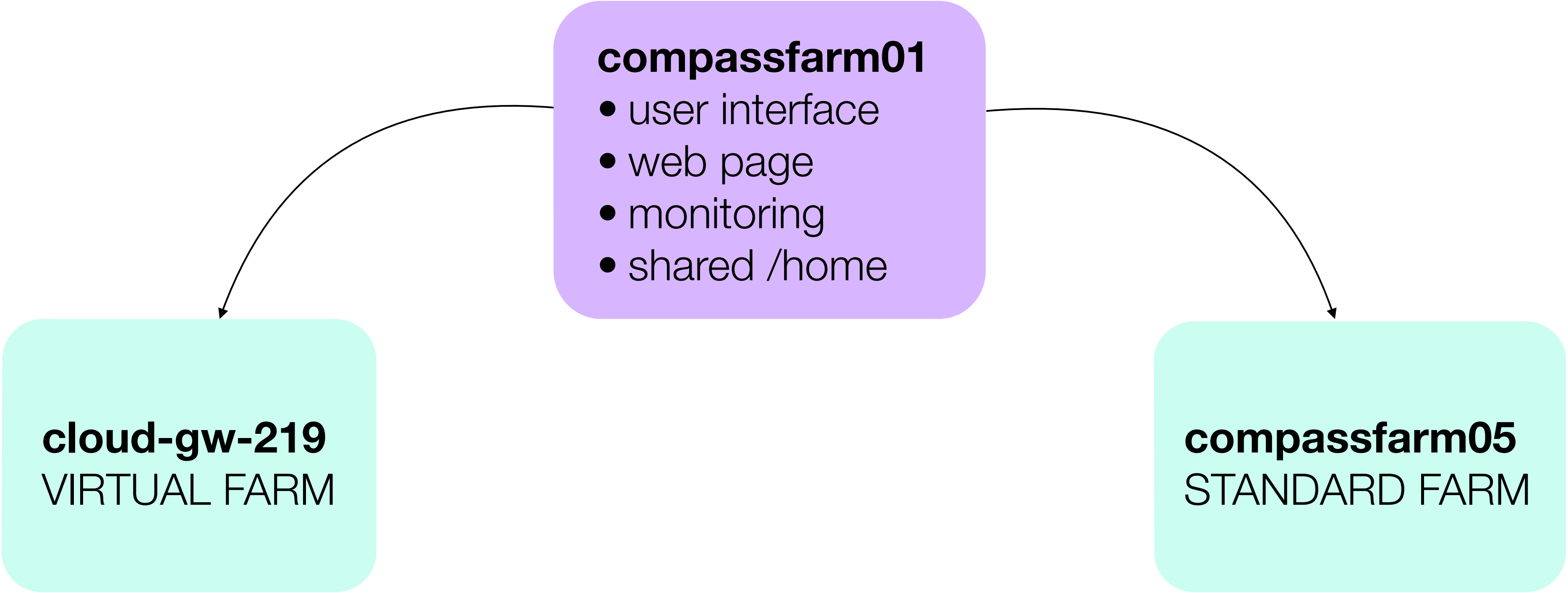
can go on cloud

COMPASS local storage

grid

The Compass case

- use-case: simple batch farm
- proprietary resources in phase-out
- new resources added to Cloud infrastructure
- slow progresses because of lack of manpower



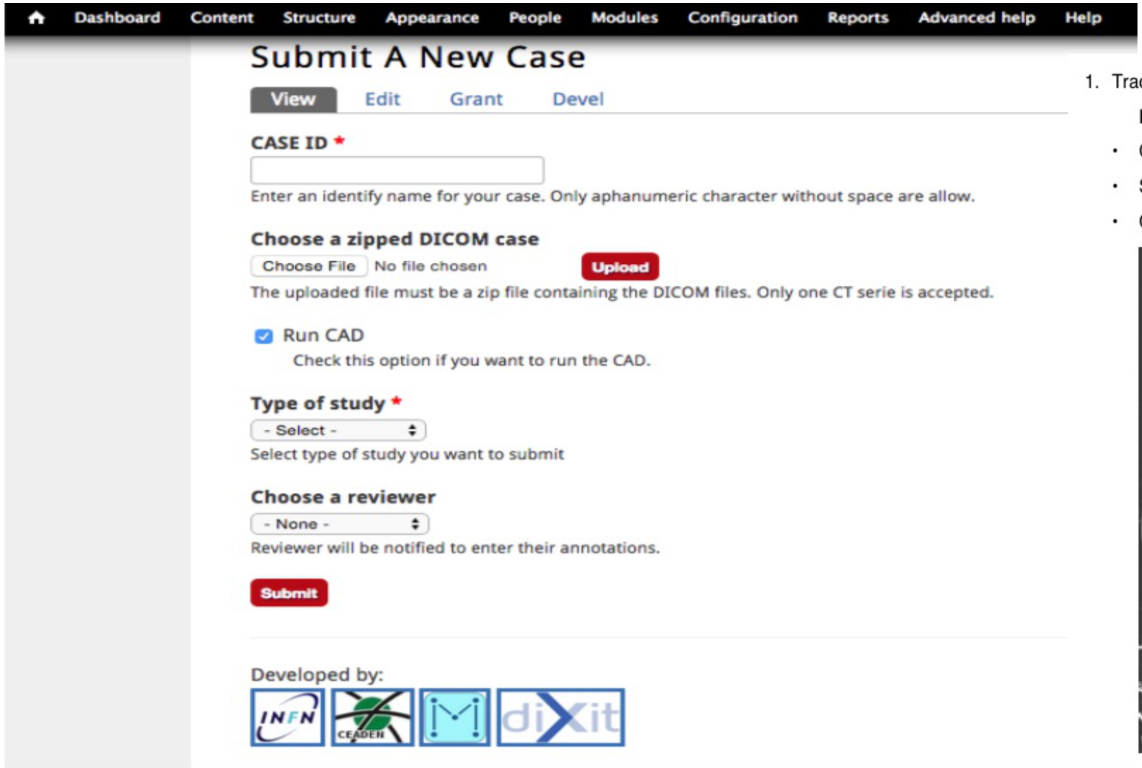
compassfarm01
- user interface
- web page
- monitoring
- shared /home

cloud-gw-219
VIRTUAL FARM

compassfarm05
STANDARD FARM

The M5L case

- use-case: on-line service for Computer Aided Detection for automatic analysis of lung CT
- physicians can submit medical exams for CAD processing through a web interface
- 2 combined CADs do the analysis
- physician and possible reviewers ara informed when processing is over
- analysis results made available through web interface in several formats (pdf, html, xml)



Submit A New Case

CASE ID *

Enter an identify name for your case. Only aphanumeric character without space are allow.

Choose a zipped DICOM case

No file chosen

The uploaded file must be a zip file containing the DICOM files. Only one CT serie is accepted.

☑ Run CAD

Check this option if you want to run the CAD.

Type of study *

- Select -

Select type of study you want to submit

Choose a reviewer

- None -

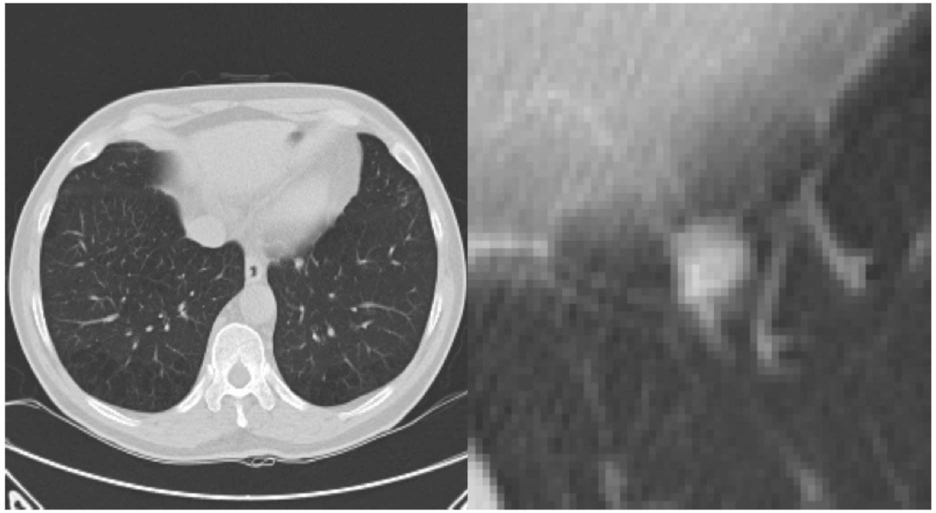
Reviewer will be notified to enter their annotations.

Developed by:

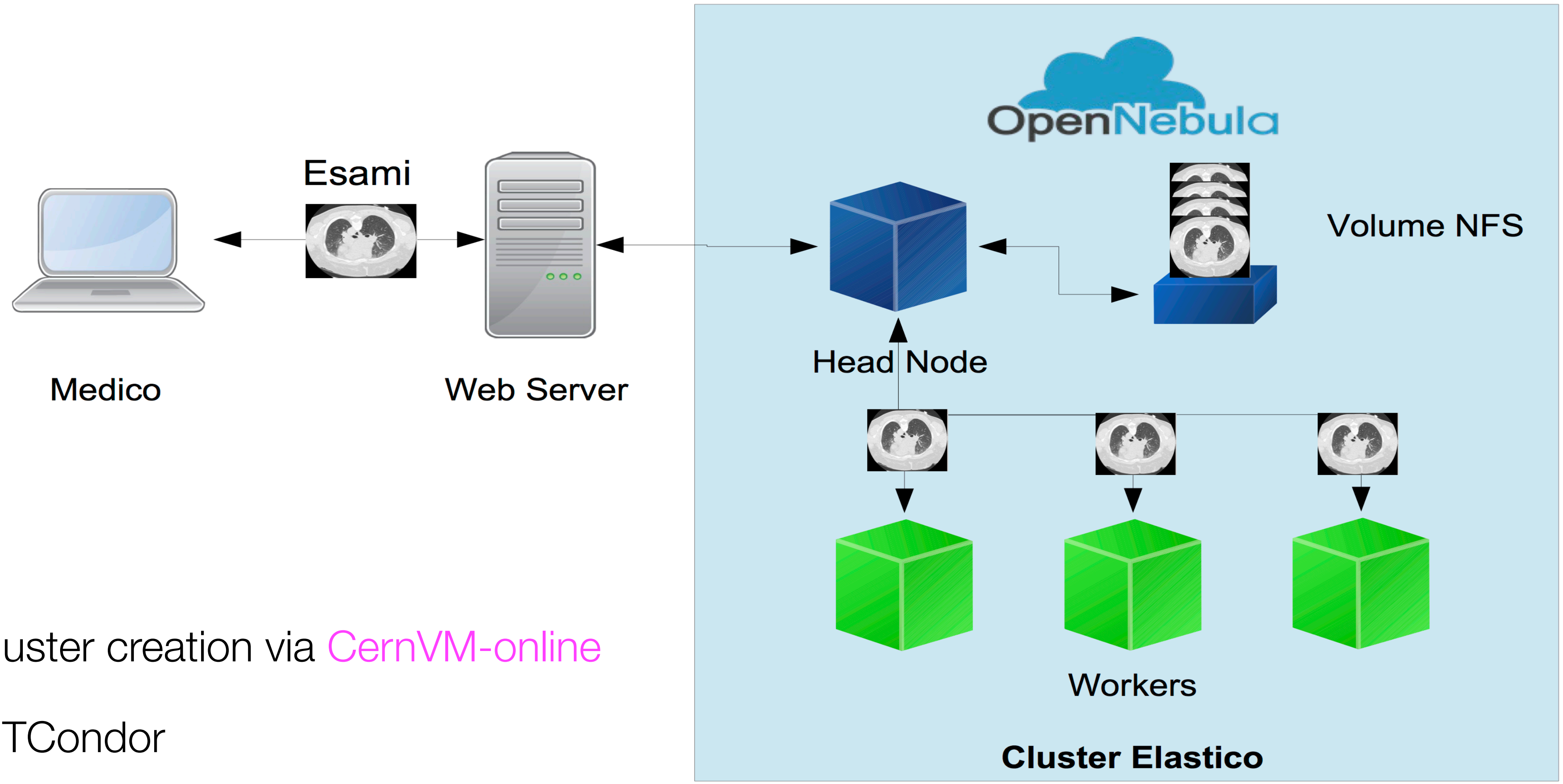
1. Tracking Identifier = "Nodule # 1"

   Properties:

   - Center: (323,261) Radius=15
   - Slice: 18, Z-Position: -267.2
   - Certainty of Feature: 95%



The M5L case



- cluster creation via CernVM-online
- HTCondor
- elasticity
- typical example of virtual-farm on-demand following the VAF model

Other use-cases

BELLEII and CTA

- both VOs use the GRID Tier2 infrastructure
- no direct contact with software developers
- about 200/300 jobs at peak time

GiuntiFarm (Theory group)

- static virtual farm
- completely self-managed
- we assume they're happy

On-demand Virtual Workstations

- UFSD, Nuclear plant simulation, electronics CAD…
- Single machine, usually large (16 core)
- Plus iSCSI disk for persistent storage
- Need a lot of support to set up

More coming

- JLAB12, Auger…

What have we learned?

- very difficult to convince users to release unused resources → force elasticity
- but some applications are non-elastic by nature…
- interaction with users/experiments is fundamental
- different flavours of users:
  - ironically small use-cases are more demanding in terms of support
  - they require very specific solutions (usually 1 VM)
  - little motivation to solve the problem themselves
  - think about them when developing PaaS (we also gain something in the game…)
- compared to GRID allows users to use a computing model they already know
- more skilled users can take full advantage of the Cloud (i.e. complex services, storage…)
- new provisioning model:
  - resources are assigned in terms of quotas and billed *a posteriori*

What have we learned?

- OpenNebula was chosen because at the time of the decision OpenStack (and CloudStack, Eucalyptus…) were not mature enough for a production-grade deployment
- we find that for our use case OpenNebula may be better suited than OpenStack, since it is simpler and more economical to manage
- we will bring the OpenNebula know-how into INDIGO to preserve the freedom to choose what fits best a given use case

The bill

- Cloud does not *necessarily* mean saving money
- resources outsourcing (few specialised sites) → economies of scale
- other economies:
  - small scale provisioning
  - manpower
  - but they do not come for free since day 1…

Manpower

- SL: physical infrastructure and GRID site management
- SV: Cloud middleware management, application support, R&D
- SB: application support and general worrying
- critical task: application support
- cannot run a production infrastructure without some R&D activity

The bill

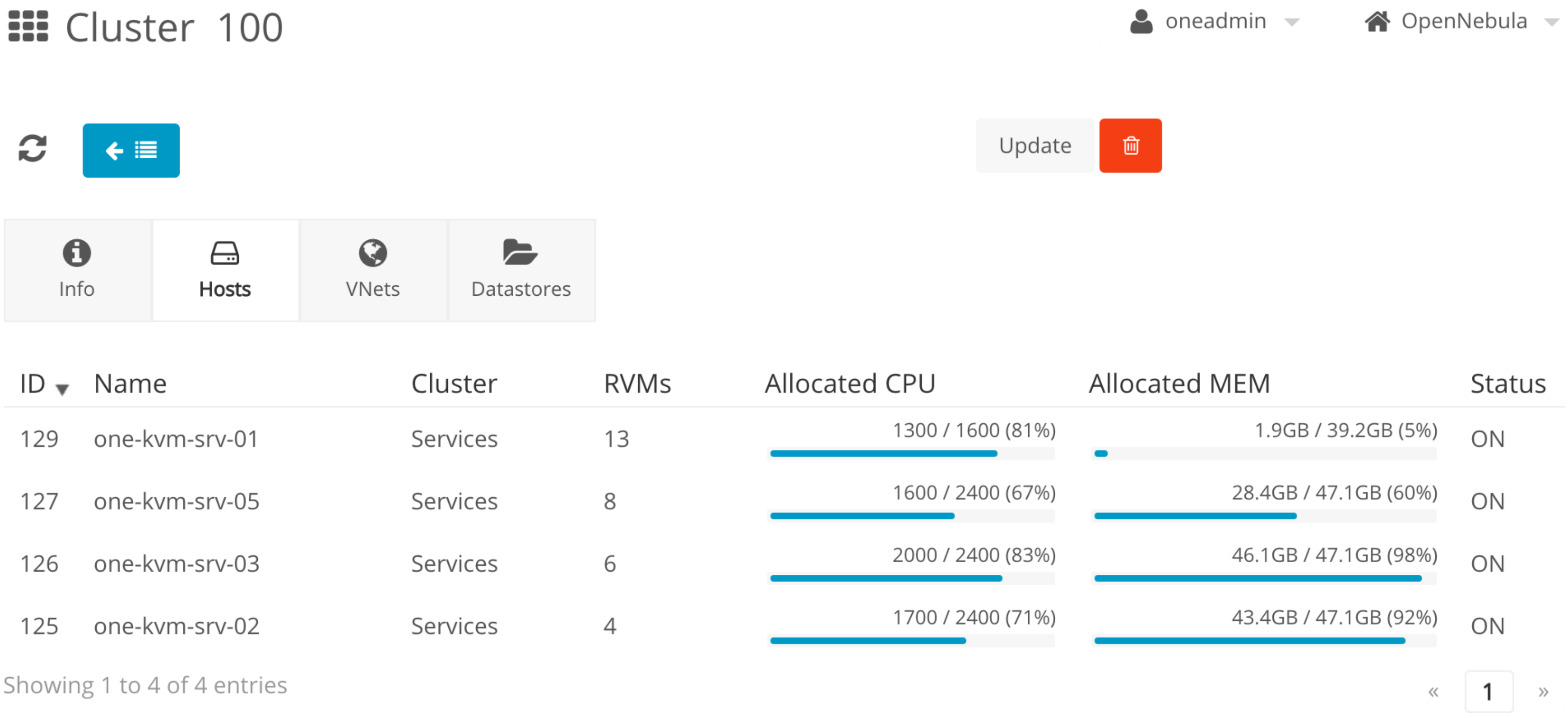
Money

- A very good infrastructure:
  - 2 HA servers for management
  - redundant high-performance storage for backend
  - iSCSI storage for persistent disk provisioning (not included)
  - Total: o(40kEUR) every 5 years
  - but pessimistic cost estimates from realistic MEPA prices, including VAT (at least 50% economies possible)
- What we have now:
  - cloud controller (2011)
  - 2 backend storage servers (2012)
  - 2 different iSCSI servers (2007, 2010)
  - disk controllers (2011, 2007)
- A production-grade infrastructure needs planned certain funding!
  - R&D can often be done on recycled hardware
  - computing power and storage (e.g. iSCSI) are funded by users
  - Who pays for the infrastructure?

The bill

Some immediate savings

- IaaS eases the use of virtualization technologies
- virtualization means consolidation:
  - we run 15 services (including 2 production Grid CEs, VOBox, BDII etc.) and 16 vRouters on 4 servers
  - no visible performance issue



ID	Name	Cluster	RVMs	Allocated CPU	Allocated MEM	Status
129	one-kvm-srv-01	Services	13	1300 / 1600 (81%)	1.9GB / 39.2GB (5%)	ON
127	one-kvm-srv-05	Services	8	1600 / 2400 (67%)	28.4GB / 47.1GB (60%)	ON
126	one-kvm-srv-03	Services	6	2000 / 2400 (83%)	46.1GB / 47.1GB (98%)	ON
125	one-kvm-srv-02	Services	4	1700 / 2400 (71%)	43.4GB / 47.1GB (92%)	ON

Showing 1 to 4 of 4 entries

« 1 »

Ongoing and planned activities

Infrastructure 2.0

- redesign a more resilient infrastructure, easier to manage
- more robust iSCSI persistent storage service
- clean-up of the network topology
- get rid of the last custom patches to ON (difficult to support)
- migrate to more mainstream Virtual Network management (i.e. VRouter → OpenVSwitch)
- more robust back-end database set-up

Finalise the *Virtual Farm Toolkit*

- web interface
- documentation, training sessions, share user experience
- contextualisation templates

Monitoring, Accounting & Billing

- test version of accounting/billing service based on the ElasticSearch ecosystem
- towards monitoring-as-a-service to complement the *Virtual Farm Toolkit*
- design and implement comprehensive modular monitoring system for infrastructure and applications

Outlook

- bring ONe experience into INDIGO
  - strong interest in automatic elasticity
- besides our specific HTC use-case, other ways could be explored…
  - GPUs
  - HPC
  - …
- Cloud computing for scientific applications is not yet mature (the Cloud was not originally conceived for that)

R&D IS FUNDAMENTAL

C3S

Centro di Competenza sul Calcolo Scientifico

- 900 kEUR funding from *Compagnia di S. Paolo* to UNITO to build a multi-purpose interdepartmental HPC Cluster
  - a tool for production-type computing
  - a platform for R&D activities in Scientific Computing
  - a forum where scientific computing know-how can coalesce and grow
- INFN is partner in the project and will host the cluster
  - the cluster will be a separate entity, but…
  - …the system (or a part thereof) will be managed as an IaaS infrastructure very similar to the existing one…
  - … hopefully fostering lots of synergies