



Esperienze su infrastrutture Cloud:

Trieste, Torino e Area Padovana

Speaker: Sara Vallero

Trieste

Lucio Strizzolo, Claudio Strizzolo, Tullio Macorini (Ceph)

Area Padovana

Paolo Andreetto, Fulvia Costa, Alberto Crescente, Alvise Dorigo, Sergio Fantinel, Federica Fanzago, Massimo Sgaravatto, Sergio Traldi, Marco Verlato, Lisa Zangrando

Torino

Stefano Lusso, Stefano Bagnasco, Sara Vallerio

- Giugno 2013: inizio attività e collaborazione con il gruppo di lavoro di CCR sul Cloud Computing.
- Piattaforma Cloud scelta su indicazione gruppo CCR sul Cloud Computing: **OpenStack**.
- Versione utilizzata attualmente: **Liberty**.
- Distribuzione: RDO su sistema operativo **CentOS7** (fino a dicembre 2014 Scientific Linux 6).
- *RDO is a community of people using and deploying OpenStack on CentOS, Fedora and Red Hat Enterprise Linux ([https:// www.rdoproject.org/](https://www.rdoproject.org/)).*

Cosa abbiamo sperimentato?

- Installazione e configurazione di tutte le versioni di OpenStack da Grizzly a Liberty (eccetto Kilo).
- Sperimentazione di diverse soluzioni per lo storage condiviso: **GlusterFS** e **Ceph**.
- Configurazione completa del networking in modalità Linux bridge. A maggio 2015 migrazione a **Open vSwitch** e utilizzo dei **Distributed Virtual Router** (DVR).

Distributed Virtual Router (DVR)

- L'utilizzo dei DVR permette ai compute node di gestire il traffico da e verso le reti pubbliche per le istanze che presentano un floating IP (indirizzo sulla rete pubblica). In questo modo **si evita di far passare il traffico attraverso i network node** che in caso contrario costituirebbero dei punti critici e potrebbero diventare dei colli di bottiglia.
- L'utilizzo dei DVR non influenza il flusso del traffico delle istanze senza floating IP che rimane gestito dagli *L3 agent* sui network node.
- I DVR permettono inoltre di scaricare i network node dal traffico *east-west* fra la istanze.

Problemi e considerazioni

- Il passaggio da Linux bridge (LB) a Open vSwitch (OVS) è stato affrontato per poter testare i DVR. **OVS è più delicato** rispetto a LB, ed abbiamo riscontrato alcuni problemi (risolti) in fase di test e configurazione allo start-up dei nodi.
- Liberty configurato per l'utilizzo dei **DVR** presenta un **problema in fase di live-migration** delle macchine virtuali: è possibile effettuare la migrazione ma eventuali comunicazioni aperte (SSH) rimangono appese (<https://bugs.launchpad.net/neutron/+bug/1456073>)

Chi ha usato la Cloud?

- Già durante la fase di sperimentazione:
 - Esperimento **Fermi/Glast**: tools del Burst Advocate per analisi rapida GammaRayBurst.
 - Esperimento **Alice**: analysis facility interattive basate su Proof e sulle tecnologie di virtualizzazione e Cloud Computing.
 - Esperimento **Compass**.
 - **Servizio Calcolo e Reti**: sperimentazione e test dei servizi.
- La **struttura di produzione** è aperta a tutta la Sezione di Trieste da **Marzo 2016**.
- Sono disponibili immagini con sistemi operativi diversi: Scientific Linux, CentOS, Debian, Ubuntu, μ CernVM-prod.

- Nodi di infrastruttura:
 - 1 cloud controller
 - 2 network nodes
 - 1 database server
 - 1 LDAP server per l'autenticazione
 - 4 compute nodes, espandibili a seconda delle richieste
- Switch con 2 porte a 10 Gb (1 verso server Ceph, 1 verso rete locale/farm di sezione)

- Lo storage della cloud è basato su **Ceph**
- Infrastruttura Ceph:
 - 8 server virtuali su un unico hypervisor VMware ESXi 6: 3 Monitor, 3 OSD, 1 Metadata, 1 Servizi & export iSCSI
 - storage disponibile: 6 dischi da 4 TB in modalità replica 2
→ poco più di 11 TB effettivi
 - collegamento 10 Gb verso lo switch

- **Networking:** configurazione *Per-tenant Routers with Private Networks* con ML2, Open vSwitch e VXLAN. Utilizziamo i DVR (Distributed Virtual Router).
- Le comunicazioni di rete verso l'esterno per le macchine virtuali senza indirizzo pubblico (NAT) utilizzano 2 network nodes in modo da ripartire il traffico dei diversi progetti (tenant) ospitati nella Cloud.
- Gestione degli utenti: viene utilizzato un **server OpenLDAP** dedicato alla Cloud. L'autenticazione avviene tramite il **Kerberos** nazionale INFN.

- Integrazione nella Cloud dei worker nodes della farm di calcolo scientifico della Sezione di Trieste, con accesso allo storage GPFS, integrazione con il cluster LSF, installazione del software GRID
- Storage: completata la sperimentazione con GPFS, stiamo iniziando una **sperimentazione con EOS**

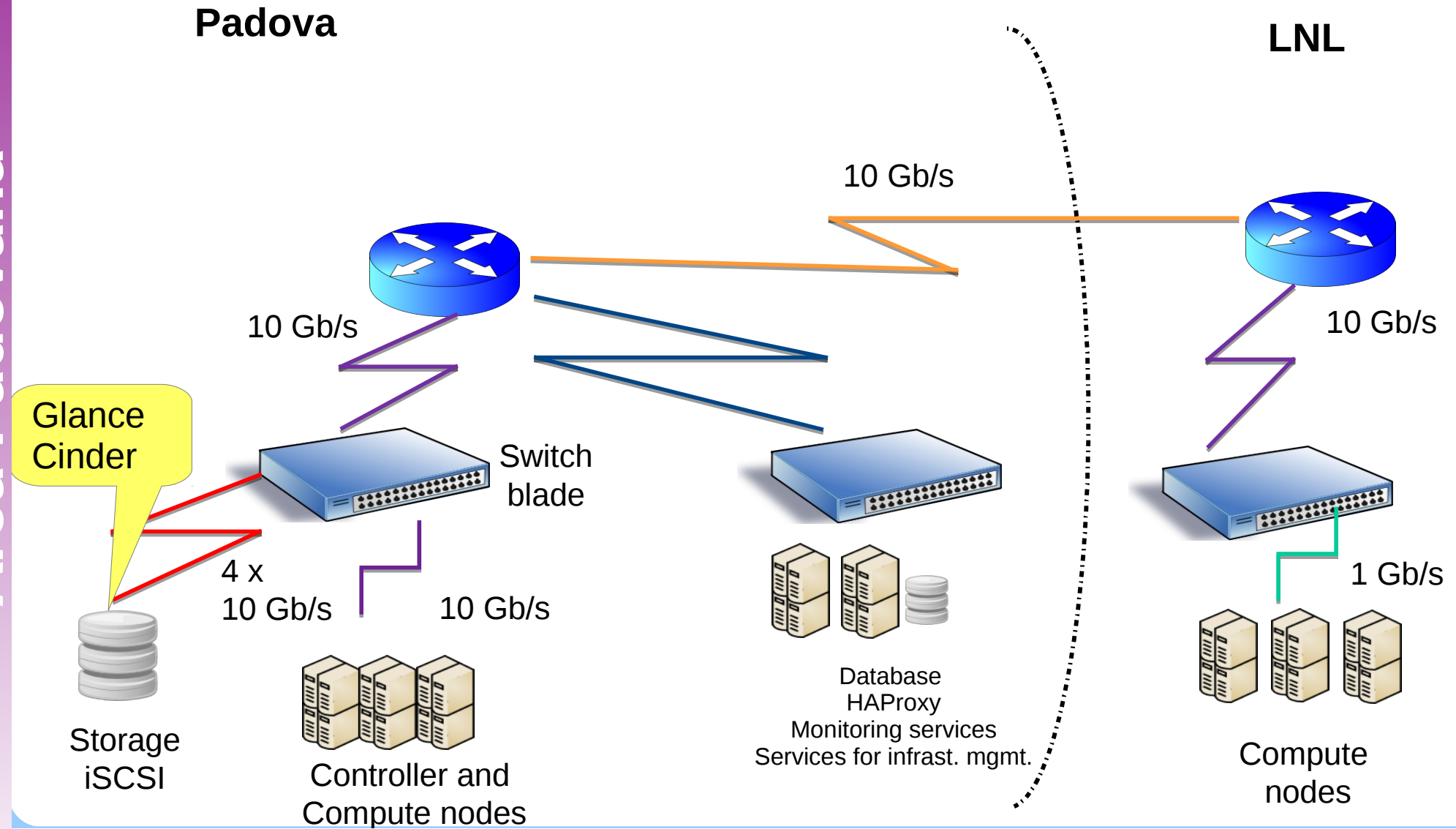
La Cloud dell'Area Padovana



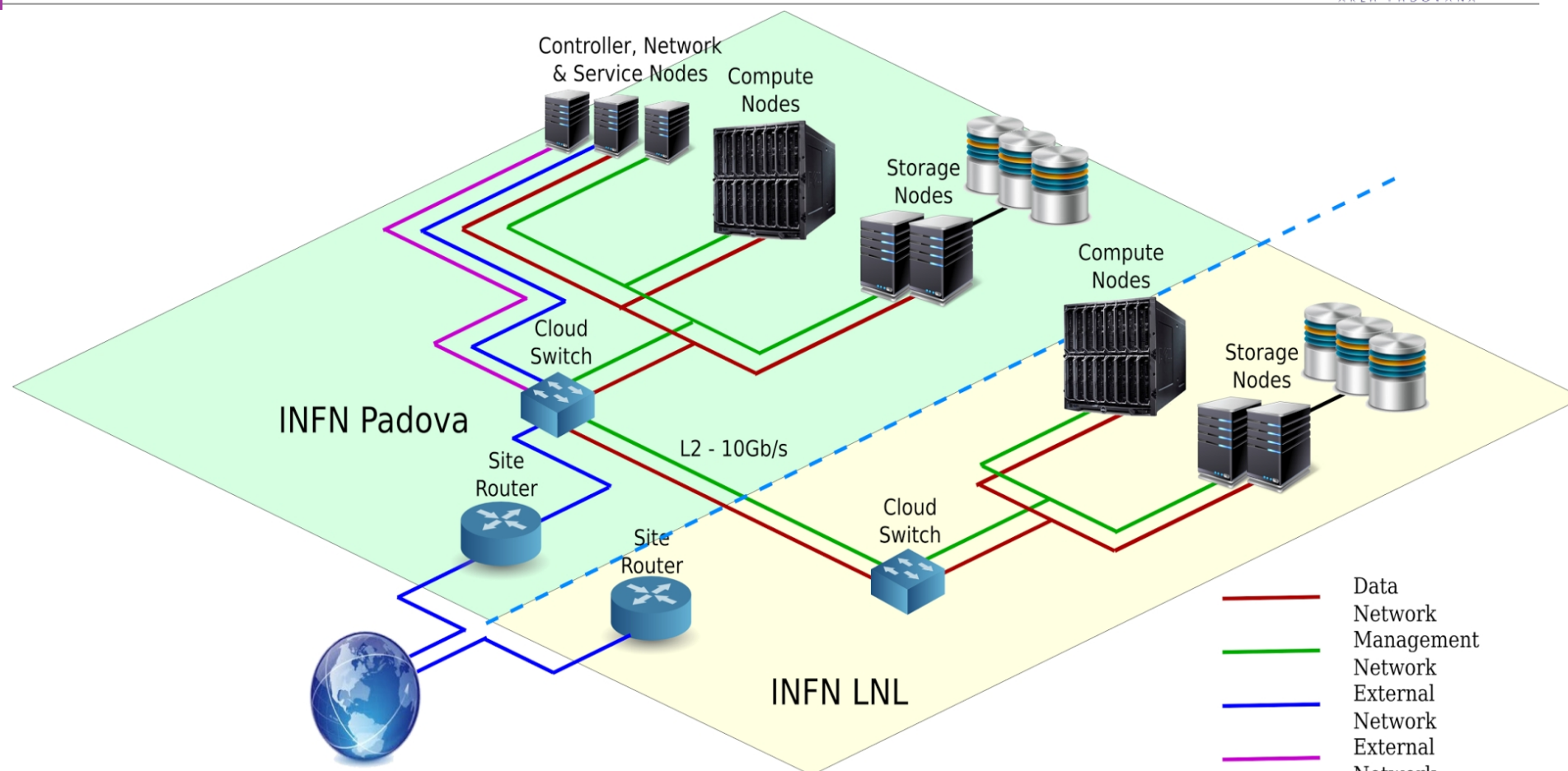
- Singola IaaS Cloud con **risorse distribuite tra INFN-Padova e INFN-LNL**
- Ora in uso la versione Kilo di **OpenStack**
- 1 OpenStack update per anno, i.e. saltando una release
- Servizio di Produzione
- 70+ utenti registrati, 20+ progetti (v. prossime slide per alcuni casi d'uso)
- Integrati alcuni **in-house developments**, e.g. per autenticazione e user registration, come mostrato nelle prossime slide

	Compute Nodes	Cores (in HT)	Storage (TB)
Padova	15	656	43 (Cinder, Glance)
LNL	13	416	
Total	28	1072	

Layout

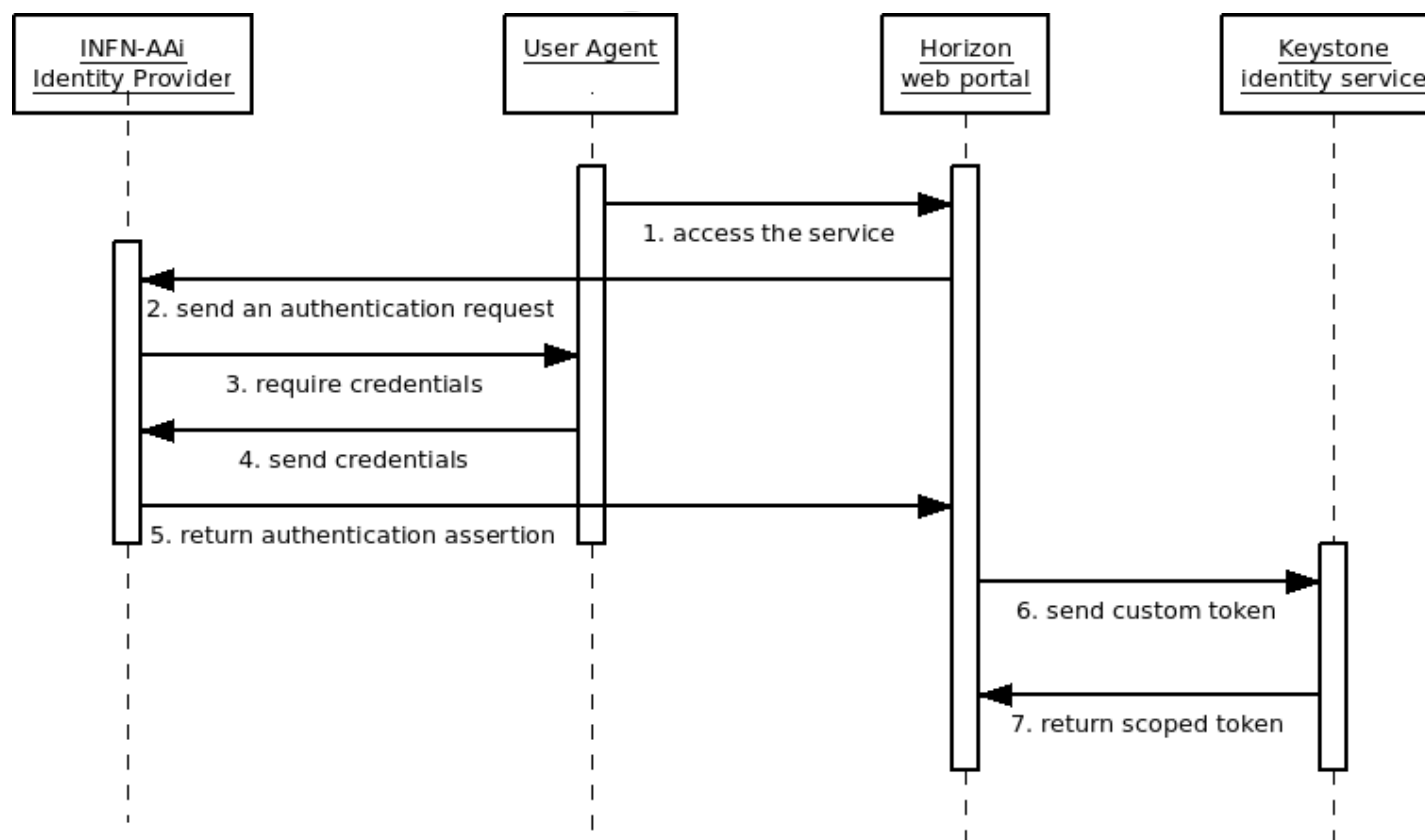


Network Layout



- **Neutron with Open vSwitch/GRE** configuration
- **Two virtual routers** with **external gateways** on public and LAN networks
- Possibility to access VMs from Padova and LNL lans even if they don't have a floating IP
- **GRE tunnels among Compute nodes and Storage** servers to allow high performance storage access (via e.g. NFS) from VMs
- Integration with Tier-2 networking (for ALICE and CMS users)

Identity and access management 1/2



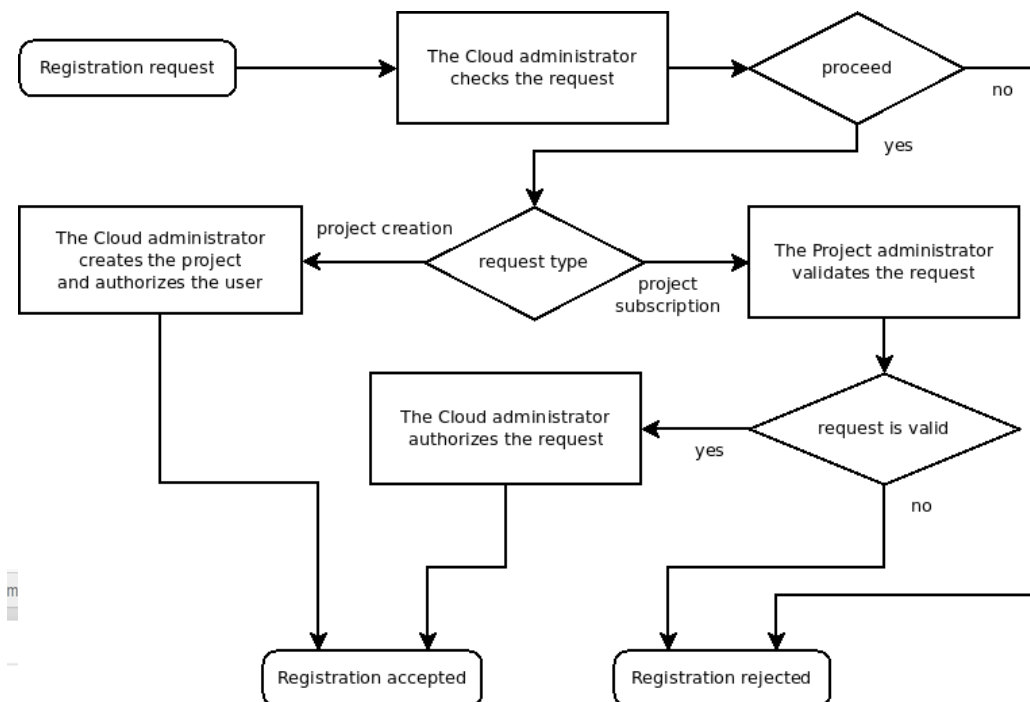
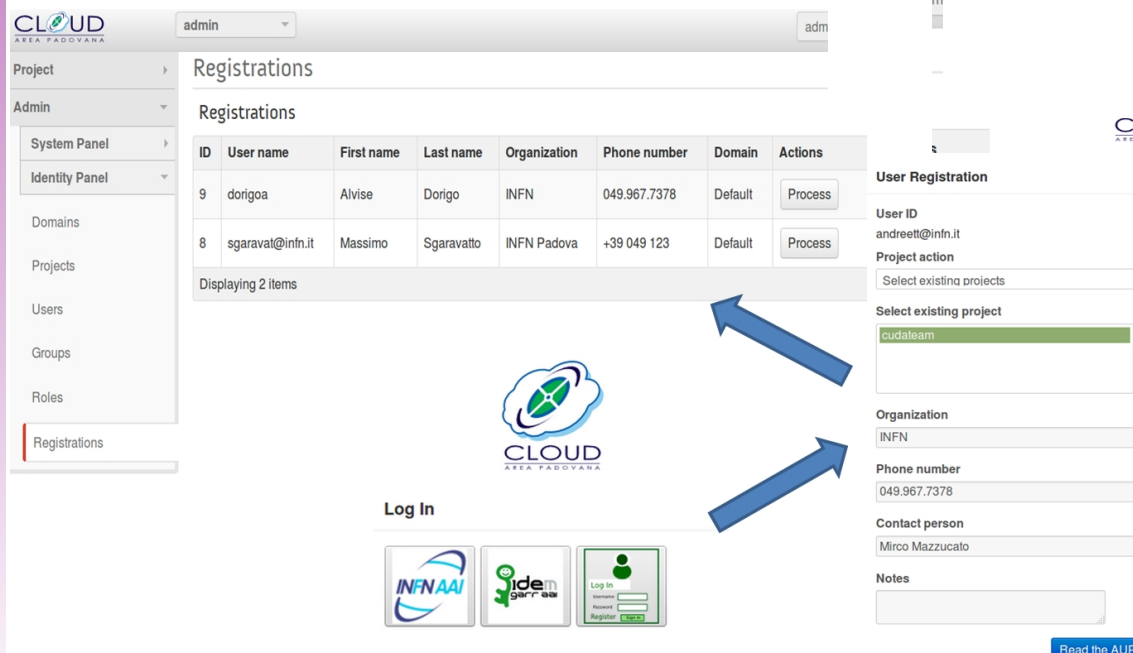
OpenStack Keystone Identity service e Horizon Dashboard estesi:

- Per **interagire con l' INFN-AAI Identity Provider**
- Facendo riferimento al Security Assertion Markup Language (SAML) OASIS standard (Web Browser SSO profile)
- Netta **separazione** tra fase di **autenticazione e autorizzazione**

Identity and access management 2/2

- Workflow per gestire le **richieste di registrazione** che coinvolge due entità con diversi ruoli e capability:

- ✓ Il cloud administrator
- ✓ Il project administrator (in genere il responsabile di quel gruppo di ricerca/esperimento/...)

The screenshot shows the Cloud Area Padovana dashboard. On the left is a sidebar with navigation links: Project, Admin, System Panel, Identity Panel, Domains, Projects, Users, Groups, Roles, and Registrations (highlighted). The main content area displays the 'Registrations' panel with a table of user registrations.

ID	User name	First name	Last name	Organization	Phone number	Domain	Actions
9	dorigoa	Alvise	Dorigo	INFN	049.967.7378	Default	Process
8	sgaravat@infn.it	Massimo	Sgaravatto	INFN Padova	+39 049 123	Default	Process

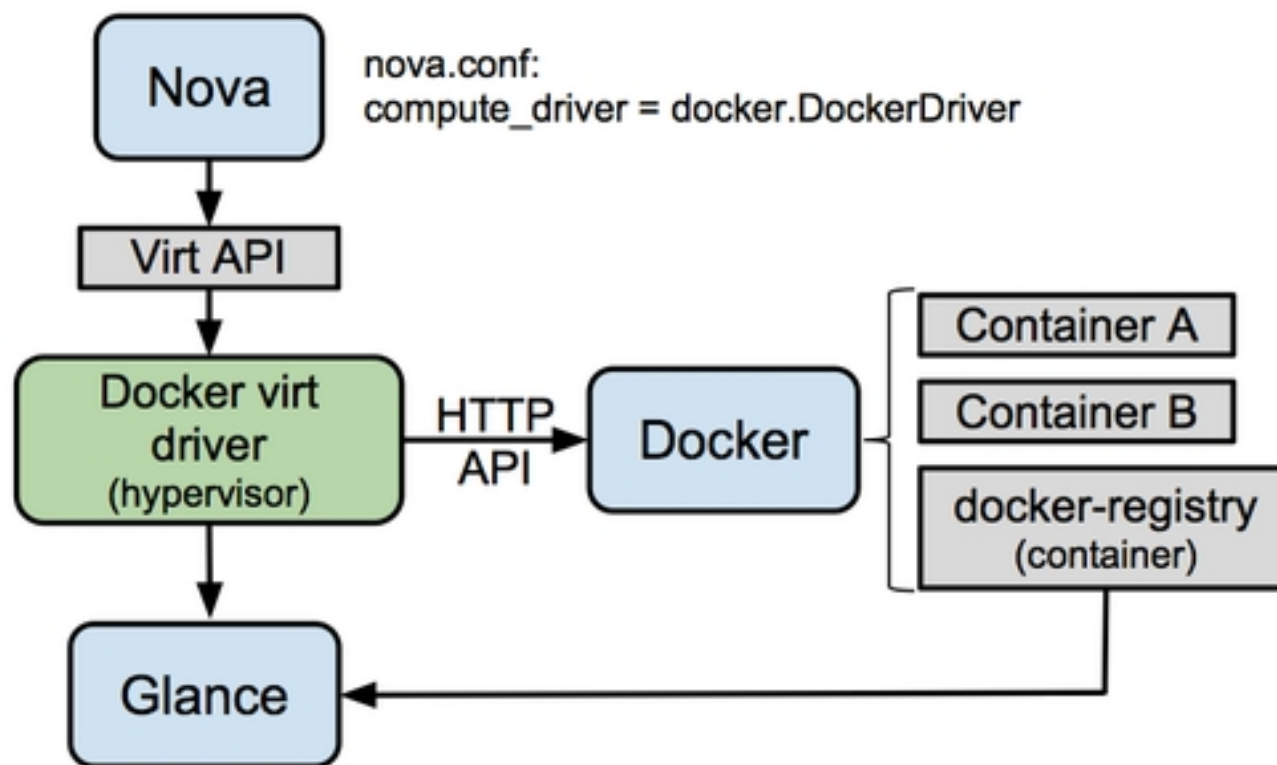
Below the table, it says 'Displaying 2 items'. To the right of the table is the 'User Registration' form with fields for: User ID (andreett@infn.it), Project action (Select existing projects), Select existing project (cudateam), Organization (INFN), Phone number (049.967.7378), Contact person (Mirco Mazzucato), and Notes. At the bottom, there are 'Log In' and 'Read the AUP' buttons.

Integrato nella dashboard:

- User registration form
- Panel per registration request management
- Panel per project subscription management

Supporto per container

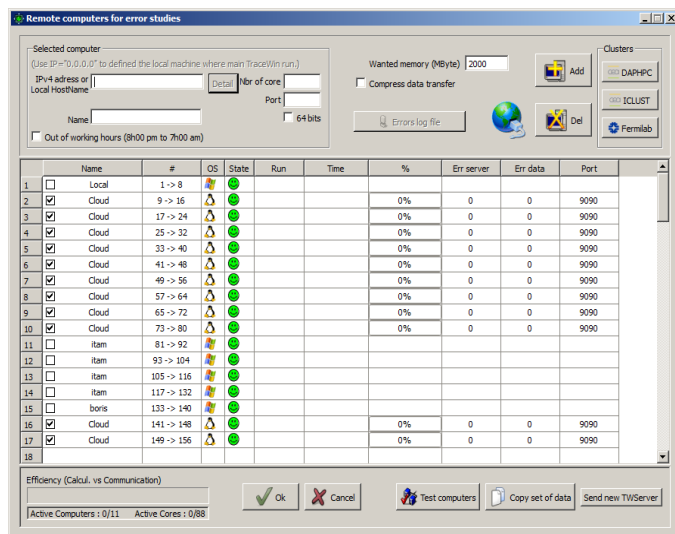
- Valutato **nova-docker** nell'ambito del progetto INDIGO DataCloud
- V. <http://tinyurl.com/hfr88xz>
- Permette di istanziare container usando gli stessi tool usati per creare Virtual Machine (dashboard, OpenStack nova CLI, OCCl, ...)



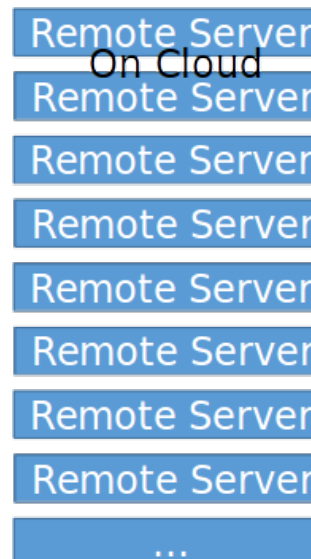
- Pronti per installarlo in produzione, per il momento su un singolo compute node

Alcuni casi d'uso (1/2): SPES

- Una delle attività di calcolo di SPES è la simulazione
 - per il tuning di parametri *perturbati* dell'acceleratore
- Necessità di numero di simulazioni molto elevato in breve tempo
- L'uso “elastico” della Cloud AreaPadovana **abbatte di un fattore dieci i tempi** per ottenere i risultati per la progettazione della macchina
- Un software di simulazione (**TraceWin**) gestisce fino ad un migliaio di VM multicore distribuendovi i job di simulazione usando un'architettura client-server



TraceWin Client
with GUI on local
desktop



Alcuni casi d'uso (2/2): CMS



- **Uso interattivo:**

- compilazione codice, produzione interattive di ntuple, analisi, sottomissione al cluster LSF del Tier-2
- ogni utente istanzia la sua VM *personale*, del flavour più opportuno, e la distrugge quando non gli serve più
- storage *esterno* alla Cloud: cluster Lustre

- **Uso batch:**

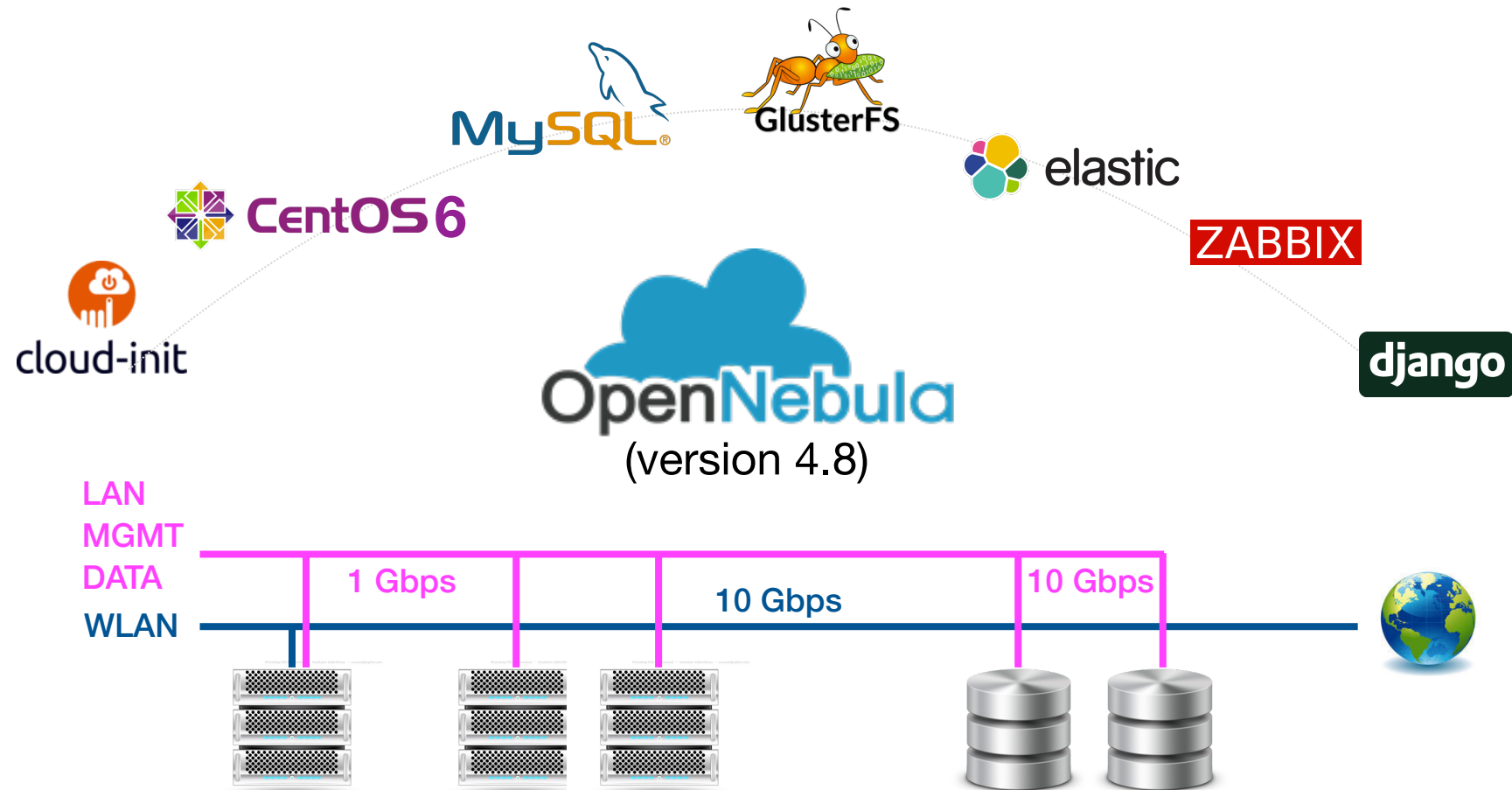
- cluster elastico (HTCondor): Elastiq automaticamente crea nuove VM/distrugge VM a seconda delle esigenze
- immagini μ cernVM based
- cluster istanziato tramite euca tool e contestualizzato con user_data
- cluster *multiutente*

- Firmata una **lettera di intenti tra dieci dipartimenti dell'Università, la Sezione di Padova dell'INFN e i Laboratori Nazionali di Legnaro** *per lo sviluppo, la messa in opera e sperimentazione di un Centro pilota di Elaborazione Dati Cloud a Padova – CED-C ad alte prestazioni a sostegno della ricerca dei partner coinvolti*
 - nell'ottica di unire gli sforzi mettendo in comune le competenze esistenti e le esperienze già fatte
 - per creare un centro di riferimento tecnologico e di competenze a livello regionale
- Modello implementativo considerato:
 - uso delle risorse Cloud dell'Università per la prima implementazione
 - ~ 500 core in HT, ~ 100 TB di storage
 - fisicamente stanno nella sala CED Dip. Fisica/INFN Padova
 - successiva integrazione delle risorse INFN, una volta verificato che tutto funziona
 - deciso di *esporre* questo servizio Cloud con un nome *neutro* (i.e. non UniPD, non INFN)
 - scelto di chiedere il dominio **cloudveneto.it**

- La prima implementazione (quella che coinvolge le sole risorse hardware di UniPd) è stata completata
- Il servizio è stato presentato (e aperto agli utenti) a fine **Novembre 2015**
- Al momento 7 progetti attivi:
 - Chimica, Scienze Biomediche, Fisica, Medicina Molecolare, Ing. Civile, Ing. Informatica, Scienze Farmaceutiche

La Cloud Torinese

Torino



Services cluster (4 hosts)

- in/out-bound connectivity
- shared system datastore (live migration)
- no special I/O requirements
- resiliency optimised FS (RAID1)

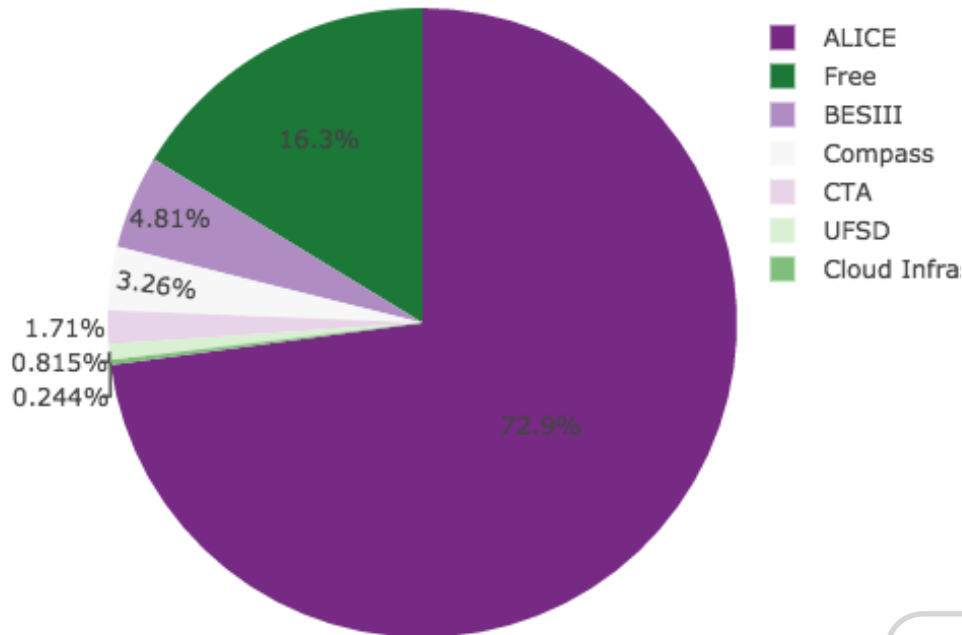
Workers cluster (69 hosts)

- private IP only
- high storage I/O performance
- locally cached image datastore
- performance optimised FS

- 1.3k cores
- about 200 running VMs
- virtualisation: KVM
- host OS: CentOS 6

Storage resources

OUT Cloud

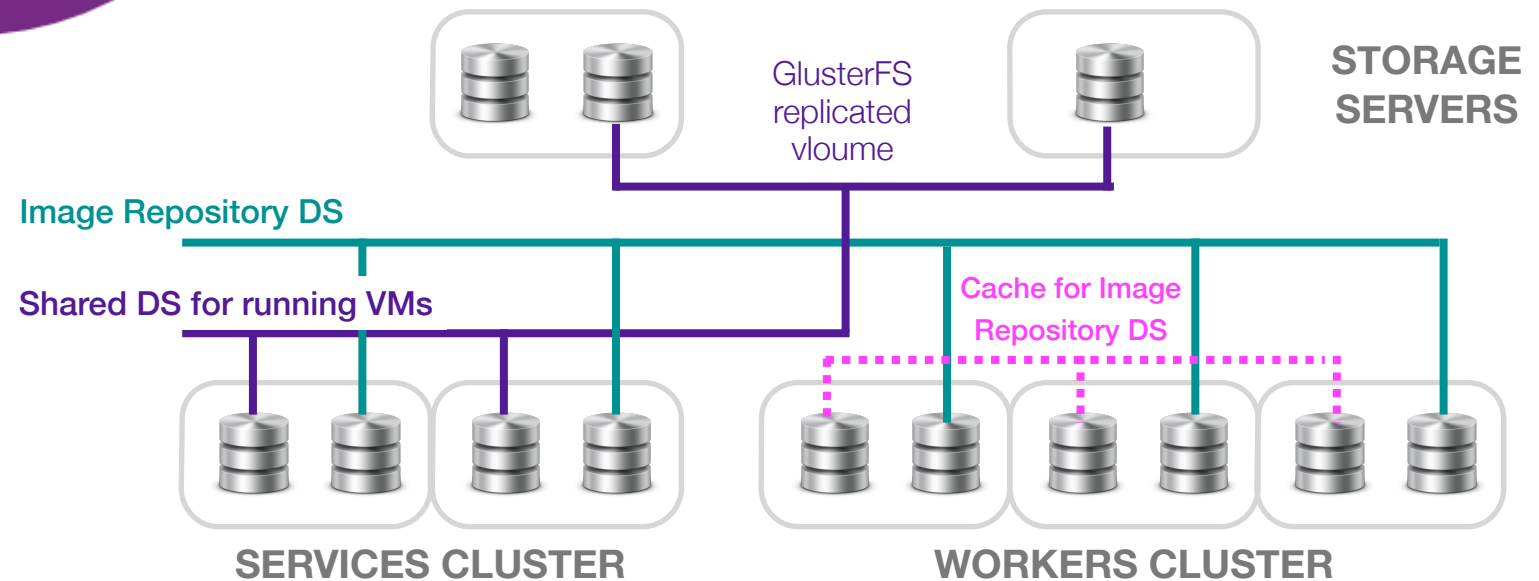


- total 1.8 PB lordi
- 2 access servers per l'infrastruttura
- 4 access servers per i dati (XRootD, StoRM)
- *image datastore*: 2 TB
- *system datastore* per i servizi: 1 TB (shared, mirrored)
- *image datastore* per i workers: circa 300 GB per host (cache locale, script custom di sincronizzazione)

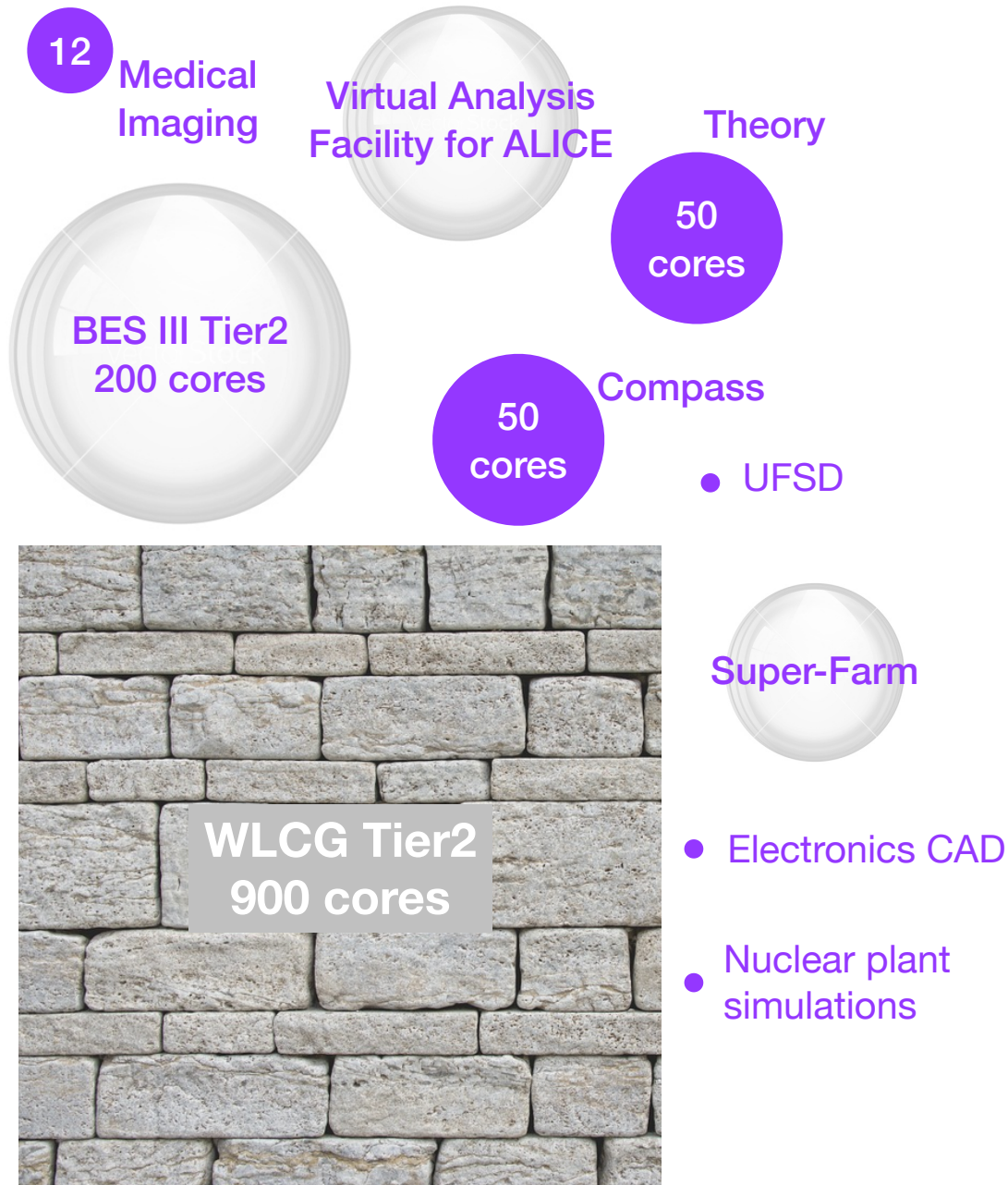
IN Cloud

- iSCSI server
 - iSCSI NAS
 - datastore scp
- Total: 28 TB**

Typical use case:
persistent home NFS
exported
within priv. net.



Multi-tenancy



Sandboxes

Network isolation (level 2):

- una rete virtuale privata per utente (255.255.255.0)
- ed una pubblica (255.255.255.255)
- reti isolate con ebttables sul bridge dell'hyprvisor

Virtual routers (level 3):

- IP pubblico e privato
- light-weight OpenWRT VM (1 CPU, 150 MB)
- funzionalità DHCP, DNS, NAT
- firewalling and port-forwarding
- management in mano all'amministratore
- IP pubblici elastici (ad 1 istanza per sandbox)

Virtual farms on-demand

- applicazione confinata in sandbox
- configurazione e deploy della farm tramite EC2 compliant tools o GUI
- spazio disco persistente iSCSI
- Amazon-like flavours

Self-service virtual farms

EVF provisioning user portal
Home Documentation Logout

You are logged in as user: svallero

You farm definitions

Id	Name	Description
15	Example Ubuntu Server 14.04	Example with Ubuntu Server 14.04 image. Exports the home via NFS.
17	Example CentOS 6.6	Example with CentOS 6.6 image. Exports the home via NFS.

My virtual farms
Create a new virtual farm
Cloud dashboard



Powered by the INFN Torino Computing Centre

EVF provisioning user portal
Home Documentation Logout

You are logged in as user: svallero

Configure a new farm

Farm name:

Farm description:

EC2 access key:

EC2 secret key:

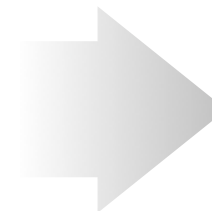
Root ssh key:

OS image:

Master flavour:

My virtual farms
Create a new virtual farm
Cloud dashboard

- realizzato con Django
- analogo a CERNVM-Online
- immagini CentOS 6 o Ubuntu
- cluster istanziato automaticamente sulla Cloud locale (boto)



Condor shared secret:
Queue polling time (s):
Minimum jobs waiting time (s):
Number of jobs per VM:
VMs polling time (s):
Maximum VM idle time (s):
Min number of workers:
Max number of workers:
VM deploy time (s):

Elasticità a livello IaaS

Elastiq

- **demone Python** custom
(<https://github.com/dberzano/elastiq>)
- comunica con ON tramite **interfaccia EC2**
(Cloud agnostic)
- plugin per **HTCondor** (Cloud aware)
- l'utente istanza una singola VM (il master)
- configurazione di Elastiq e dei workers specificata nel *context* del master
- SCALE-UP: se ci sono job in coda
- SCALE-DOWN: quando la VM è idle
- casi d'uso:
 - Virtual Analysis Facility per ALICE
 - Super-Farm
- **migrazione in corso anche per le farm utenti**

OneFlow

- ON tool per applicazioni tiered e auto-scaling
- ideale **per applicazioni stateless** (load balancing)
- ma nel nostro caso d'uso (BESIII Tier2):
 - bunch di jobs con durata simile (da alcune ore a giorni)
 - risorse allocate quasi sempre idle
 - applicazione **stateful**
 - LRMS non Cloud-aware (Torque)
- OneFlow Master: CREAM compute element
- OneFlow Slaves: DIRAC GRID worker-nodes
- CE e WN pubblicano lo stato della coda su OneGate
- SCALE-UP: 1 nodo alla volta se job in coda
- SCALE-DOWN: tutti i nodi quando la coda è vuota

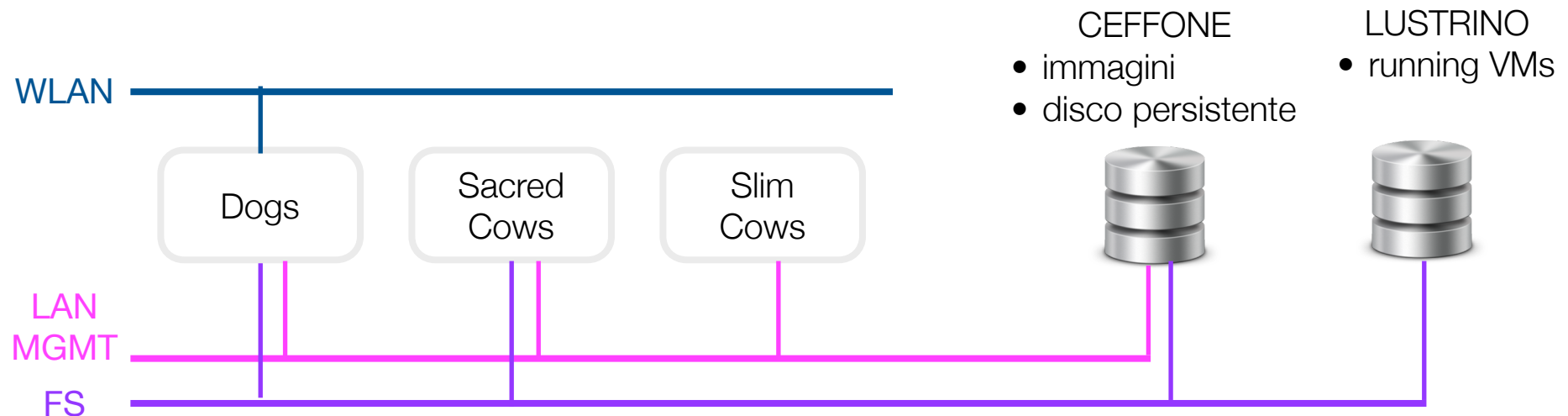
Cosa crediamo di aver imparato...

- passati al paradigma Cloud IaaS nel 2011 per la condivisione delle risorse tra sito GRID e Analysis Facility
- ne seguono una **semplificazione della gestione del sito** ed una **maggiore flessibilità**, che portano un numero di casi d'uso minori
- ma le risorse non sono infinite... gli utenti vanno **incentivati al rilascio delle risorse**
- inoltre alcuni casi d'uso sono per natura non elastici (il Tier2 WLCG non è mai idle)
- l'**interazione con gli utenti e gli esperimenti** è fondamentale
- ci sono diverse tipologie di utenti:
 - ironicamente gli use-cases minori sono più demanding
 - richiedono soluzioni molto specifiche (singola VM)
 - in generale scarsa motivazione a risolvere i problemi in maniera autosufficiente
- un **maggiore livello di astrazione** potrebbe portare beneficio ad utenti ed amministratori
- da un punto di vista tecnico:
 - si è optato per **OpenNebula** perché al momento della decisione OpenStack non era sufficientemente maturo: scelta vincente per il nostro specifico caso d'uso (semplicità ed economia di gestione)
 - necessità di una **classe intermedia di server** con filesystem condiviso (meno di un servizio critico, più di un semplice worker-node)
 - necessario un **refurbishment dell'infrastruttura di rete**

La Cloud Torinese 2.0

Progettare un'infrastruttura più resiliente e facile da gestire:

- non più patch custom ad OpenNebula (difficili da supportare)
- back-end database set-up più robusto
- servizio di storage persistente iSCSI più robusto
- provisioning semplificato (Cobbler+Puppet → Ansible?)
- ridefinizione della topologia di rete ed accesso allo storage
- migrazione ad un modello di gestione delle reti virtuali più mainstream (i.e. VRouter → OpenVSwitch)
- 3 classi di server (clusters)



Un nuovo approccio alla multi-tenancy

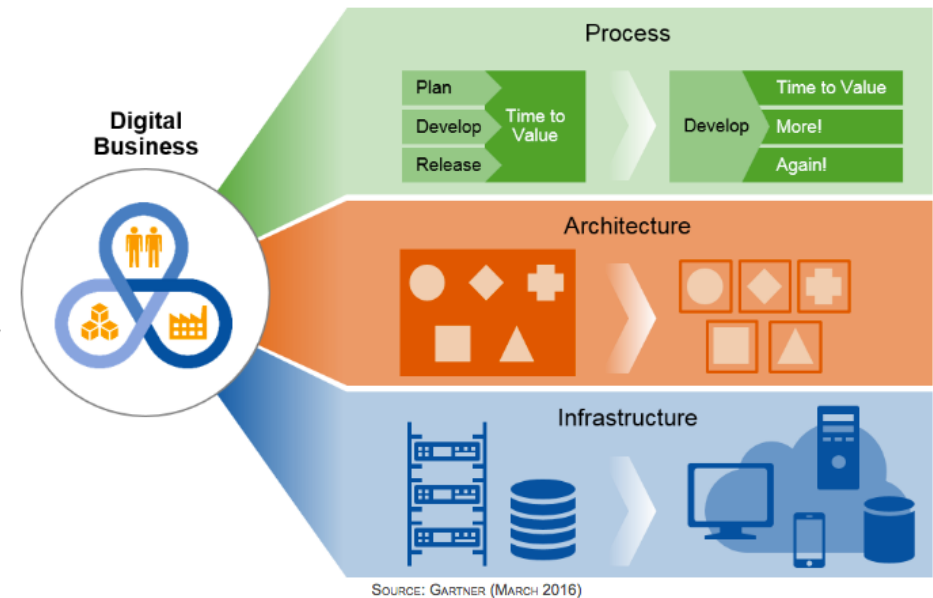
Leggendo qua e là sul WEB...

4 Reasons to use Apache Mesos frameworks

Encapsulating tasks into higher and higher abstractions is being shown by various large companies to provide a competitive advantage...

...virtual machines with little to no democratic process over how resources are utilised...

...free a developer from antiquated bureaucracy so that they are able to spend time on developing products and services, not work-arounds for legacy systems.



Containerizing application workloads (GARTNER)

...to package, ship and run distributed application components with guaranteed platform parity across different environments.

... "democratizing" virtualization by providing it to developers in a usable, application-focused form. Whereas access to virtual machine virtualisation tends to be provided through, and governed by, gatekeepers in infrastructure and operations, Docker is being adopted from the ground up by developers using a DevOps approach.

Our next step (in the INDIGO context): the dockerized Virtual Farm

- provide batch farm on-demand in a Linux **container**
- **HTCondor**
- application orchestration/elasticity with Apache **Mesos**
- Calico for sandboxing (**L3 approach**)

Nuova infrastruttura HPC: strategie e sinergie



Finanziamento di 900 kEUR dalla Compagnia di S. Paolo a UNITO per la realizzazione di un cluster HPC multiuso inter-dipartimentale



- un cluster “di produzione” per le esigenze di calcolo scientifico dei dipartimenti
- una piattaforma per attività di R&D nell’ambito del calcolo scientifico
- un forum per raccogliere le esperienze e il know-how
- una “entità per la presentazione di progetti: il centro dovrà autosostenersi per non morire.

“Supercomputer a geometria variabile” - Richieste minime da capitolato:

- 32 nodi “light” (convenzionali 2-socket)
- 4 nodi “fat” (4-socket, 712 GB)
- 4 nodi “GPU” (2-socket + 2x nVidia K40)
- 80 TB storage High Performance
- 800 TB storage “archival”
- Connettività InfiniBand (56 Mb/s, Fat Tree) & Ethernet 10Gb/s
- 2 nodi management e rete Ethernet 1Gb/s per IPMI



La nuova sala macchine

Torino



Prima

Adesso



Per concludere

- integrazione di farm locali e risorse GRID nella Cloud
- l'esperienza acquisita nella gestione di sistemi di calcolo, in particolare con tecnologie Cloud, permette di iniziare nuove **collaborazioni con le università**
- il lavoro nel contesto di **INDIGO-DataCloud** aiuta già da adesso la gestione dei centri
- le infrastrutture stanno arrivando a maturità, ma la tecnologia è in sviluppo e trasformazione continua (Linux containers), per fortuna sempre più basata su **tool mainstream e non sviluppati in proprio**