Alvise Dorigo
INFN Padova
On behalf of the "Cloud dell'area Padovana" team

INFN CCR Workshop
Frascati May 2015

Cloud Area Padovana
an update...

We remind you that ...

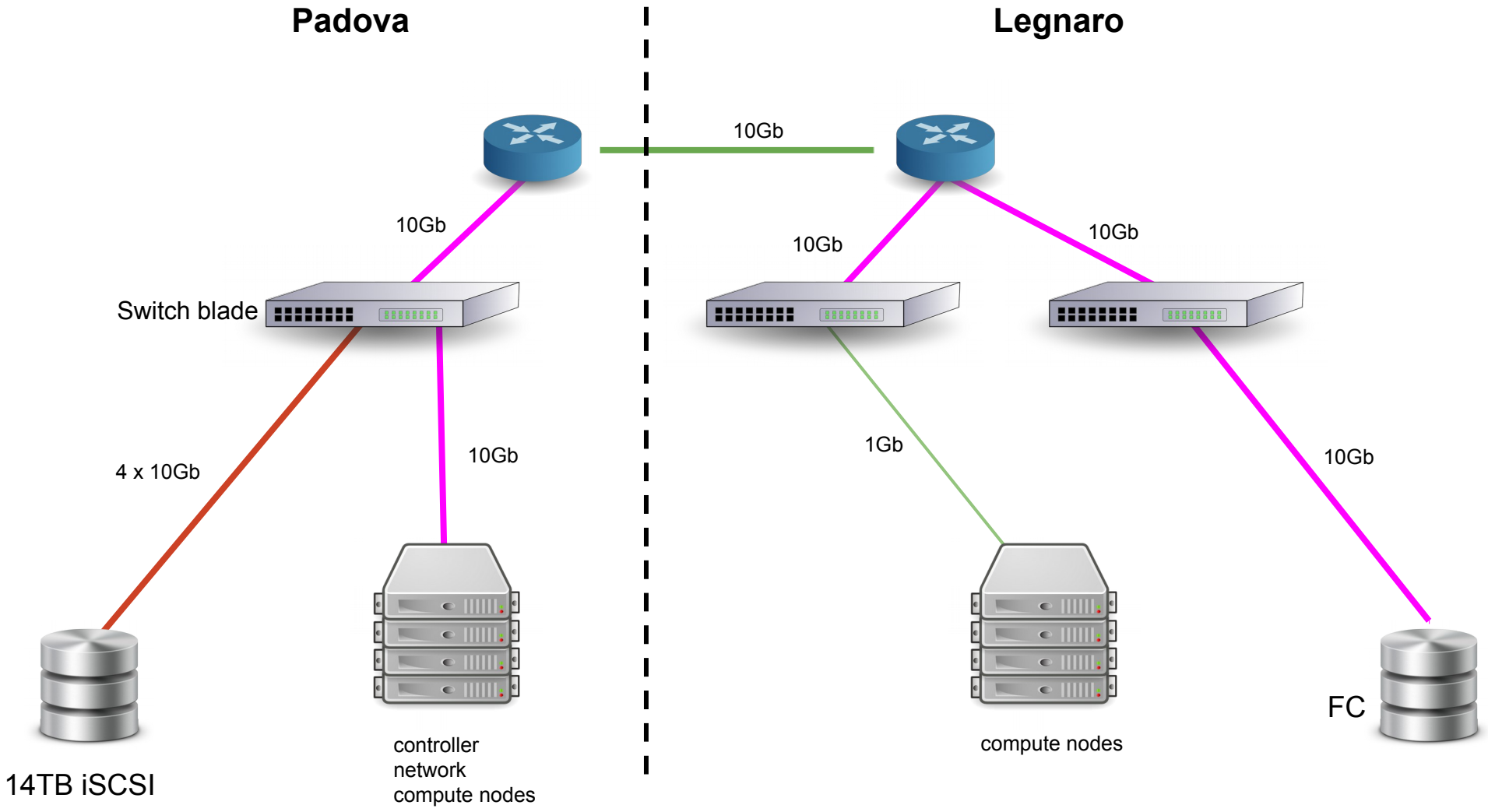
At the end of 2014 PD+LNL invested money and manpower to

- consolidate the computing resources among several research groups for a more efficient usage
- reduce the experiment owned farms' proliferation
- Sysadmin optimization
- May → October 2014 Pre-Production
- October 2014 Produzione

Current resources (May 2015)

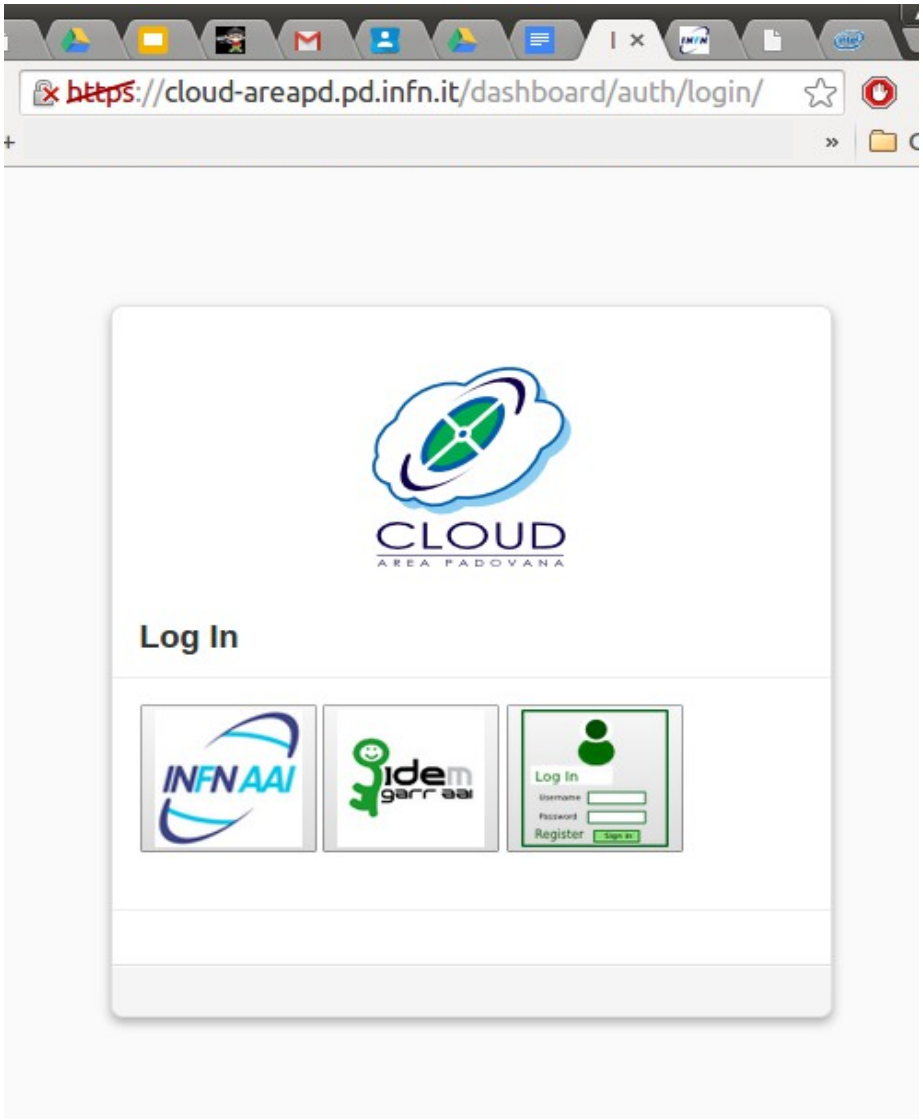
- 736 cores for VM (overcommit 4:1, 2944 VCPU), 2240 GB RAM (overcommit 1.5:1, 3360 GB RAM)
  - **PD**: 5 blades DELL M620, 2 CPU E5-2670v2 each one, 96 GB RAM each one
  - **PD**: 3 blades DELL M630, 2 CPU E5-2650v3 each one, 96 GB RAM
  - **LNL**: 6 Fujitsu Primergy RX300S8, 2 CPU E5-2650v2 each one, 96 GB RAM each one
    - To be installed: 7 Dell PowerEdge R430, 2 x E5-2640v3, 128GB, mixed financed (SPES, IFMIF, STI) – 224 core 896GB RAM
- 14 TB iSCSI DELL MD3620i in Padova (images, VM, volumes)
  - 48 TB to be installed
- 96 TB to be installed in LNL
- Less powerful resources for API services, database, high-availability, part of them is physical, another part is virtual

Current resources (May 2015)

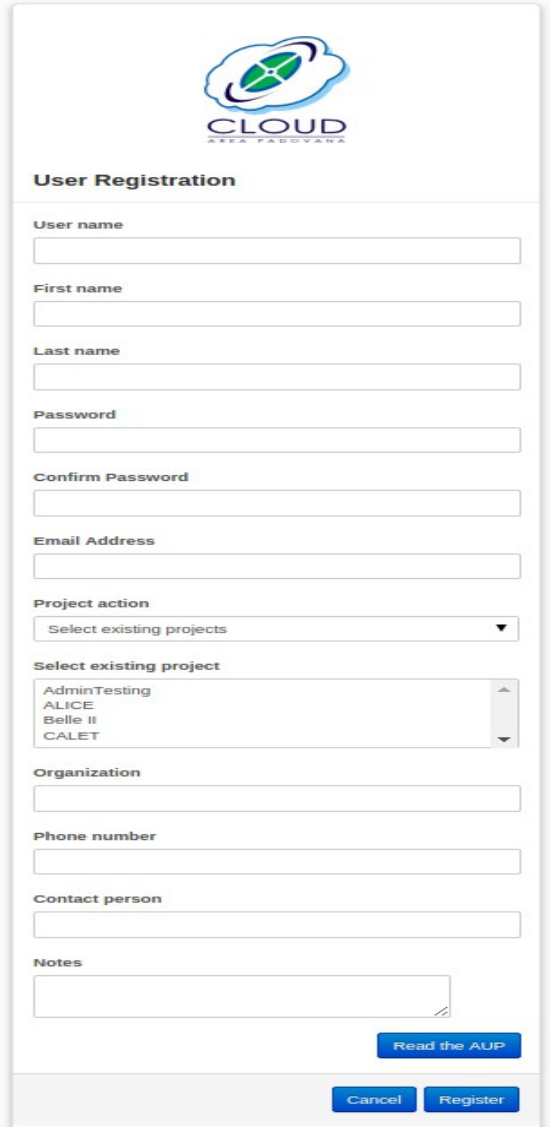


AuthN in the Cloud Area Padovana

- Integration with the INFN's IdP (SAML2) to manage authentication
- Integrated in the portal a module to manage user and project registrations (in-house development)



AuthN in the Cloud Area Padovana



Registration form in the Cloud Area-Padovana (home made)

OpenStack is the "de facto" stack – 1 / 2

- We use **Havana**
  - several bugs, but stable enough
- Migration to IceHouse scheduled for next week
  - We will also perform some changes in the OpenStack deployment configuration (see later)
- When planning the update, we were thinking to go directly to Juno, but:
  - No straightforward to update from release N to release N+2
  - IceHouse was (till a few days ago) the reference OpenStack release for OCP
- For the migration to Juno we first need to port the module for user/project registration

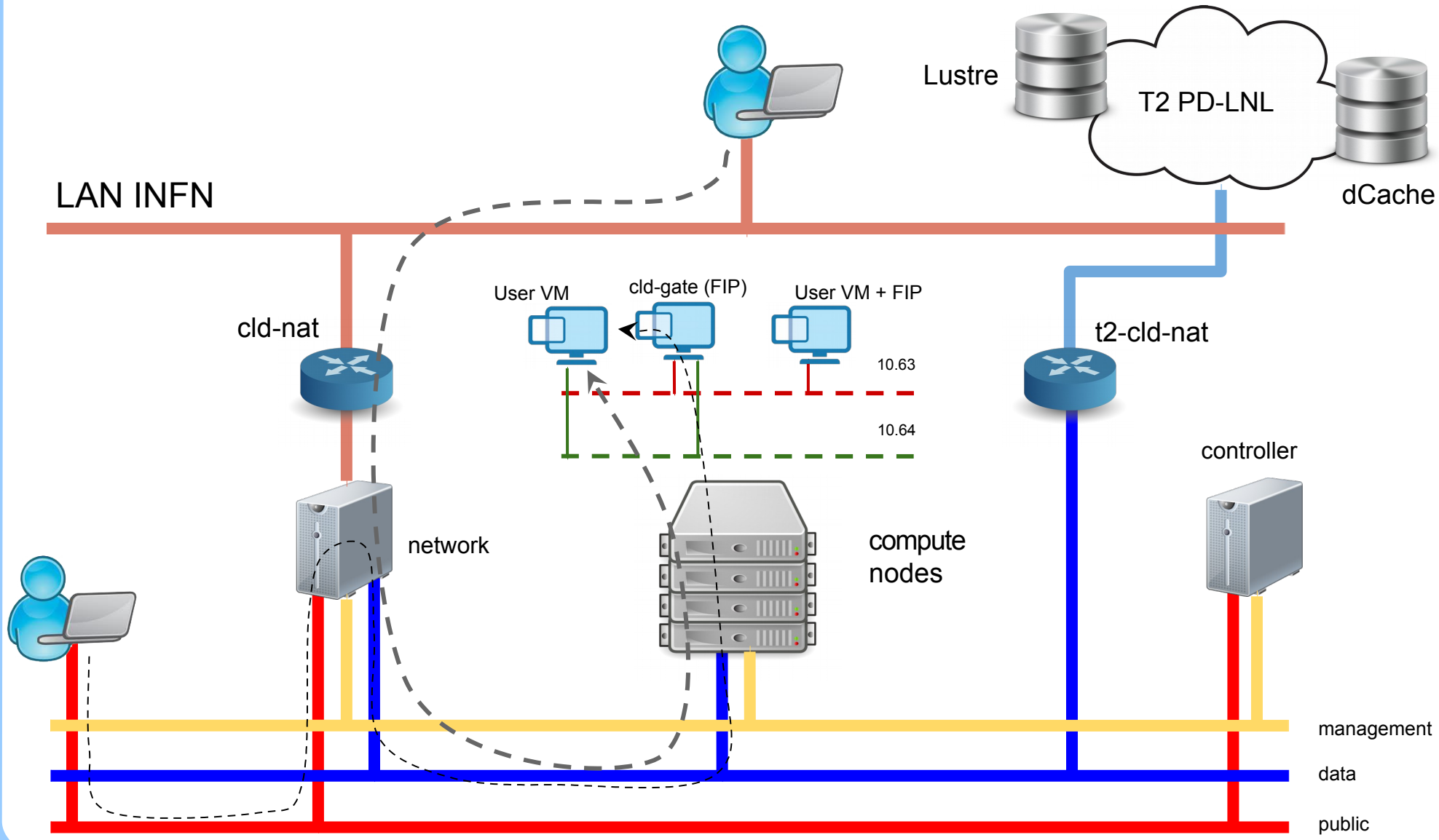
OpenStack is the "de facto" stack – 2 / 2

- Manual fine tuned installation (for services)
- active/active HA based on HAProxy/Keepalived (as suggested by OpenStack developers)
  - RabbitMQ uses a native HA clusterization
- All (not only keystone) SSL-enabled interfaces
- by mean of HAProxy configured as a SSL-terminator
- Padova: controller/network/compute
- Legnaro: compute/storage
  - Gluster to be configured in IceHouse

Networking (1 / 2)

- OpenvSwitch driver with GRE tunneling
  - “Provider router with private networks”
    - 10.64. (1 class C : 1 project) **without FloatingIP**
    - 10.63. (1 class C : 1 project) **with FloatingIP**
- **cld-nat** linux box **acting as router** to allow VM access from INFN LAN without FloatingIP
- **cld-gate** VM **acting as bastion** for VM access from Internet without using FloatingIP

Networking (2 / 2)



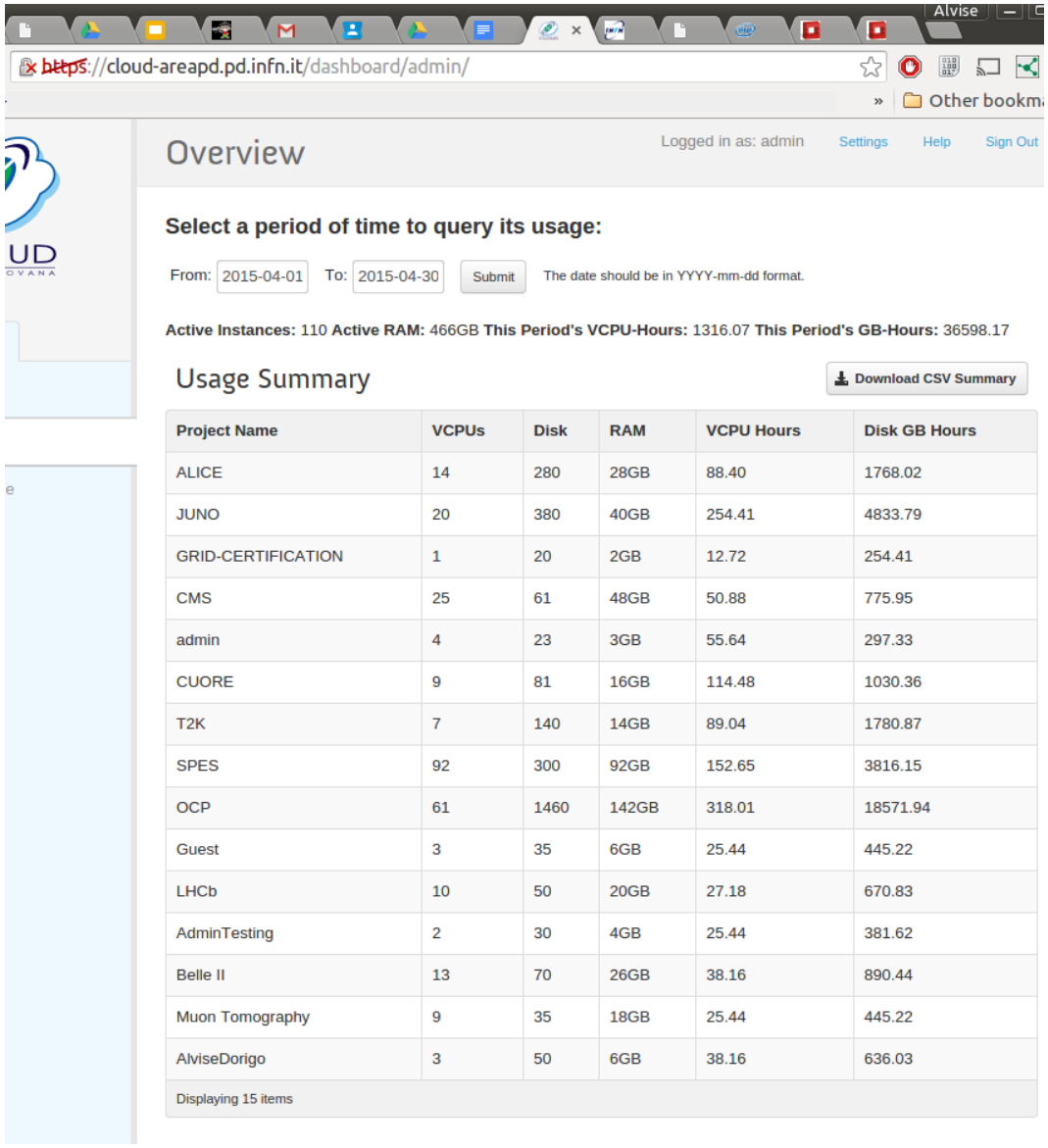
Storage (1 / 2)

- Storage iSCSI in Padova (images, VM, volumes)
  - Connected to the 2 controller nodes that act also as storage (Gluster) servers
  - This proved to be a bad strategy (e.g. VMs affected when performing a reboot of the controller to apply some updates)
  - With the migration to IceHouse we'll merge controller e network nodes on the same machine; the ex-controller nodes will became storage servers
- GlusterFS
  - 2 bricks in distributed method (for now)
  - We're evaluating an HA solution for storage (see later)
- **Shared FS for VM ⇒ Live Migration**
- No Swift (for the time being)

Storage (2 / 2)

- Possibility to access to "external" storage
  - Avoided possible bottleneck in the network node
  - installed L2 agent on the storage servers which become part of the IaaS (http://goo.gl/xseu61)
- Clients at the moment:
  - Muon Tomography (GlusterFS)
  - LHCb (Lustre)

Experiments using the Cloud



Project Name	VCPUs	Disk	RAM	VCPU Hours	Disk GB Hours
ALICE	14	280	28GB	88.40	1768.02
JUNO	20	380	40GB	254.41	4833.79
GRID-CERTIFICATION	1	20	2GB	12.72	254.41
CMS	25	61	48GB	50.88	775.95
admin	4	23	3GB	55.64	297.33
CUORE	9	81	16GB	114.48	1030.36
T2K	7	140	14GB	89.04	1780.87
SPES	92	300	92GB	152.65	3816.15
OCP	61	1460	142GB	318.01	18571.94
Guest	3	35	6GB	25.44	445.22
LHCb	10	50	20GB	27.18	670.83
AdminTesting	2	30	4GB	25.44	381.62
Belle II	13	70	26GB	38.16	890.44
Muon Tomography	9	35	18GB	25.44	445.22
AlviseDorigo	3	50	6GB	38.16	636.03

Displaying 15 items

Belle II

Current usage of the Cloud:

- Software release test
- Code benchmarking for TOP detector
- TOP reconstruction check
- MonteCarlo production
  - on average 2 cores for 24h, ~ 18k evts per core per day
- Development and testing of a system for building the software release and checking for compilation errors/warnings and memory leaks
- Import and analyze cosmic rays data as they are available
- Integrating the Cloud Padovana in the Belle II's computing model as it is ready for the Cloud

CMS

Replicated in the Cloud the physical production environment (UI cluster for interactive job cpu-demanding):

4x32 cores, 256 GB RAM

- Grid UI cluster with LFS clients
- Analysis/prod code available via CVMFS
- Users's home imported from shared LustreFS
- LDAP users management
- Shared machine among users

Current usage of the Cloud:
Execution of CPU demanding jobs for:

- software compilation
- interactive small ntuple production
- multiprocess analysis

Desiderata on the Cloud:

- Large VMs for massive production
- Interested on elastic batch cluster (VM-WN on demand)

Make the users able to create their own VM (no sharing anymore) with required CPU/RAM.

Manager of the CMS cloud project provides a suitable qcow image to the users with all installed/configured software
Integrated Cloud with T2, so that Cloud users can efficiently access T2 storage (dCache)

LHCb

- High level of parallelism (multiprocess) required to skim 460M evts to barely 2M
- Initial copious evts grouped by N of source files
  - ⇒ N VMs are required
- Comp. time drops from days to hours
- Cloud's benefits:
  - No hardware management; self-service provisioning when needed; release of resources when not used
  - For some applications simpler user interaction compared to Grid (e.g. no middleware layers to interact with)
  - Easily customizable images to span VMs compared to O.S. installation/configuration/update on several machines

SPES

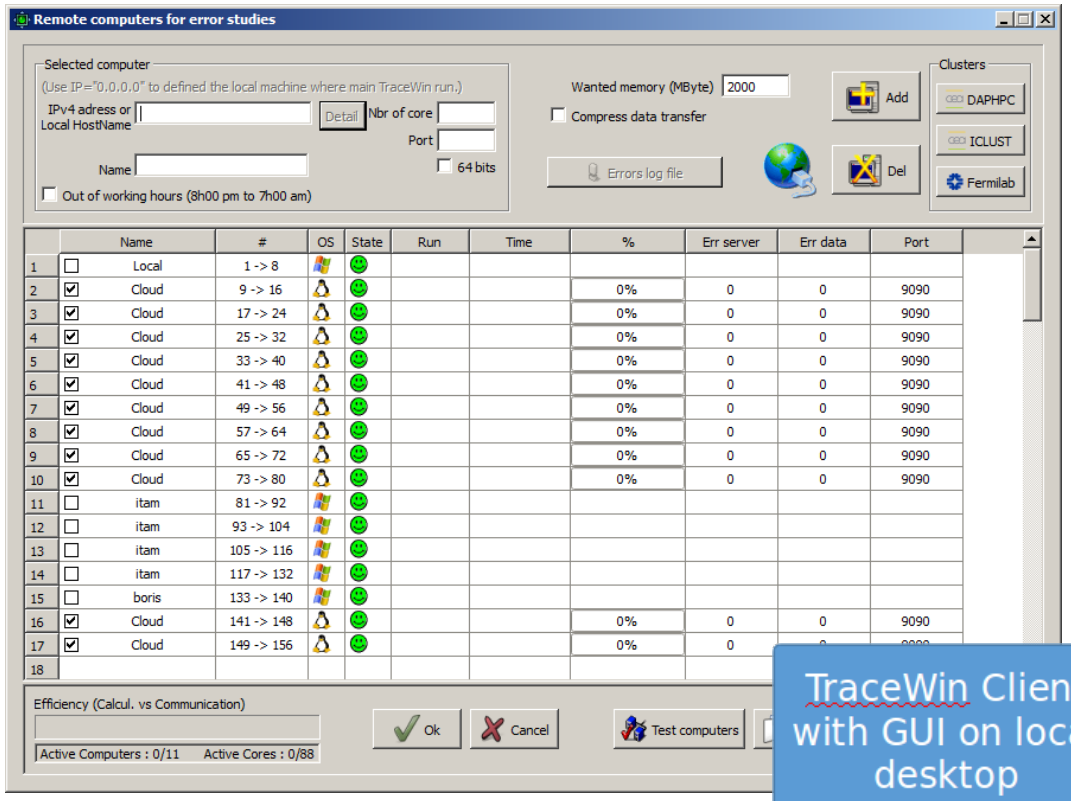
Use case:

- Beam dynamics errors studies need a lot of “short” runs (less than 5 minutes for single run, with less than 1 GB of RAM for single run)
  - ~200k runs (varying several params, changing resolution)
- Total run time ~16.6k hrs ~694 days (~2yrs!)
  - **MANDATORY PARALLELISM**

With Cloud:

- several computational servers (statically) spawned on the cloud
  - compute nodes funded by SPES exp.
  - 8 CPU demanding computational processes per VM (~40 VMs, 8 VCPU 8GB RAM)
- client sends run to them bringing down the total computational time

SPES



TraceWin Client with GUI on local desktop

Local data taking, storage and analisis

TCP/UDP comm

Remote Server On Cloud
Remote Server
Remote Server
Remote Server
Remote Server
Remote Server
Remote Server
Remote Server
...

- A remote server is installed as service on Cloud (any operating system), listen on a defined UDP port.
- A main local client send commands for **N, asynchronous runs** to be done on Cloud, typically 100 – 1000 runs.
- When the runs are finished the main local client collects all the results.

ALICE

- Cloud used for ALICE's **Virtual Analysis Facility** related activities
- Currently used by INFN Trieste people

Sinergy with OCP

- Cloud Area Padovana leverages on the experience and solutions developed by planning the OpenStack deployment in the OCP reference regions
- Converging in the use of the OCP blessed Foreman/Puppet tools for automatic installation (Padova's contribution to these implementations)
- OCP project made available to OCP personnel for PaaS testing/evaluation (CloudFondry + OpenShift) and other activities

Operations

- Full automatic installation of compute nodes by mean of **Foreman/puppet**
  - Puppet also used for other services (Nagios and Ganglia nodes)
- **Nagios**-based monitoring infrastructure
  - alert system for every single failing service/daemon on controller, network and compute nodes
  - customized nagios sensor for known problems
  - as a new kind of problem appears, a new ad-hoc sensor is implemented
- **Shift-based support** (helpdesk, and monitor of services' health)

Sinergy with Padova's University (1 / 3)

- Reminder:
- Signed a memorandum of understanding between 10 University's departments, INFN-PD and LNL *“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”*
  - share the existing know-how and the acquired experience with the INFN Cloud
  - create a regional technological center with strong Cloud know-how, for the research and public administration environments
- Agreement on the possible implementation
  - First installation of a Cloud with exclusive UniPD's resources
  - Subsequent integration with INFN's Cloud

Sinergy with Padova's University (2 / 3)

- Hardware (136 phys. cores, 2048GB RAM, 122 TB disk)
  - 4 blades for services, two E5-2609, 32 GB RAM
  - 12 blades for computation, with two E5-2670v2, 160 GB RAM
  - Equallogic with 17x1.2TB SAS 10000 rpm (20 TB) + 7 x 800GB SSD (6 TB)
  - Equallogic containing 24 x 4TB SAS 7200 rpm (96 TB)
- Currently hosted in INFN/Physics Dept. CED

Sinergy with Padova's University (3 / 3)

Cloud implementation using the resources being finalized:

- OpenStack IceHouse using CentOS7
- Separating Storage and OpenStack services.Controller and network services merged on the same machine (redundant for HA)
  - *lesson learned from the Cloud Area Padovana*
- Using Equalogic driver for Cinder
- Percona MySQL DB, distributing the OpenStack services (accessing the DB) among three Percona instances
- Assessing possibile HA solutions for storage for images and instances

END