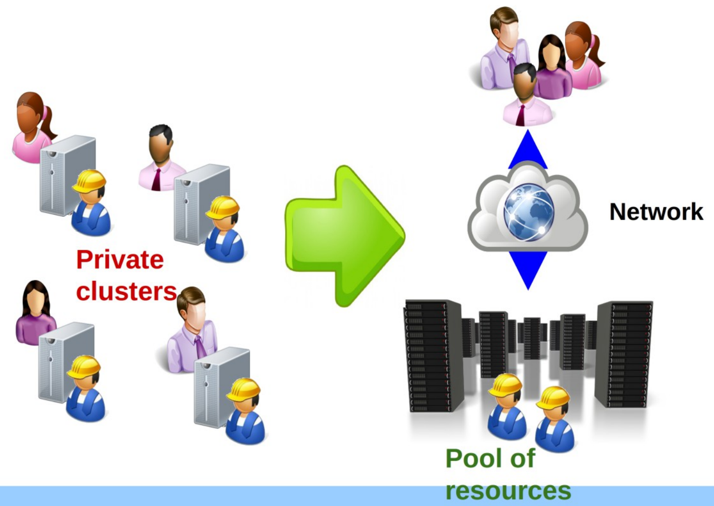
Cloud Area Padovana: Implementation and operations of a distributed IaaS

Massimo Sgaravatto
INFN Padova

On behalf of the Cloud Area Padovana team

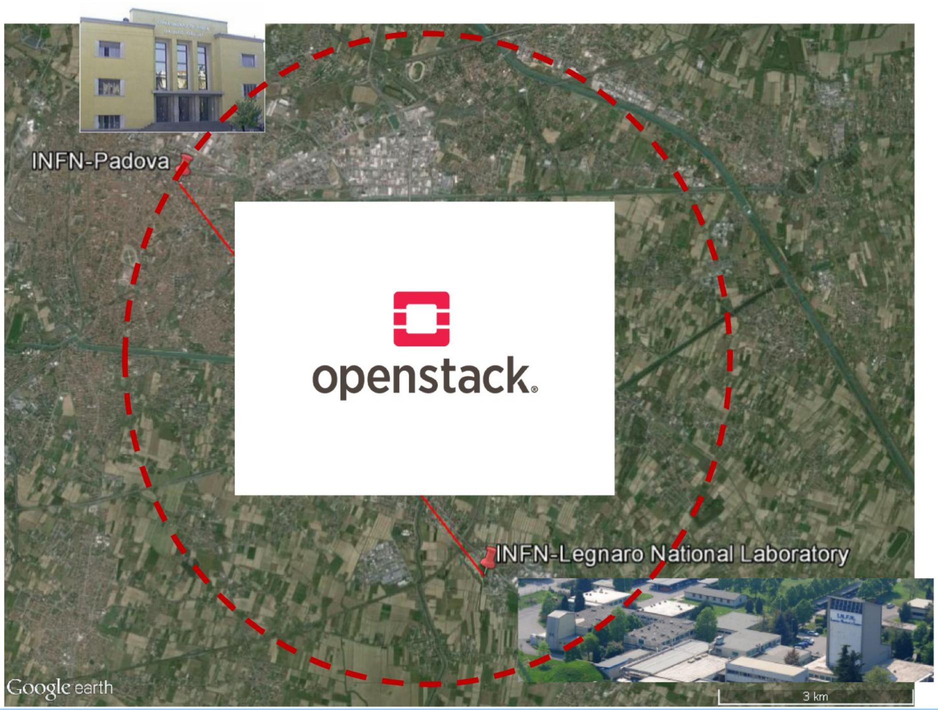
Cloud Area Padovana

- Project started at the end of 2013 for realizing a Cloud infrastructure
- Targeted in particular to use cases that can not be easily ported in the existing Grid infrastructure
- Main goal: improve the overall computing resources usage and decrease their maintenance costs
- Share of resources, manpower, competences between INFN Padova and INFN-Legnaro
  - As done for the Padova-Legnaro WLCG Tier-2



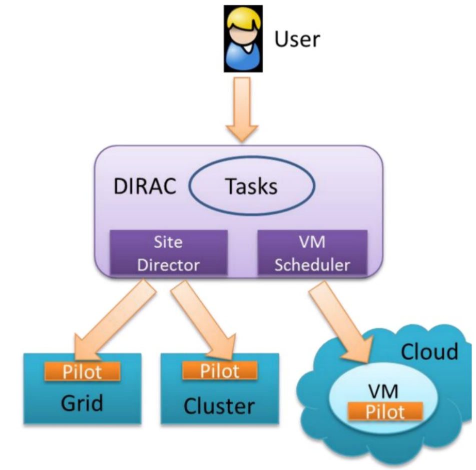
Cloud Area Padovana: current status

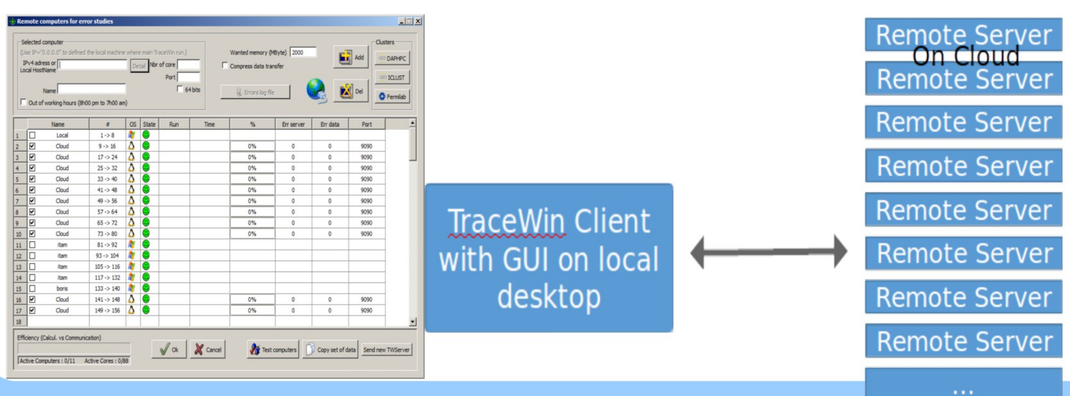
- Single OpenStack based IaaS with compute nodes spread among the two sites
- Services are instead deployed in a single location (Padova)
- In production since the end of 2014, after a pilot phase

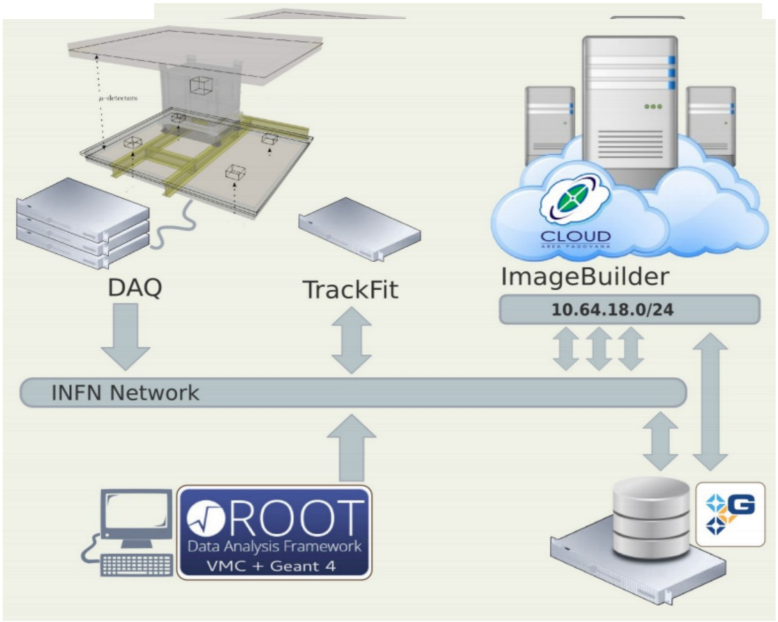


Cloud Area Padovana: usage

- ~ 120 registered users belonging to ~ 30 projects
  - Interactive activities, batch like jobs, services, ...
- Research groups invited to use (and invest) on this infrastructure instead of buying new private machines

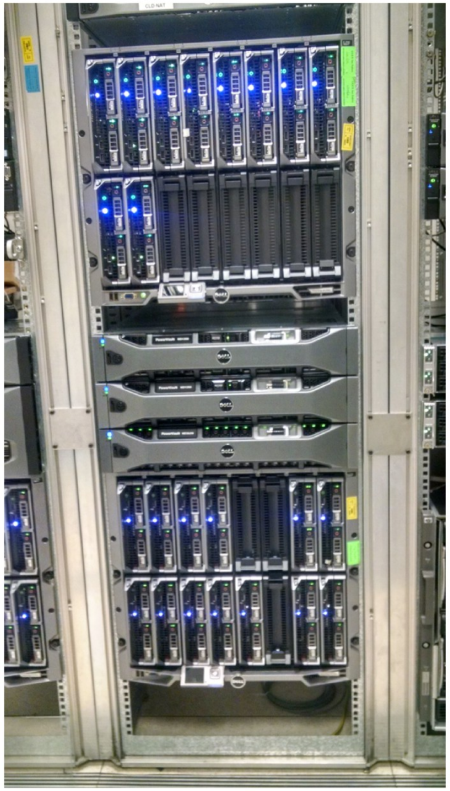




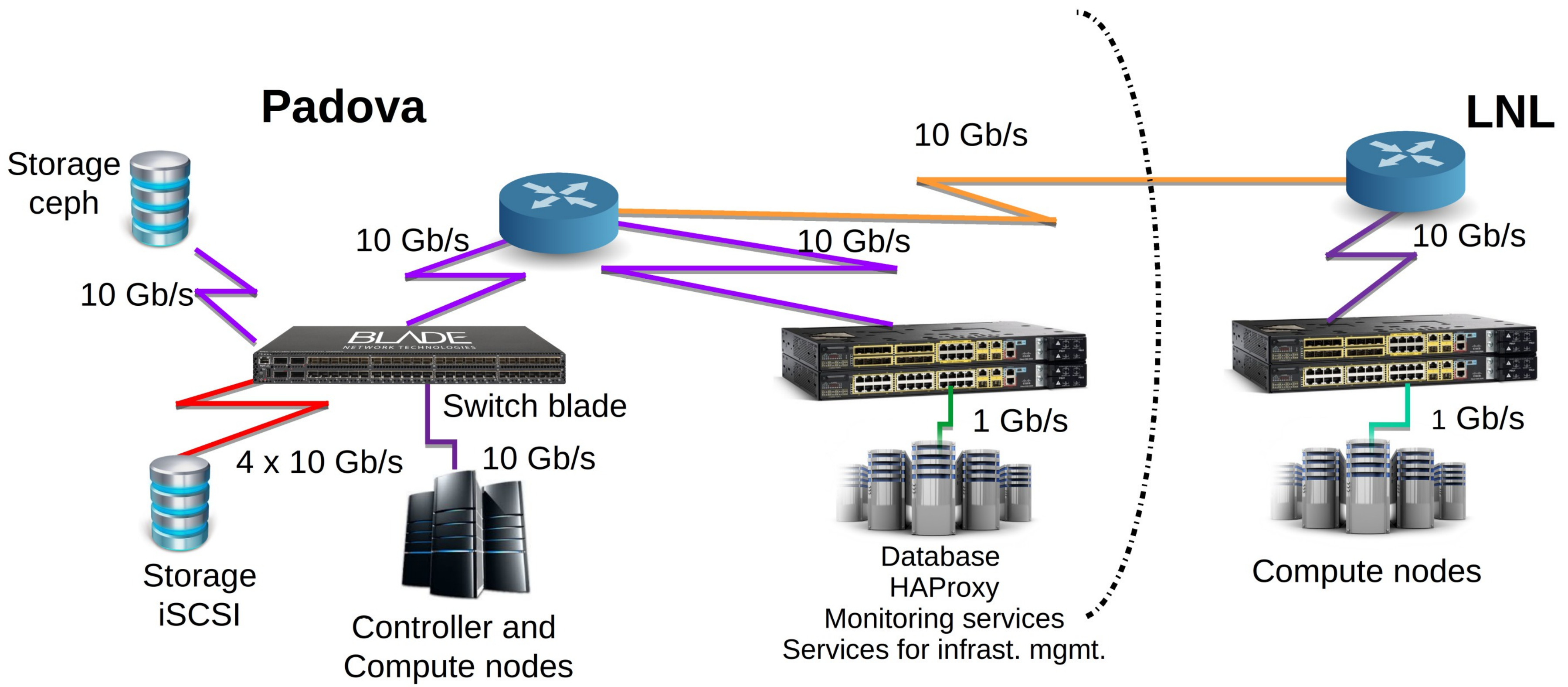


Resources

	Compute Nodes	Cores (in HT)	RAM (GB)	HS06	Storage for images and volumes (TB)
Padova	15	656	2048	7060	187
LNL	13	416	1472	4390	
Total	**28**	**1072**	**3520**	**11450**	**187**



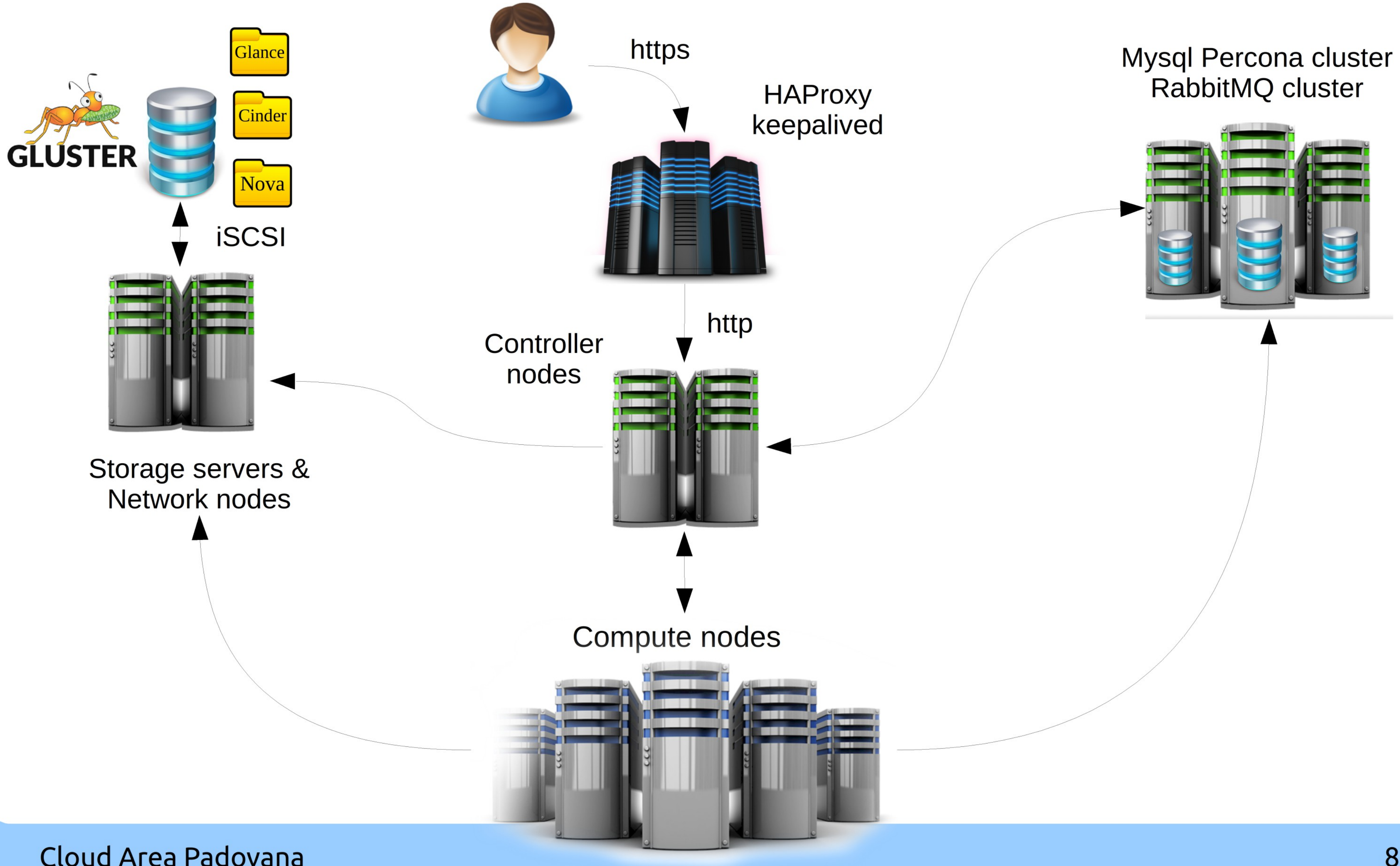
“High level” topology



Service deployment

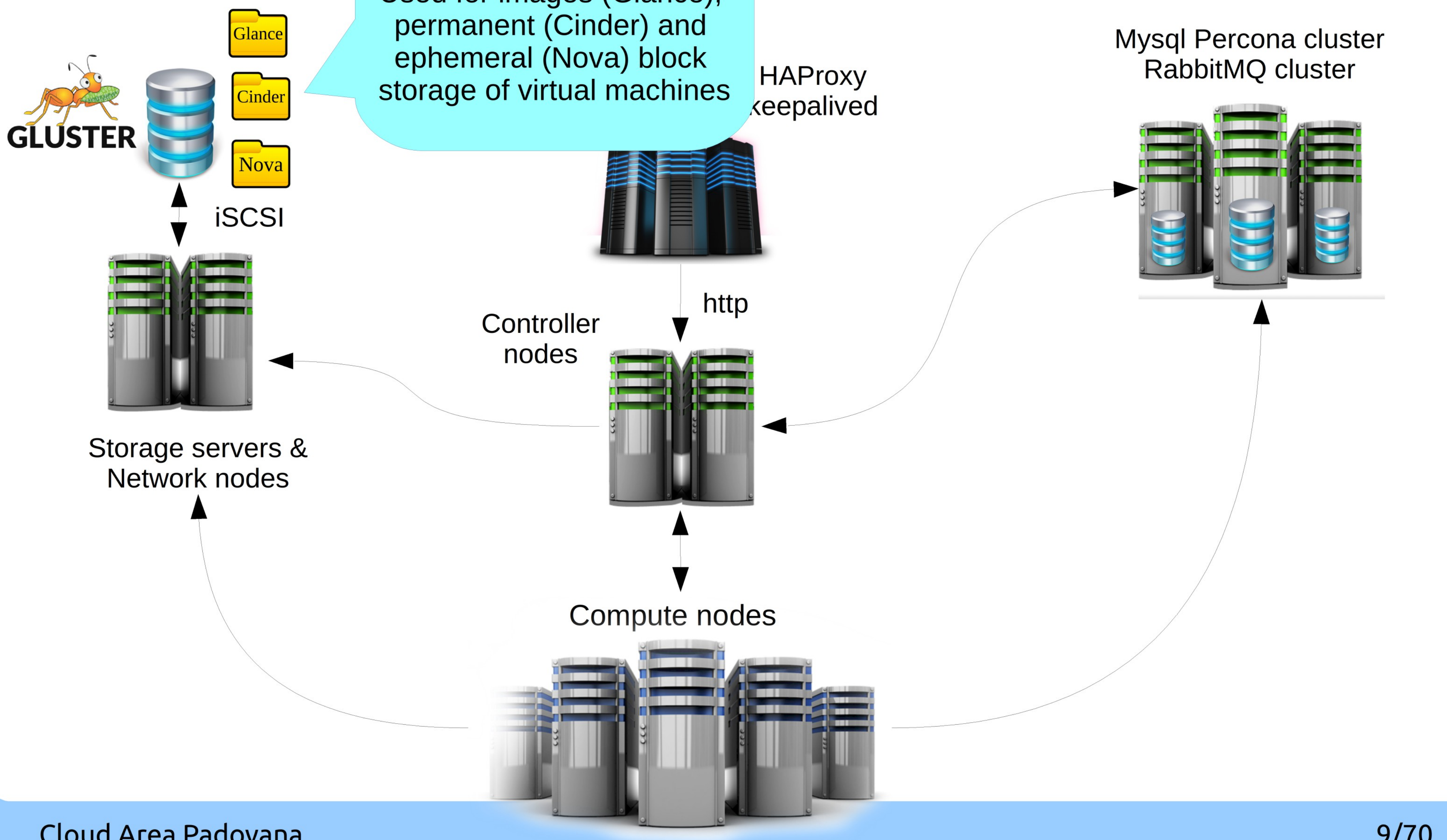
- Deployed OpenStack services:
  - “Core” services (keystone, glance, nova, neutron, cinder, horizon)
  - Heat (orchestration)
  - Ec2-service (to provide an Amazon EC2 compatible interface)
  - Nova-docker (to instantiate docker containers)
  - Ceilometer (accounting)
- Integrated also some in-house developments
  - To manage authentication relying on existing identity providers
  - To manage project and user registration
  - To manage accounting data
- Service deployment evolved and revised as needed

Service deployment (I version)

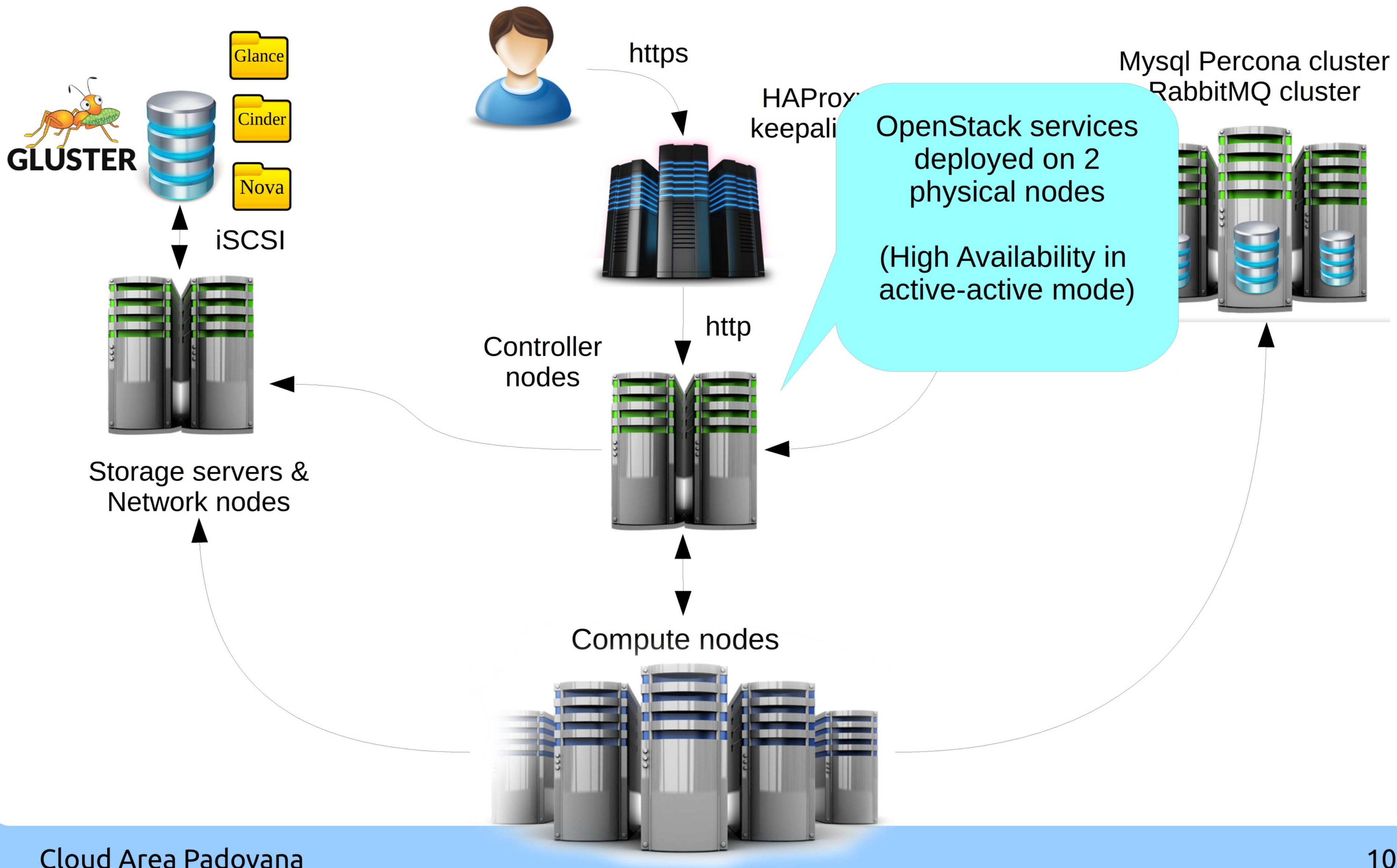


Se… ent (I version)

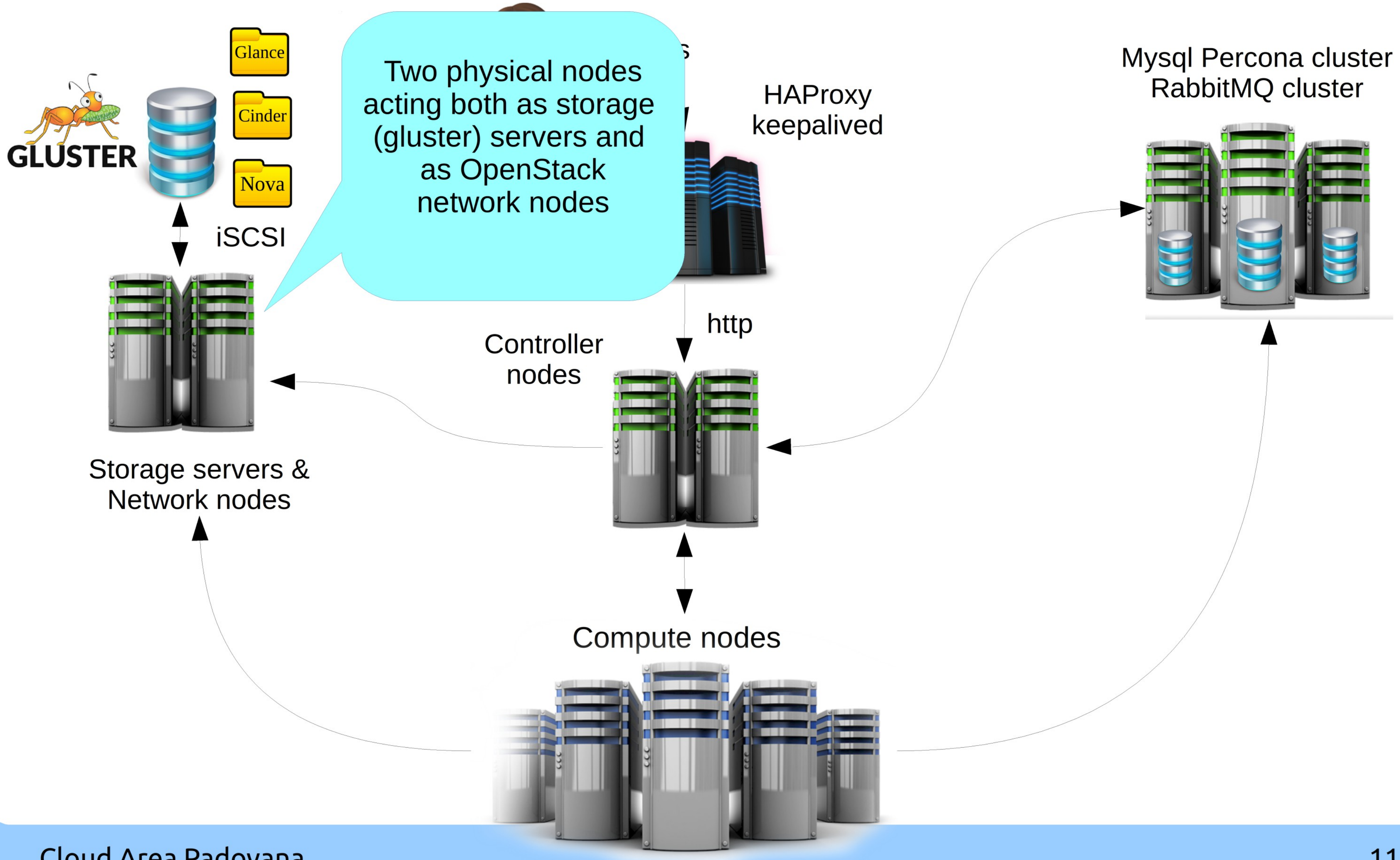
iSCSI storage
exposed via GlusterFS
(2 servers).
No replica.
Used for images (Glance),
permanent (Cinder) and
ephemeral (Nova) block
storage of virtual machines



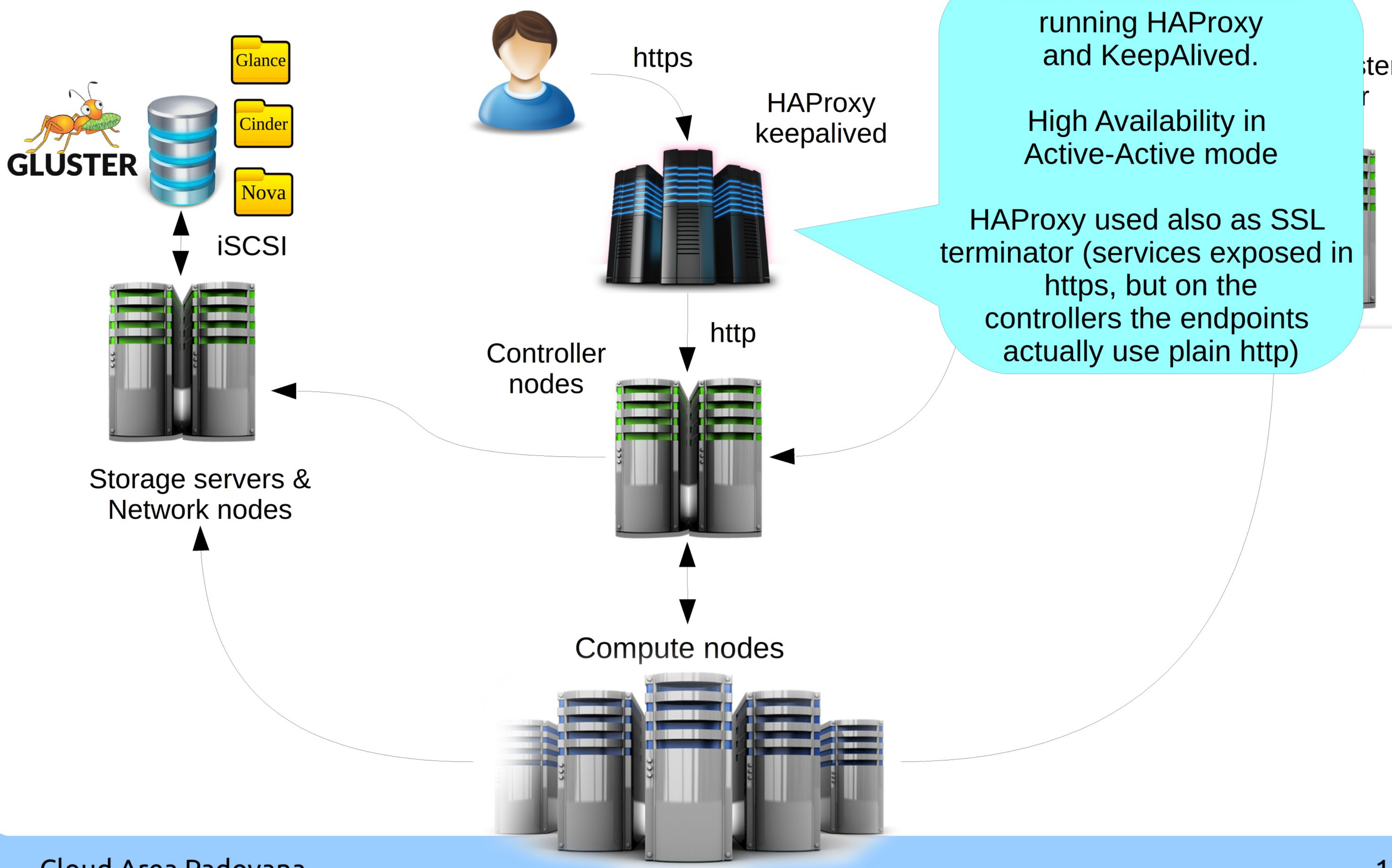
Service deployment (I version)



Service deployment (I version)



Service deployment (I version)

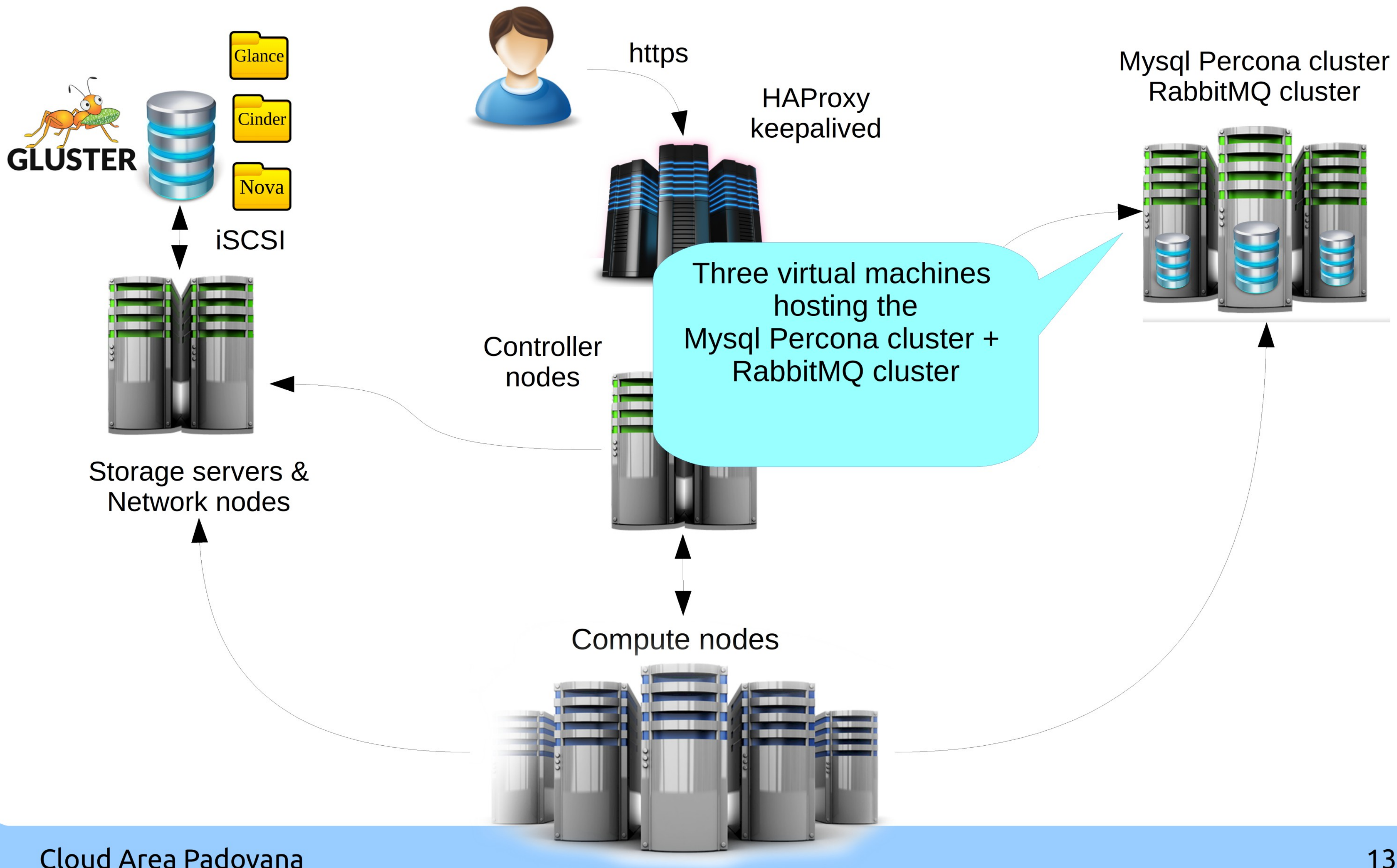


Three virtual machines running HAProxy and KeepAlived.

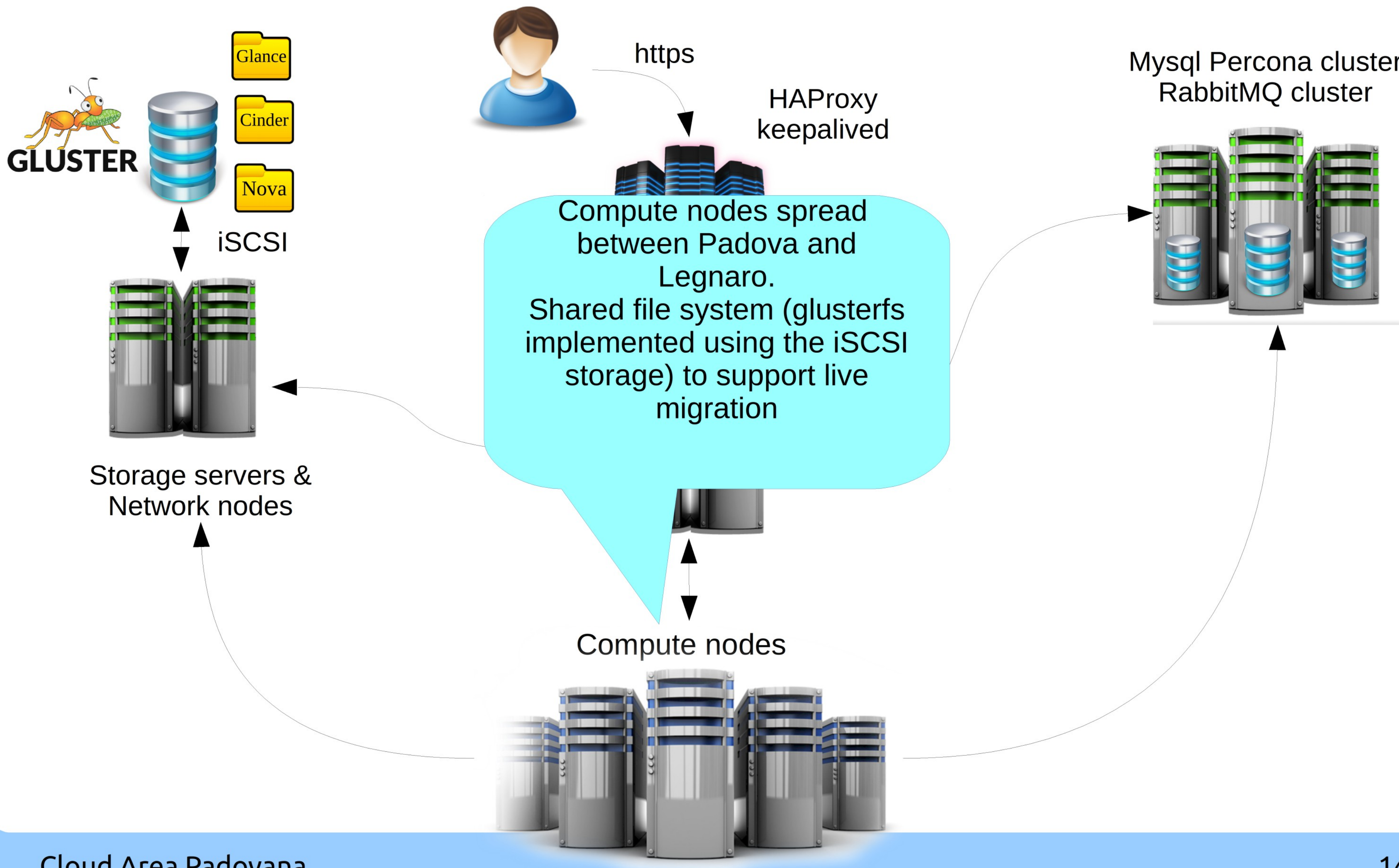
High Availability in Active-Active mode

HAProxy used also as SSL terminator (services exposed in https, but on the controllers the endpoints actually use plain http)

Service deployment (I version)



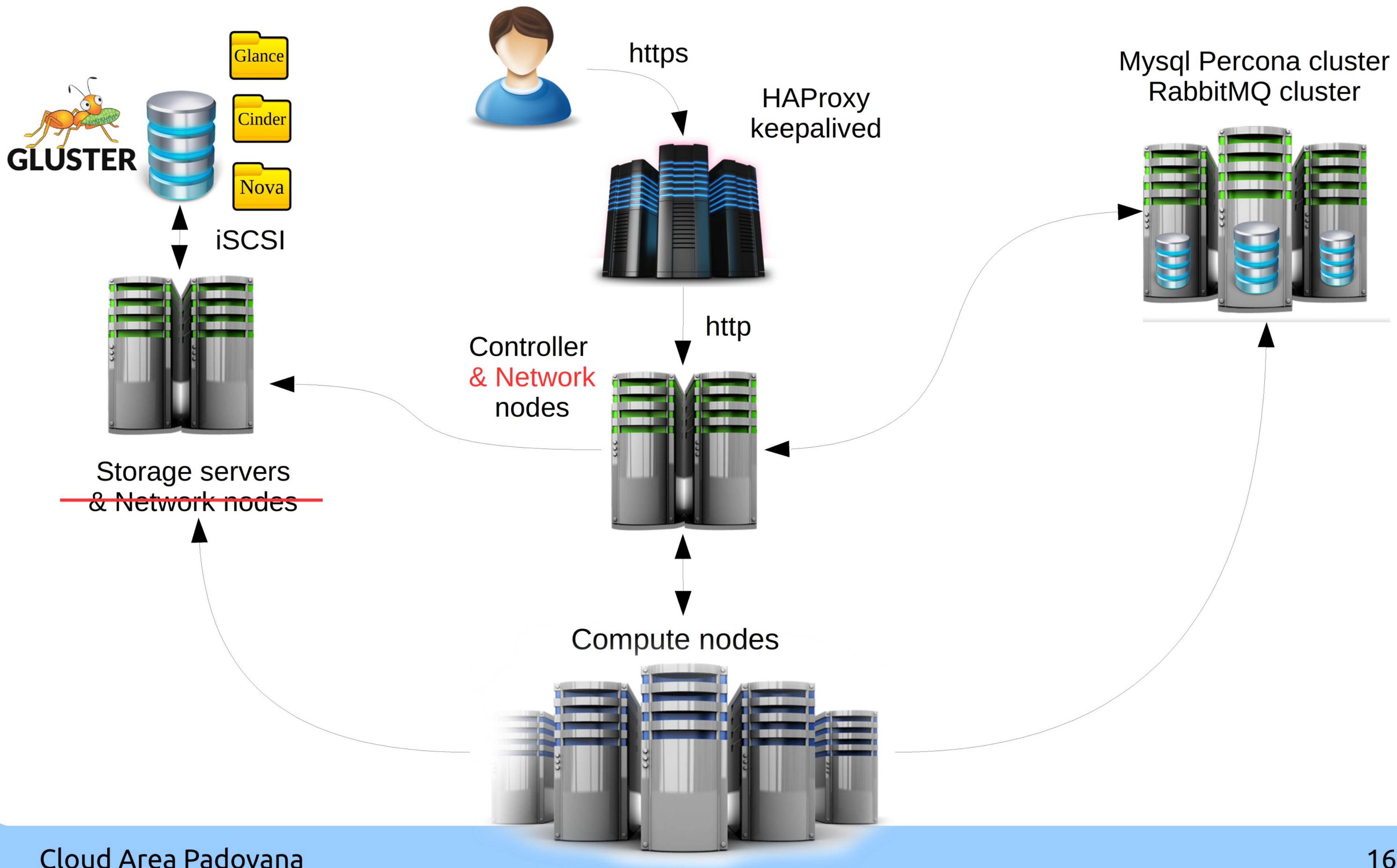
Service deployment (I version)



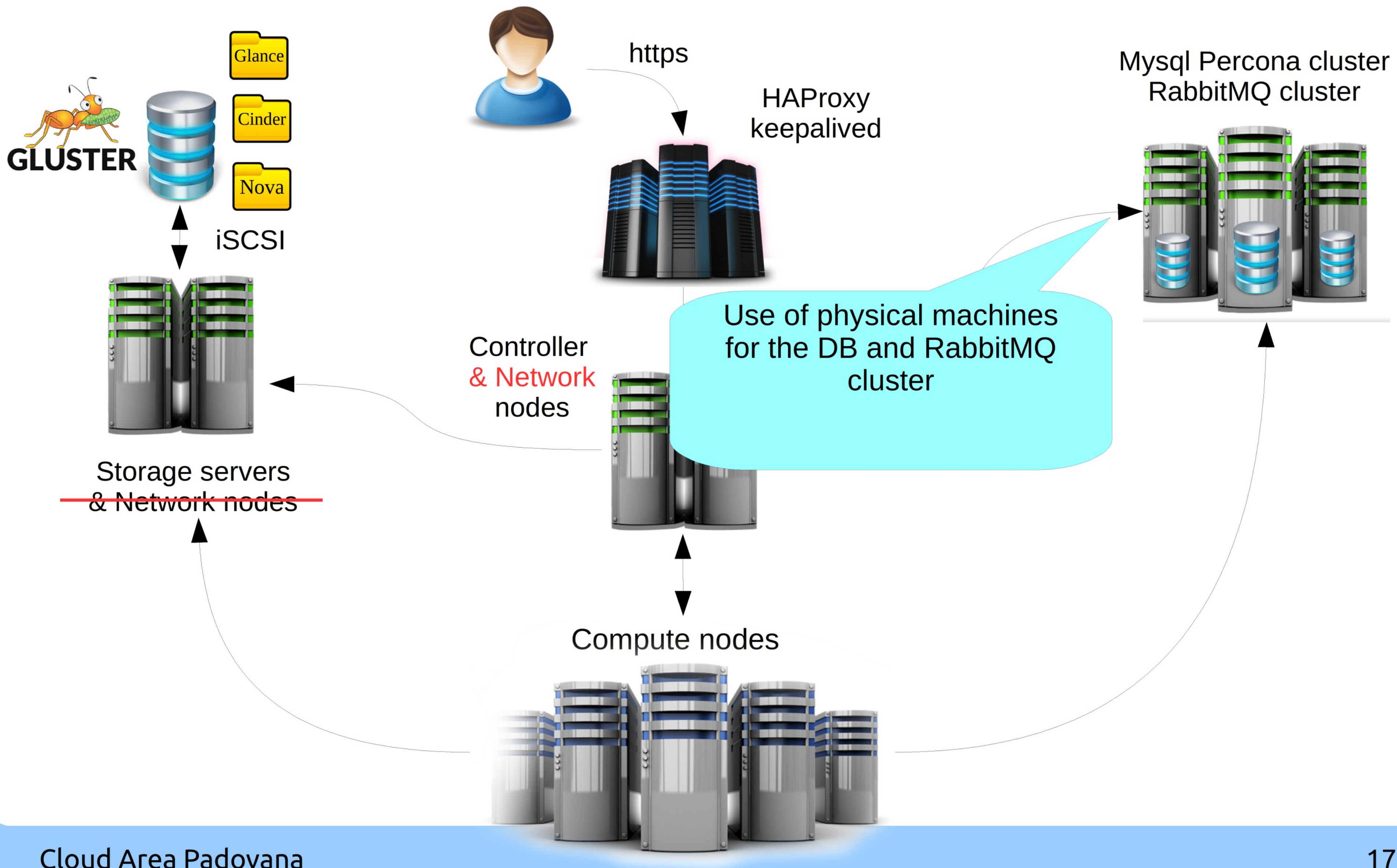
Issues with this 1st deployment

- Using the same box as storage server and network node proved to be a problem
  - E.g if you need to reboot a network node, impact also on the storage exposed by that node
- In general mixing storage servers with other services is not a great idea
- We also experienced DB performance problems
  - ➡ Move to physical nodes
  - ➡ DB tuning (e.g. use of different primary DB for the different services)

Service deployment (II version)



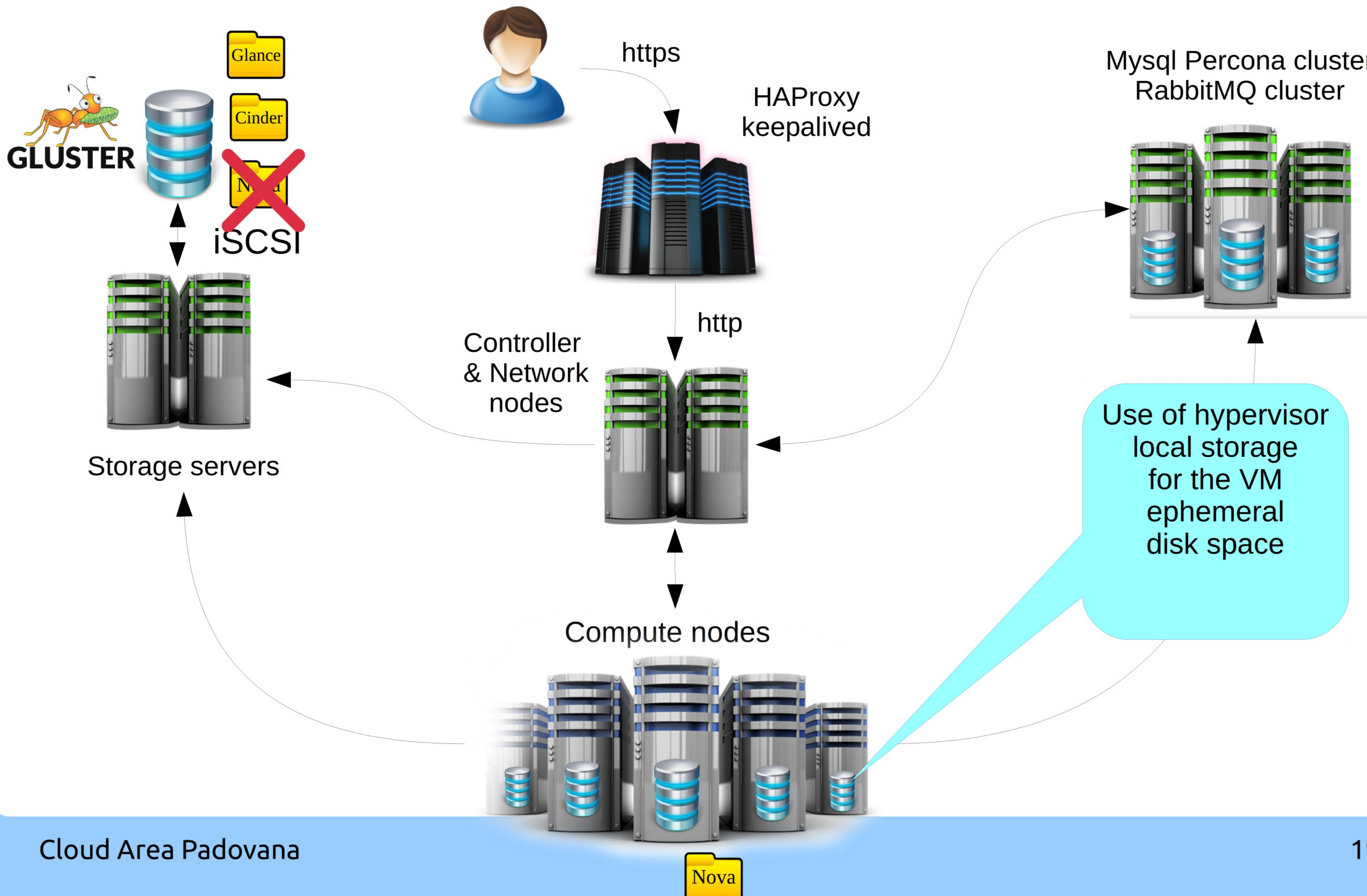
Service deployment (II version)



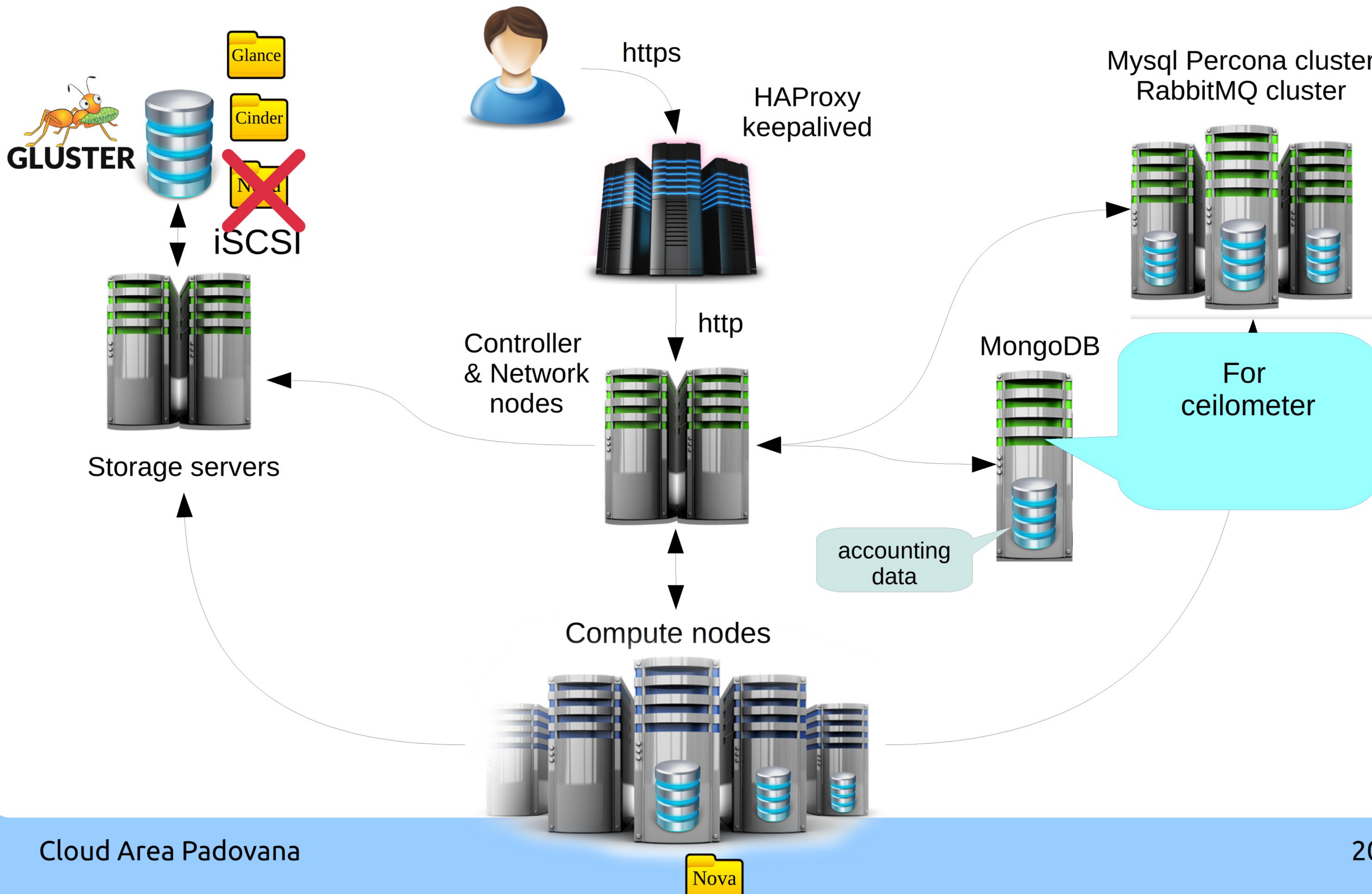
Storage problems

- Very high load on the storage system with some particular workloads
- Impact on the whole infrastucture because of the shared file system
  - Which was chosen to support live migration of instances
- ➡ Decided to use the hypervisor storage for the ephemeral storage of the virtual machines
- For “critical” virtual machines, for which realiability is important, users are suggested to consider “boot from volumes” instances

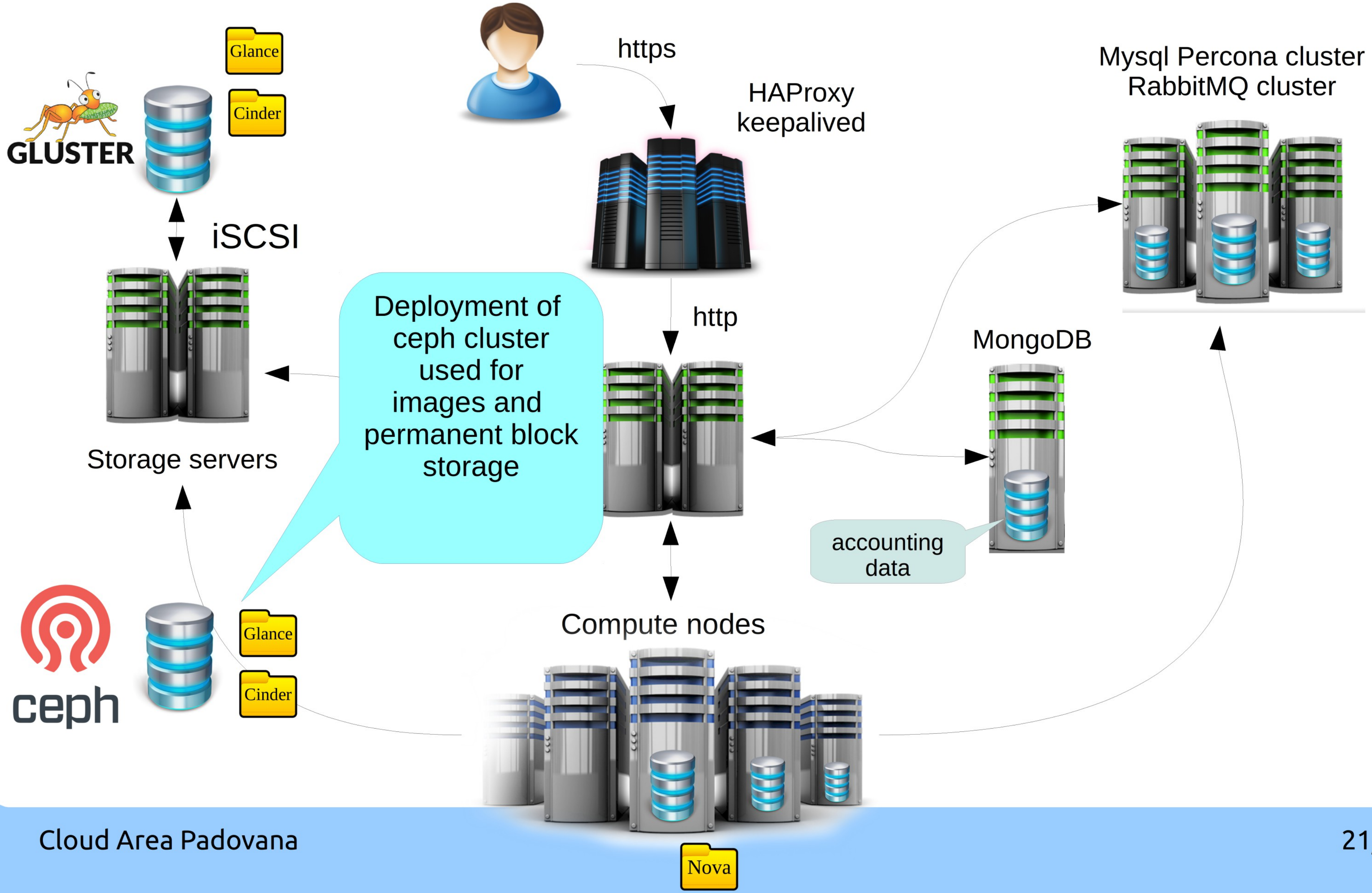
Service deployment (III version)



Service deployment (III version)



Service deployment (IV version)



CEPH cluster

~ 55 TB
3 MON (on 3 hosts)
30 OSD (on 3 hosts)
replica 3

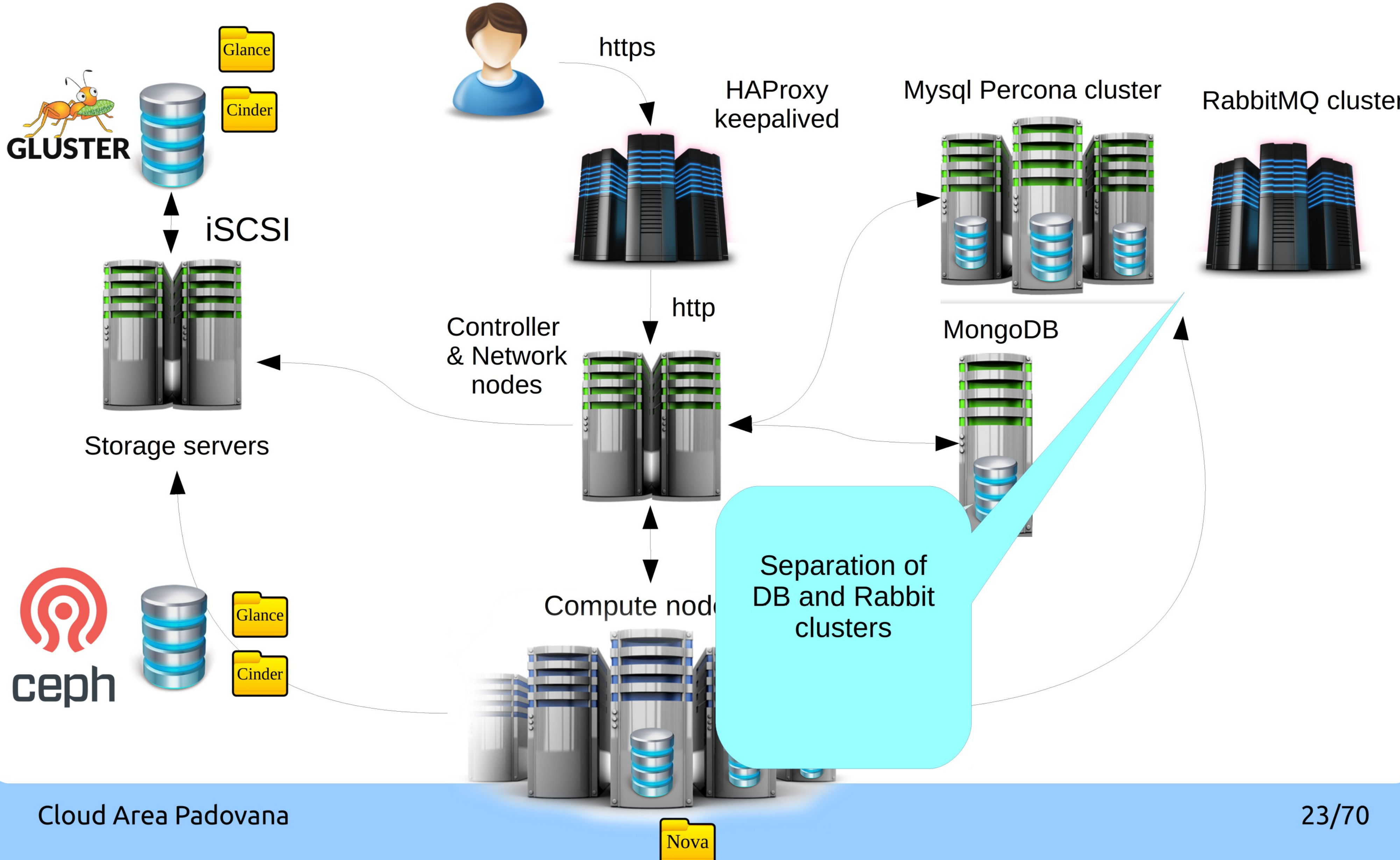
3 MON hosts:
- 2 x E5520 @ 2.27GHz
- 24 GB RAM
- 1 Gbps

3 OSD hosts:
- 2 x E5-2620 v4 @ 2.10GHz
- 64 GB RAM
- 10 x 6 TB SATA (CEPH OSD)
- 2 SSD (CEPH journal)
- 10 Gbps

MON hosts

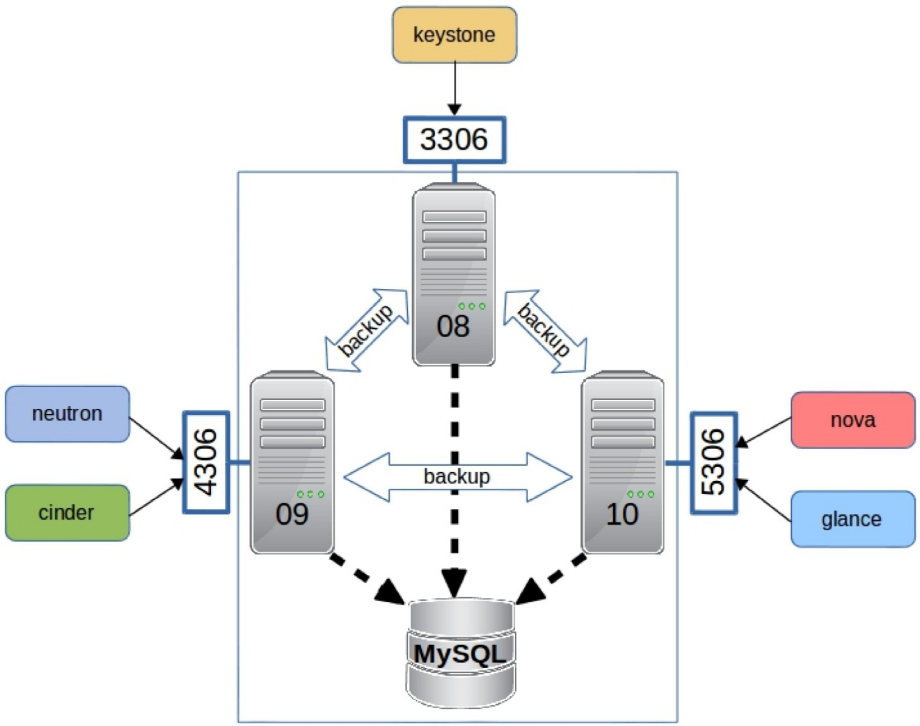
OSD hosts

Service deployment (V version)



MySQL database

- Implemented through a Percona XtraDB cluster (3 instances)
- Access to database through the HAproxy service
- To avoid deadlock issues:
  - configured the cluster in active/backup mode
  - distribute the various OpenStac services between the three different active servers
  - configured each other server as backup



MySQL Database

lazionale
Nucleare

Openstack conf files

```
connection = mysql+pymysql://keystone:<KEYSTONE_DB_PASSWORD>@192.162.30.1:3306/keystone
connection = mysql+pymysql://cinder:<CINDER_DB_PASSWORD>@192.162.30.1:4306/cinder
connection = mysql+pymysql://glance:<GLANCE_DB_PASSWORD>@192.162.30.1:5306/glance
...
```

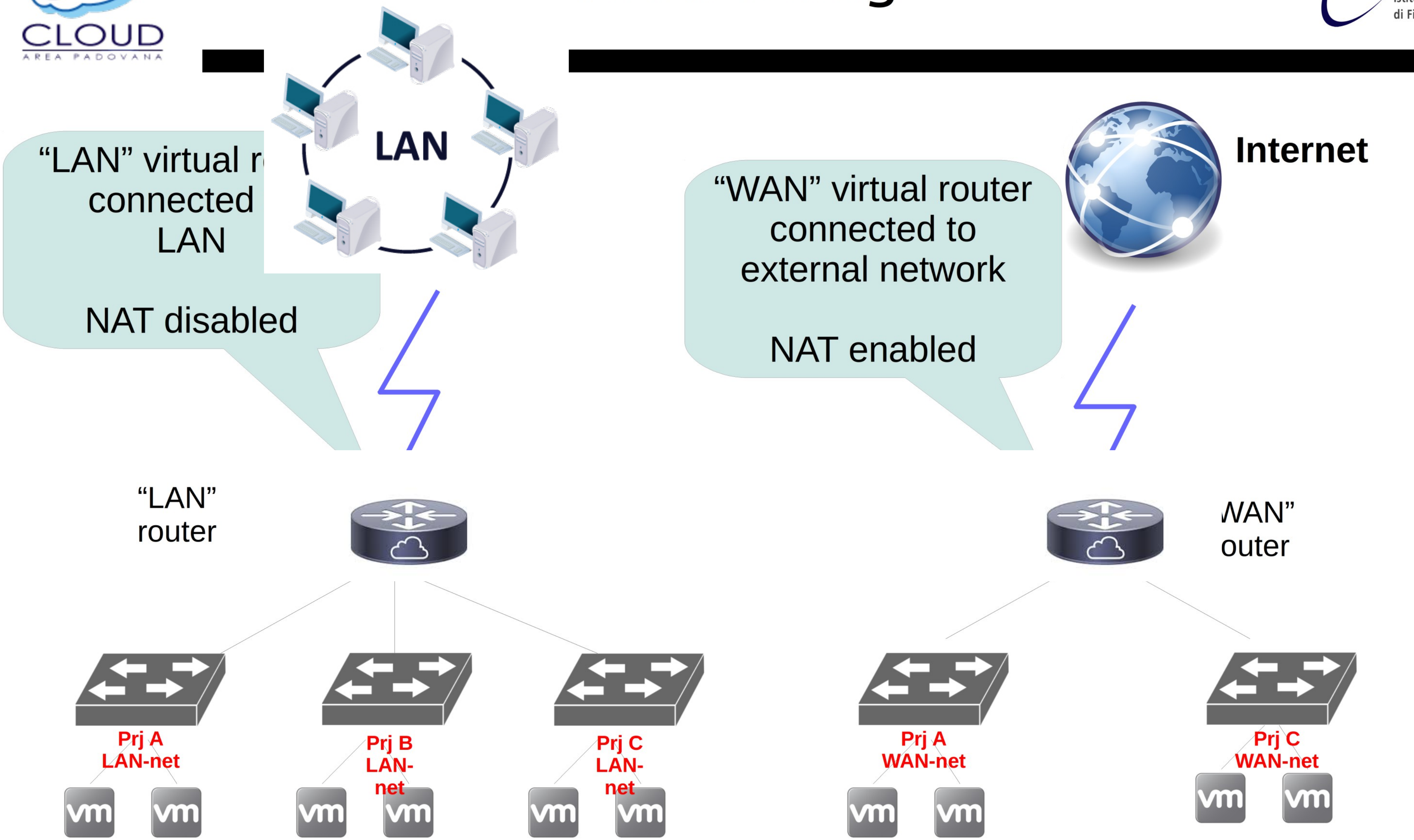
HAproxy conf file

```
...
listen mysql_cluster_one
        bind    192.162.30.1:3306
        mode    tcp
        balance leastconn
        option  httpchk
        default-server on-marked-down shutdown-sessions on-marked-up shutdown-backup-sessions
        server  db-08.cloud.pd.infn.it 192.162.30.14:3306 check port 9200 inter 12000 rise 3 fall 3
        server  db-09.cloud.pd.infn.it 192.162.30.15:3306 check port 9200 inter 12000 rise 3 fall 3 backup
        server  db-10.cloud.pd.infn.it 192.162.30.16:3306 check port 9200 inter 12000 rise 3 fall 3 backup
listen mysql_cluster_two
        bind    192.162.30.1:4306
        mode    tcp
        balance leastconn
        option  httpchk
        default-server on-marked-down shutdown-sessions on-marked-up shutdown-backup-sessions
        server  db-08.cloud.pd.infn.it 192.162.30.14:3306 check port 9200 inter 12000 rise 3 fall 3 backup
        server  db-09.cloud.pd.infn.it 192.162.30.15:3306 check port 9200 inter 12000 rise 3 fall 3
        server  db-10.cloud.pd.infn.it 192.162.30.16:3306 check port 9200 inter 12000 rise 3 fall 3 backup
...
```

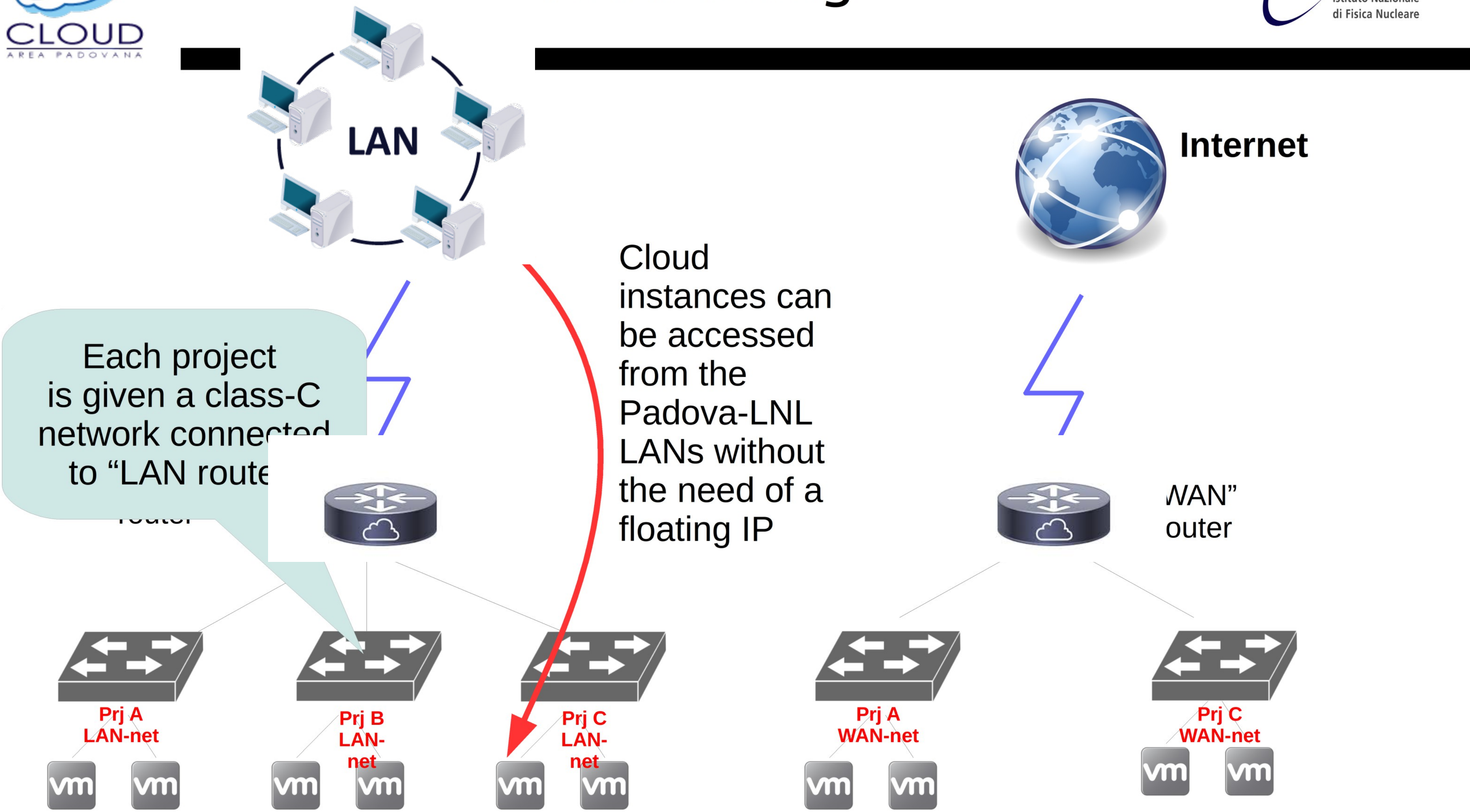
Networking

- Two main requirements
  - Keep Cloud separated from the rest of the Local Area Network
  - Easy access to Cloud instances
    - In particular allow accessing Cloud virtual machines from the Padova-LNL LANs without the need of floating IPs

Networking

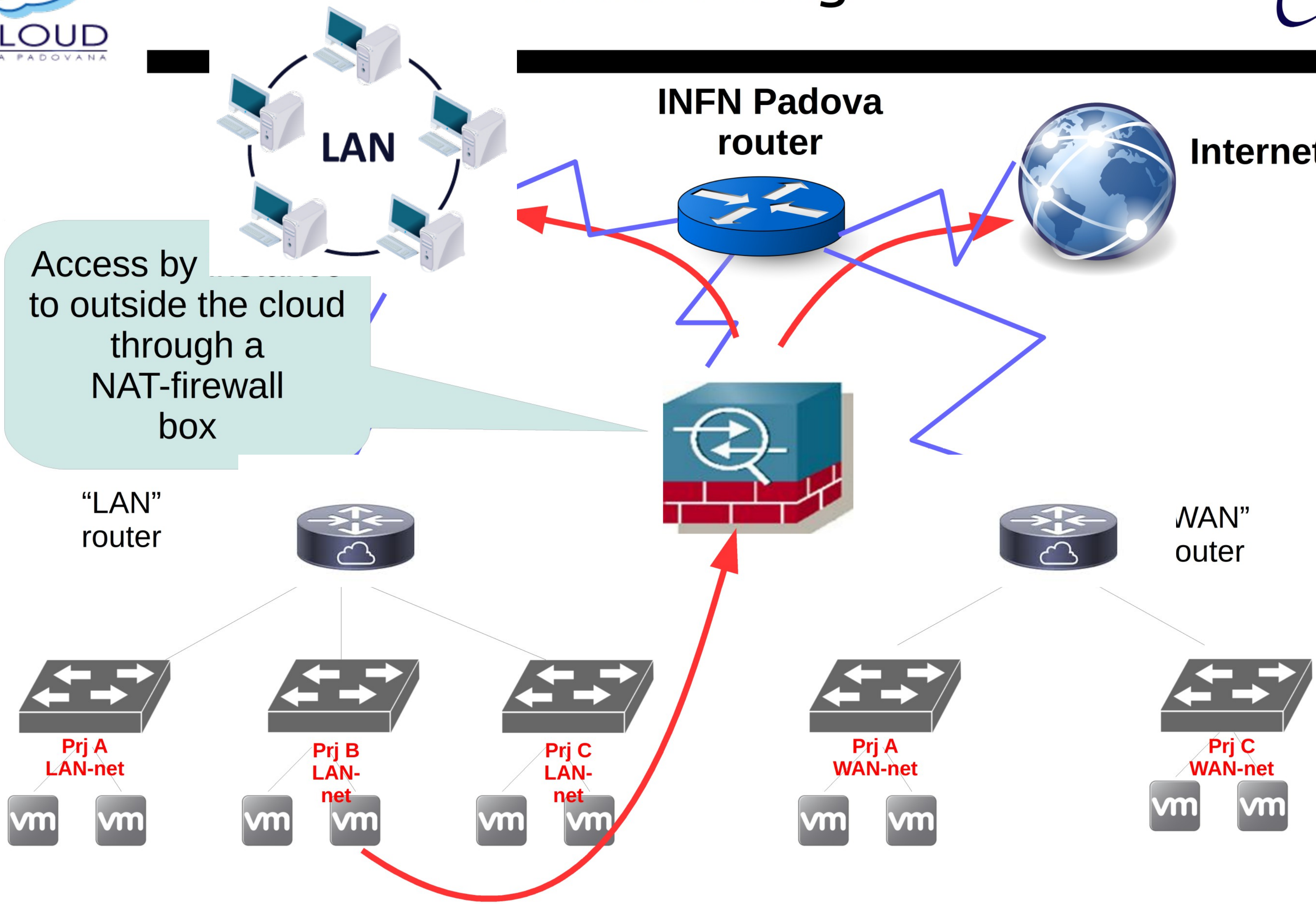

“LAN” virtual r
connected
LAN
NAT disabled
LAN
“WAN” virtual router connected to external network
NAT enabled
Internet
“LAN” router
WAN”
outer
Prj A LAN-net
Prj B LAN-net
Prj C LAN-net
Prj A WAN-net
Prj C WAN-net
vm

Networking

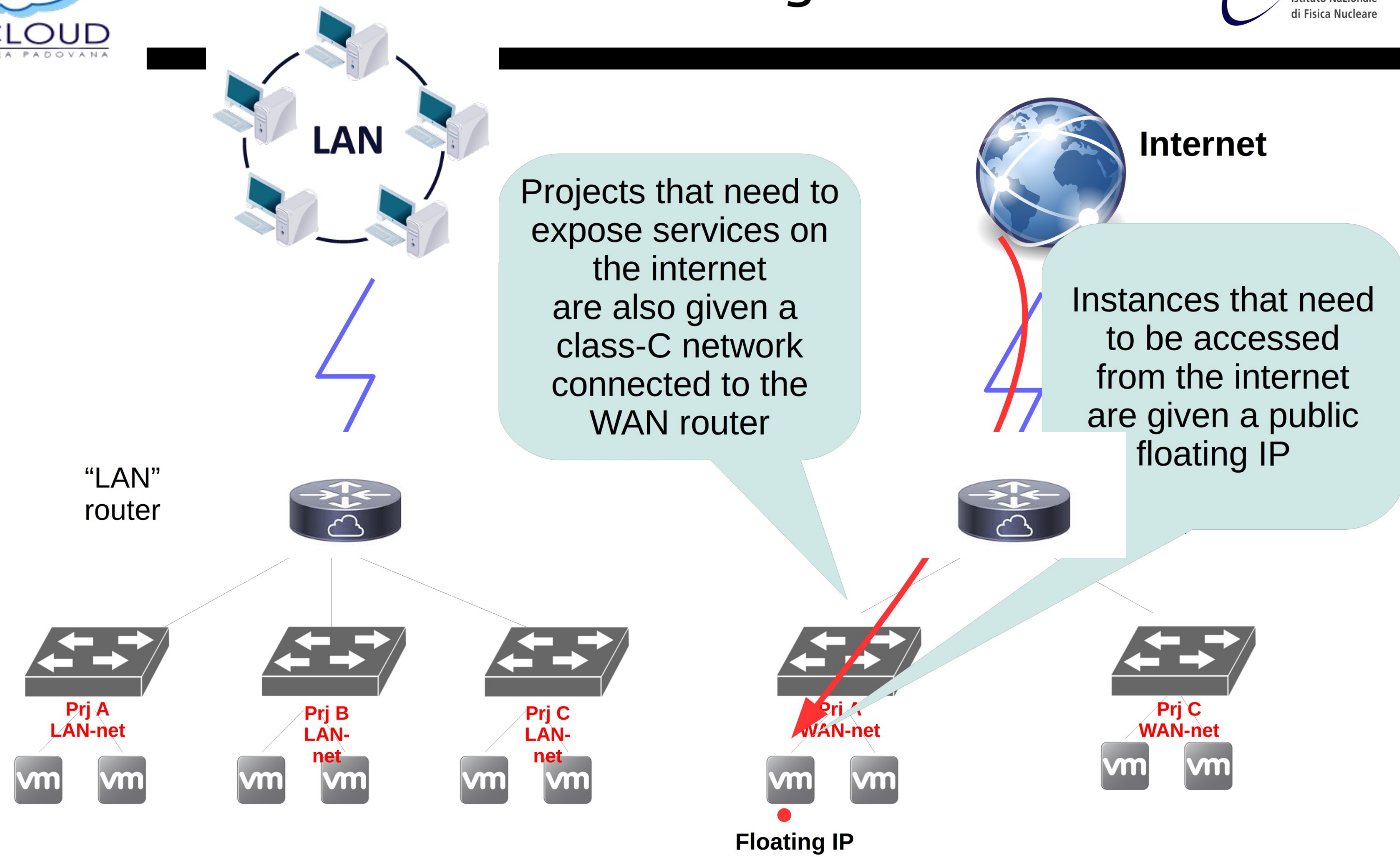


Cloud instances can be accessed from the Padova-LNL LANs without the need of a floating IP

Networking


LAN
INFN Padova router
Internet
Access by
to outside the cloud through a NAT-firewall box
“LAN” router
WAN”
outer
Prj A LAN-net
Prj B LAN-net
Prj C LAN-net
Prj A WAN-net
Prj C WAN-net
vm

Networking


LAN
Internet
Projects that need to expose services on the internet are also given a class-C network connected to the WAN router
Instances that need to be accessed from the internet are given a public floating IP
"LAN" router
Prj A LAN-net
Prj B LAN-net
Prj C LAN-net
Prj A WAN-net
Prj C WAN-net
vm
Floating IP

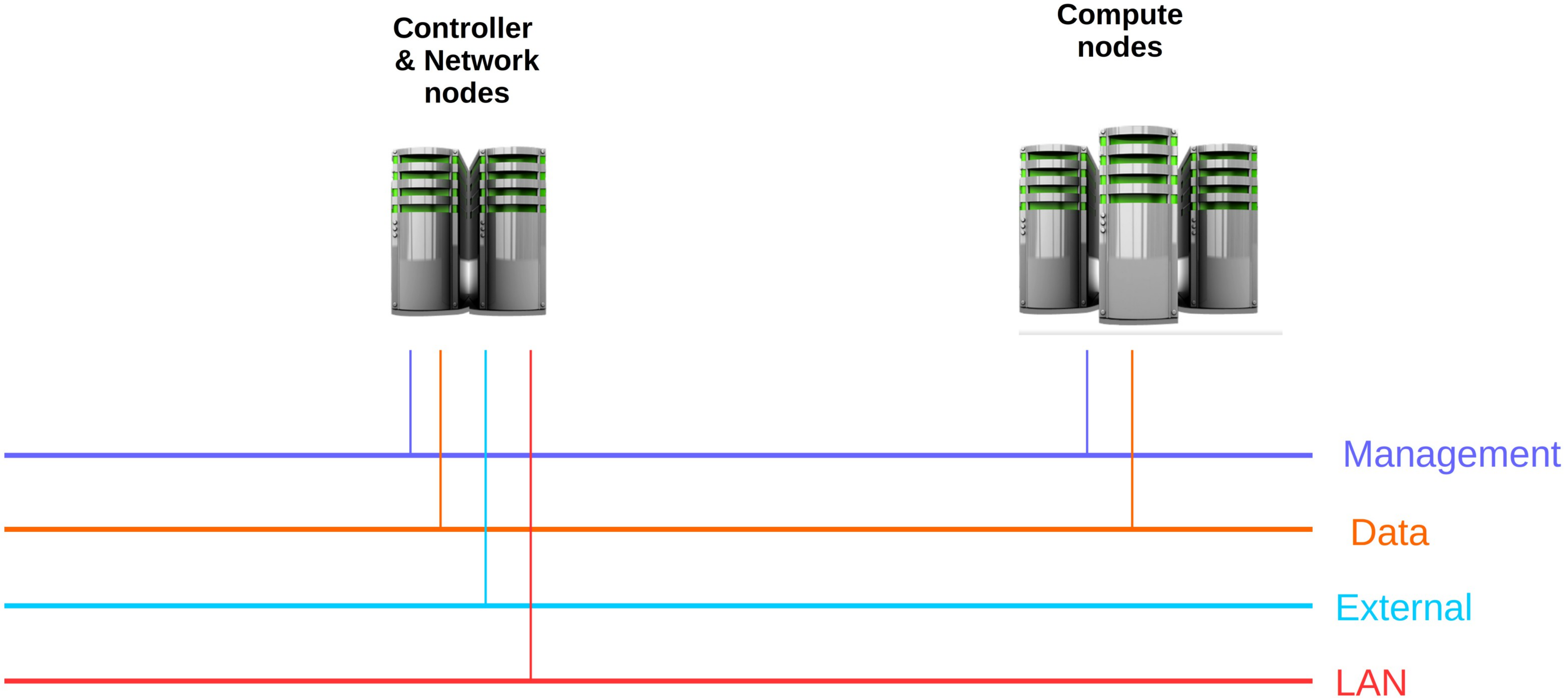
Networking

- Neutron with Open vSwitch/GRE configuration
- Neutron configured with two virtual routers (used by all projects)
  - One with external gateway on public network (“WAN router”)
    - Configured with SNAT enabled
  - One with external gateway on LAN network (“LAN router”)
    - Configured with SNAT disabled
- Each OpenStack project is given
  - A class-C network connected to the “LAN router”
  - A class-C network connected to the “WAN router”, if needed (if the instance has to be exposed on the internet)

Networking

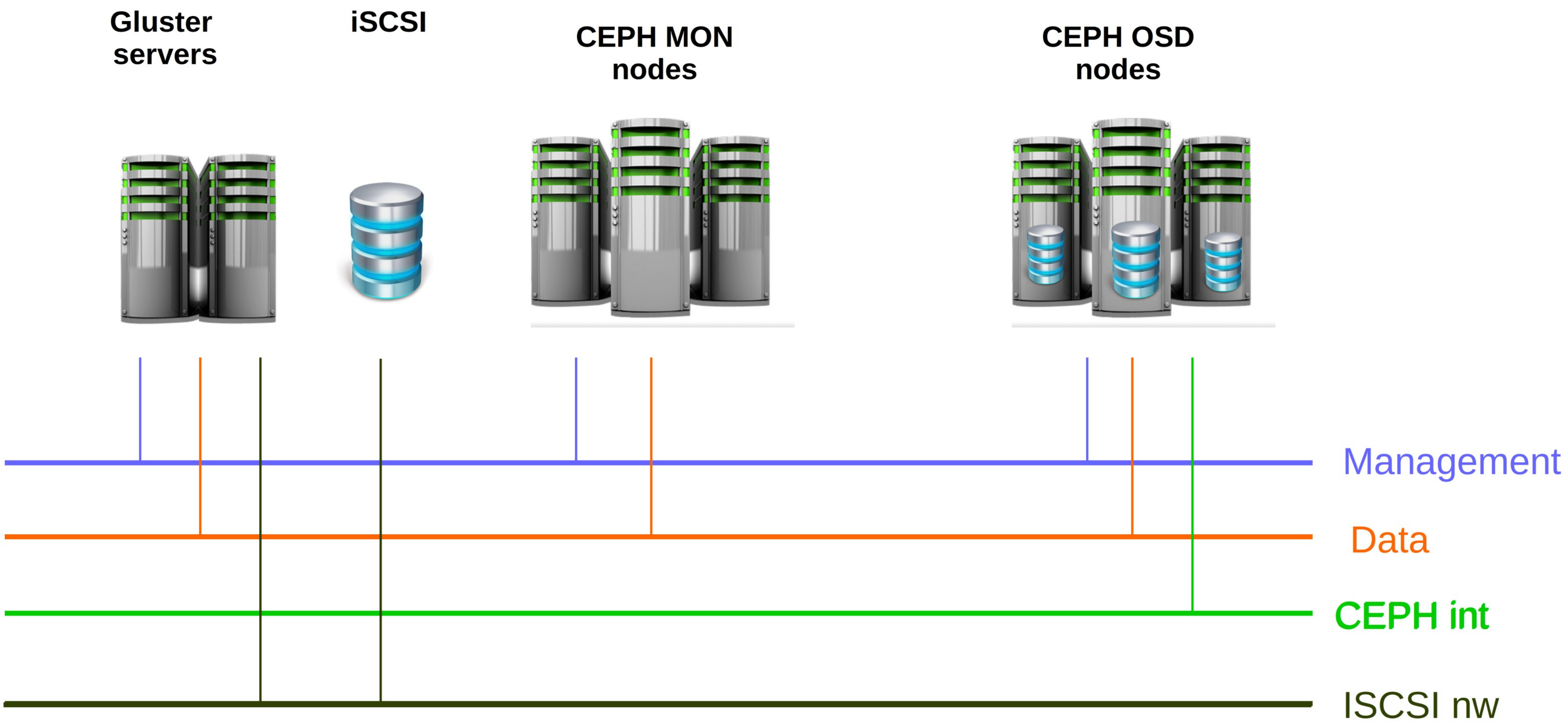
- Each OpenStack project is given
  - a class-C network connected to the “LAN router”
    - This allows accessing the instances from the Padova/LNL LANs (without the need of using floating Ips)
      - From the internet through a gate
    - Instances are identified
      - on the LAN with their private IP
      - on the Internet with the IP of a NAT
  - a class-C network connected to the “WAN router”, if needed
    - These instances can be given a floating (public) IP
    - Used for instances hosting services that have to be exposed on the internet

Networking (OpenStack services)



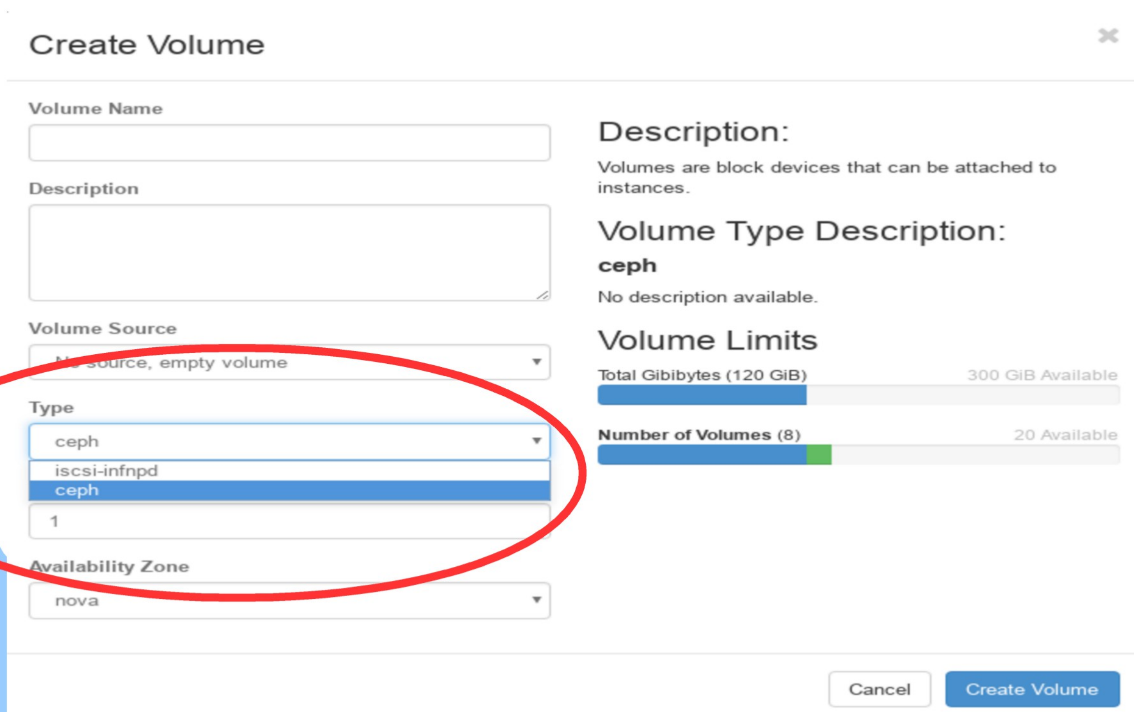
A VLAN per each network
VLANs shared between Padova and Legnaro

Networking (storage services)



Cloud Storage

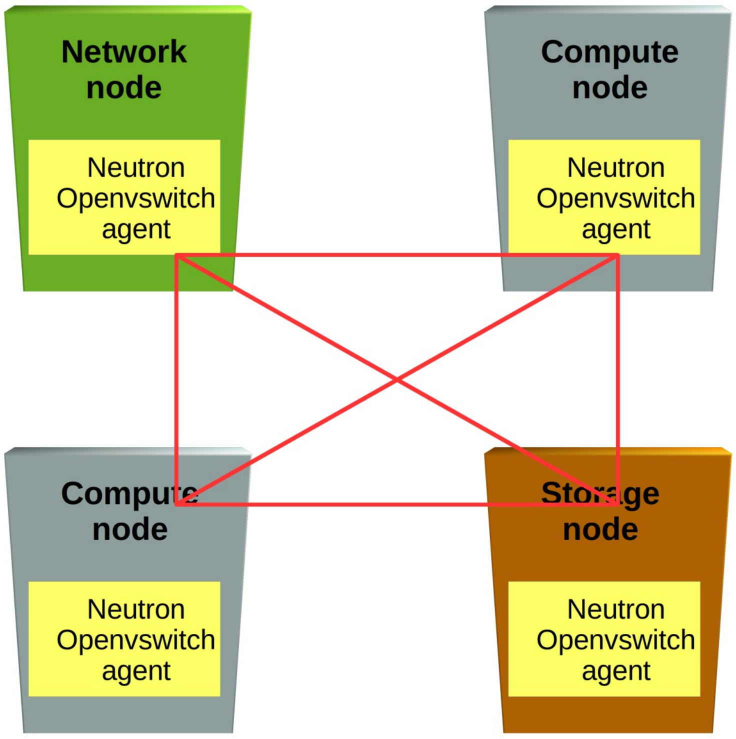
- Hypervisor local disk for the ephemeral block storage
- Cinder (persistent block storage) with two backends: iSCSI and ceph
- For the iSCSI gluster cinder driver (not supported anymore) replaced with the NFS driver
- The two backends are exposed using two volume types
- Different quotas for the different backends




```
# cinder quota-usage b38a0dab349e42bdbb469274b20a91b4
+-----------------------+--------+----------+-------+
|          Type         | In_use | Reserved | Limit |
+-----------------------+--------+----------+-------+
|   backup_gigabytes    |   0    |    0     |  1000 |
|        backups        |   0    |    0     |   10  |
|       gigabytes       |  120   |    0     |  300  |
|     gigabytes_ceph    |   85   |    0     |  100  |
| gigabytes_iscsi-infnpd|   35   |    0     |  200  |
|  per_volume_gigabytes |   0    |    0     |  1000 |
|       snapshots       |   0    |    0     |   10  |
|     snapshots_ceph    |   0    |    0     |   -1  |
| snapshots_iscsi-infnpd|   0    |    0     |   -1  |
|        volumes        |   8    |    0     |   20  |
|      volumes_ceph     |   6    |    0     |   -1  |
|  volumes_iscsi-infnpd |   2    |    0     |   -1  |
+-----------------------+--------+----------+-------+
```

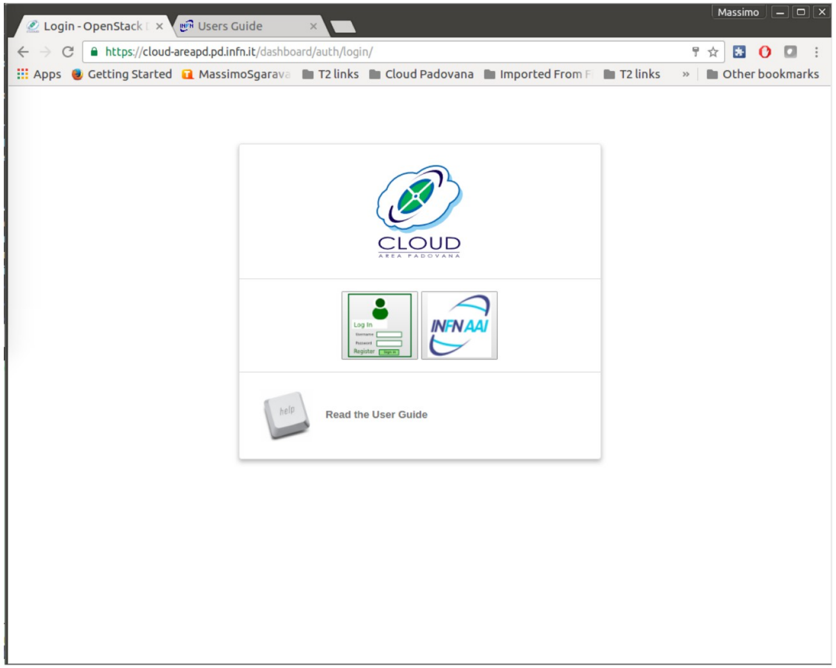
Access to external storage

- Some users need to access to some external (non-cloud) storage
  - E.g Lustre clusters for CMS and LHCb, GlusterFS storage for MuTom
- Used approach: install and configure the Neutron L2 stuff on the storage servers to create GRE tunnels among the Compute nodes and the storage servers
  - Good performance in accessing such external storage
  - Not straightforward to implement and maintain
  - Exploring other solutions (OneData, ...)



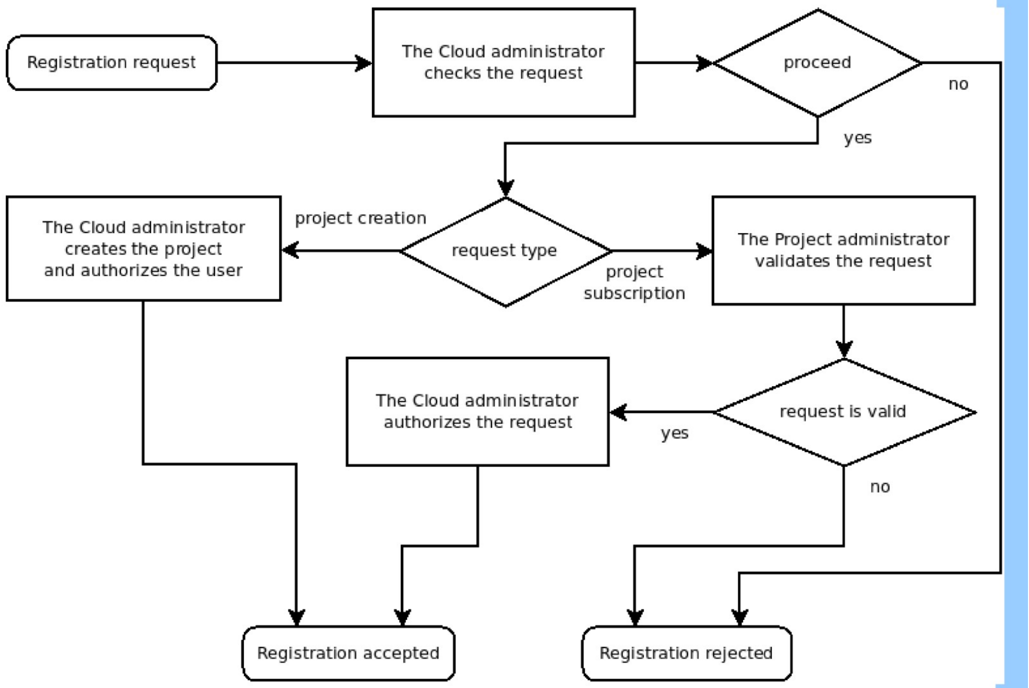
Authentication to the Cloud

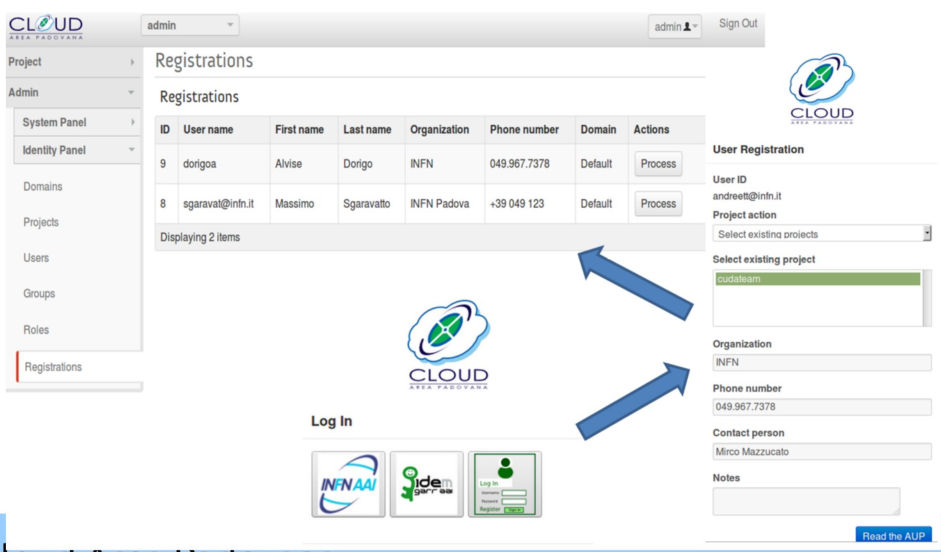
- Users can authenticate to the Cloud using username-password, or using credentials provided by external providers
- Support for SAML and OpenID identity providers
  - Done extending keystone and horizon, since at that time no support was available in OpenStack
  - This implementation has now been revised relying on OS-Federation
- Implemented integration with INFN-AAI and Unipd-SSO (both SAML2)
- Planned to integrate also Indigo IAM (OpenID) and GARR IDEM federation



User and project registration

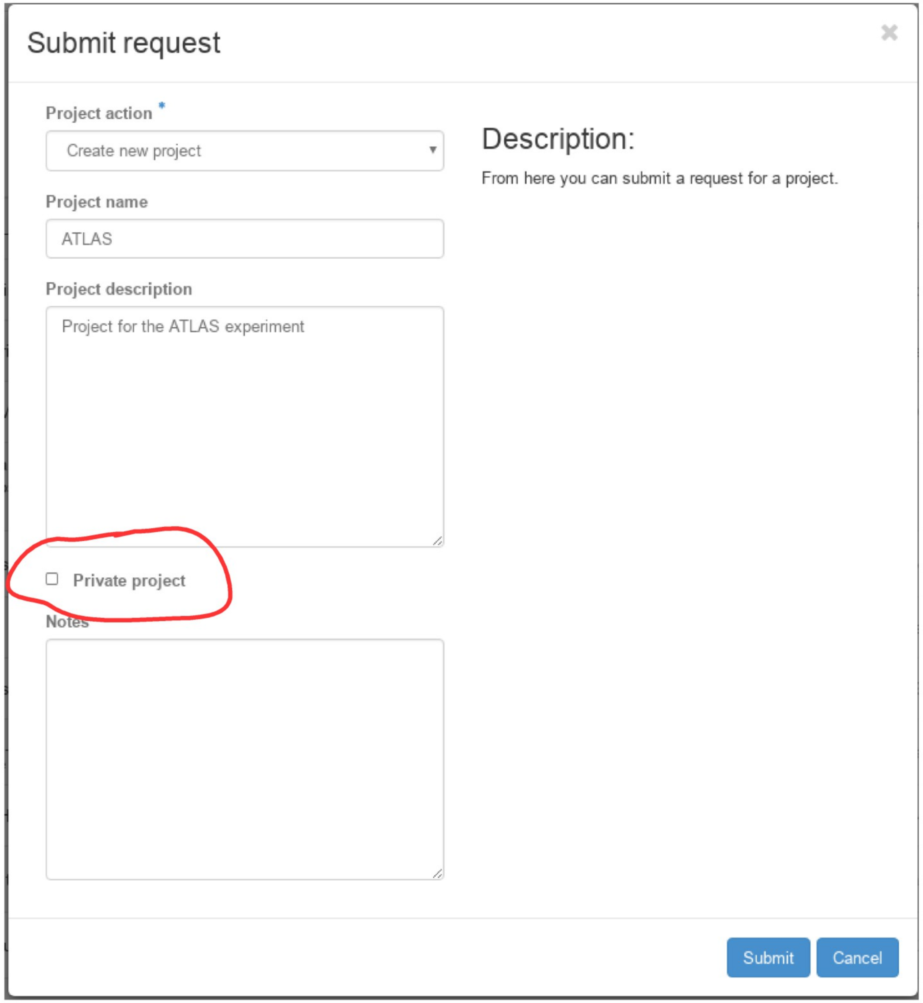
- Complete separation of authentication from authorization
  - A user registration step is needed to grant project membership, access and privileges on the infrastructure
- OpenStack dashboard extended for user and project registrations management
- Requests for registration pass through a workflow (involving the Cloud administrator and the project manager)
- The project manager is responsible to manage (accepting or rejecting) affiliation (and renewals) requests for his/her group
  - Per project account expirations





Projects

- Support for personal (private) and shared (group) projects
- Personal projects are discouraged because they imply a “fragmentation” of resources
- Usually one project per experiment/research group

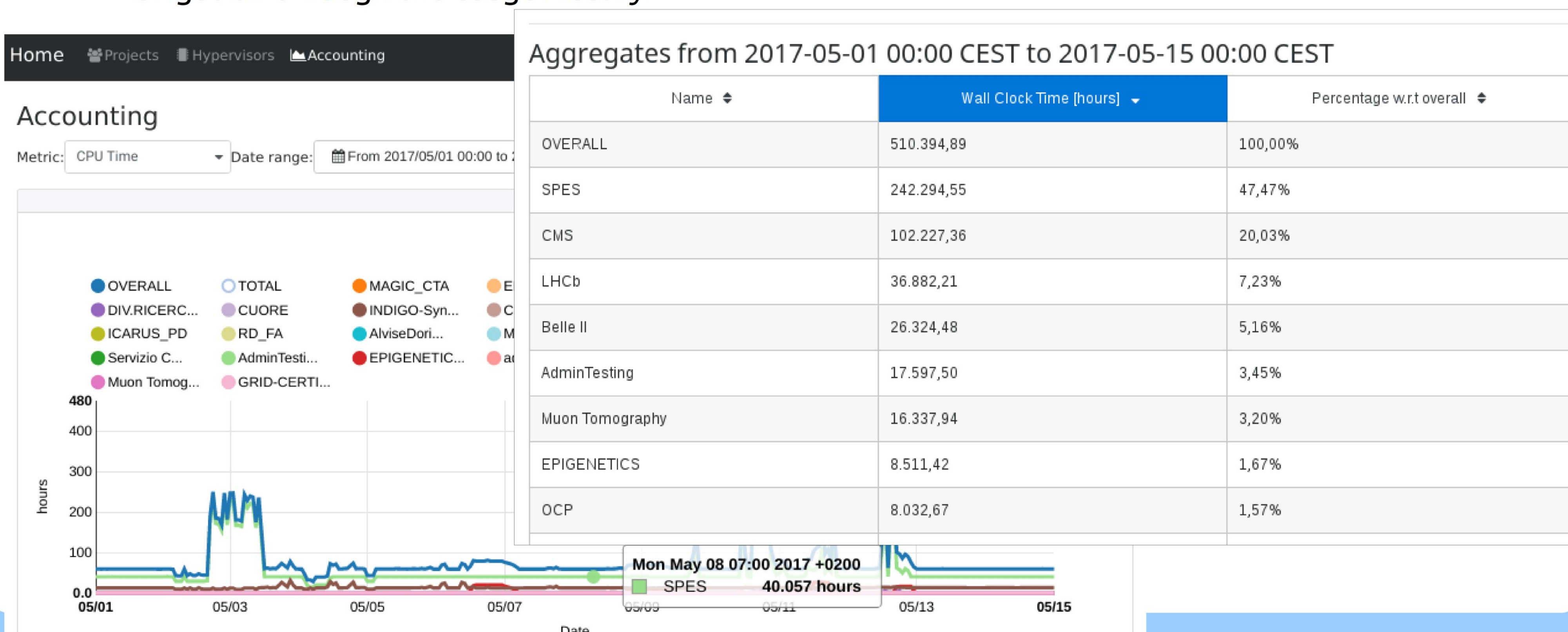


Accounting

- Raw accounting information collected by the OpenStack telemetry service (ceilometer)
- Many problems using Ceilometer also for managing and presenting the accounting data
- ➡ an in-house developed tool (CAOS, Control Application for OpenStack) to:
  - collect and handle accounting information
    - resource usage by projects
    - usages on compute nodes
    - quotas
  - easy management and presentation of accounting data
    - actual usages
    - usages in a specified time range
    - custom metrics not provided by ceilometer

CAOS: accounting

- Comprehensive graphs of accounting data
  - on-the-fly generation
  - cover any of the tracked metrics
  - customizable time range / resolution
  - navigation through the usage history



Aggregates from 2017-05-01 00:00 CEST to 2017-05-15 00:00 CEST

Name	Wall Clock Time [hours]	Percentage w.r.t overall
OVERALL	510.394,89	100,00%
SPES	242.294,55	47,47%
CMS	102.227,36	20,03%
LHCb	36.882,21	7,23%
Belle II	26.324,48	5,16%
AdminTesting	17.597,50	3,45%
Muon Tomography	16.337,94	3,20%
EPIGENETICS	8.511,42	1,67%
OCP	8.032,67	1,57%

CAOS: Resource usage by projects

Home | Projects | Hypervisors | Accounting | Log out

			VCPUs	VMs	RAM
CLOUD-LNL	367f4c8918fc4d658d49be0b79ac9daf	3 minutes ago	Usage: 0,00 Quota: 15,00	Usage: 0 Quota: 15	Usage: 0,00 Quota: 15,36
Belle II	36b1ddb5dab8404dbe7fc359ec95ecf5	3 minutes ago	Usage: 83,00 41,50% Quota: 200,00	Usage: 13 6,50% Quota: 200	Usage: 120,83 59,00% Quota: 204,80
JUNO	37c6f66ffa0047898811f1c3b9d3e2cb	3 minutes ago	Usage: 7,42 12,37% Quota: 60,00	Usage: 7 7,00% Quota: 100	Usage: 15,21 24,75% Quota: 61,44
OCP	3beba6dd3f2648378263bc04d9c205fa	3 minutes ago	Usage: 29,00 29,00% Quota: 100,00	Usage: 14 70,00% Quota: 20	Usage: 59,39 90,63% Quota: 65,54
CALET	3f13d911e5a0448db1ad8363d0d264d5	3 minutes ago	Usage: 4,00 26,67% Quota: 15,00	Usage: 1 6,67% Quota: 15	Usage: 8,19 53,33% Quota: 15,36
ICARUS_PD	4acc5c73693d4b8f909a5271f3b09a53	3 minutes ago	Usage: 7,38 14,76% Quota: 50,00	Usage: 2 4,00% Quota: 50	Usage: 15,11 29,51% Quota: 51,20
AlviseDorigo	529769bcecf64277b7447a3c11088f2a	3 minutes ago	Usage: 22,00 27,50% Quota: 80,00	Usage: 10 50,00% Quota: 20	Usage: 45,06 90,11% Quota: 50,00
Multi-tenant Services	535c148bff524f5ca7ac1cb4dc6ef08b	3 minutes ago	Usage: 9,00 45,00% Quota: 20,00	Usage: 5 25,00% Quota: 20	Usage: 18,43 92,16% Quota: 20,00
T2K	73986e45bd624099bb4f4c5f03e56105	3 minutes ago	Usage: 0,00 Quota: 40,00	Usage: 0 Quota: 40	Usage: 0,00 Quota: 40,96
GERDA	8c8f846533a24f2abd632c7f7a8d5fac	3 minutes ago	Usage: 2,00 13,33% Quota: 15,00	Usage: 1 6,67% Quota: 15	Usage: 4,10 26,67% Quota: 15,36
Servizio Calcolo e Ret	b3280982c25146c8b7dccbde74ba1532	3 minutes ago	Usage: 0,00 Quota: 15,00	Usage: 0 Quota: 15	Usage: 0,00 Quota: 15,36
AdminTesting	b38a0dab349e42bdbb469274b20a91b4	3 minutes ago	Usage: 61,00 30,50% Quota: 200,00	Usage: 23 46,00% Quota: 50	Usage: 92,16 45,00% Quota: 204,80
EPIGENETICS	bdd04490175e4462a4c258b2cfed6ca0	3 minutes ago	Usage: 32,00 100,00% Quota: 32,00	Usage: 2 6,25% Quota: 32	Usage: 32,77 100,00% Quota: 32,77
admin	beaeede3841b47efb6b665a1a667e5b1	3 minutes ago	Usage: 22,15 11,08% Quota: 200,00	Usage: 19 47,50% Quota: 40	Usage: 17,49 28,68% Quota: 60,96
			Usage: 98,00	Usage: 8	Usage: 102,40

CAOS: usage on compute nodes

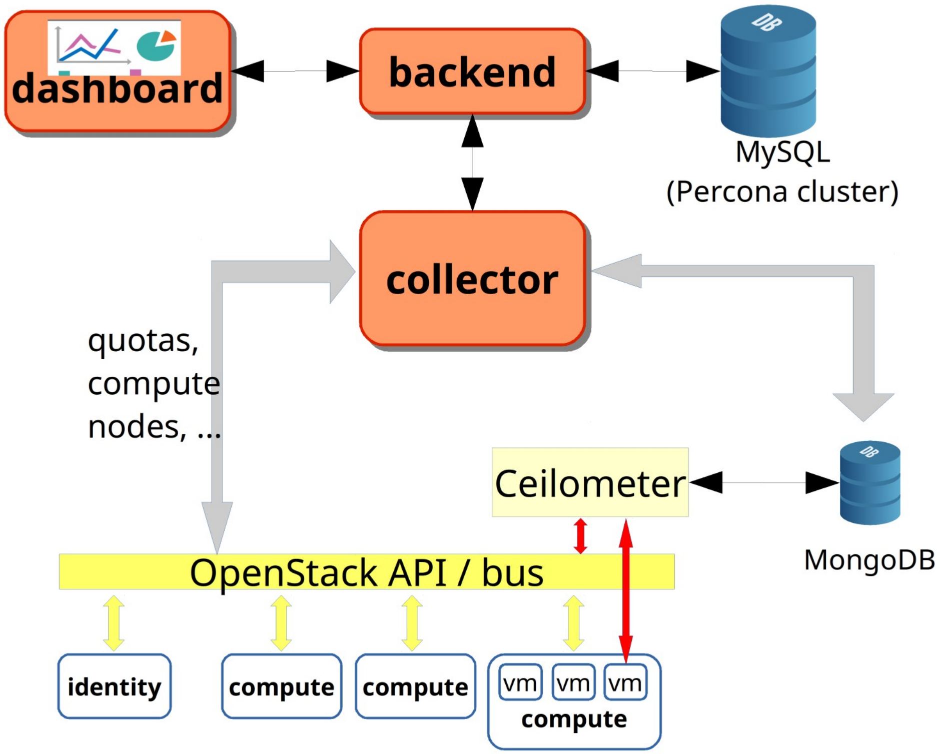
Home | Projects | Hypervisors | Accounting | Log out

Active VMs	Allocated VRAM (TB)		Allocated VCPUs	
	Virtual	On bare	Virtual	On bare
259	Virtual: 3,15 62,34% Virtualizable: 5,05	Virtual: 3,15 88,96% Total Bare: 3,54	Virtual: 1926 46,48% Virtualizable: 4144	Virtual: 1926 179,66% Total Bare: 1072

Hypervisors

Name	Last seen	State / Status	VMs	Allocated VRAM (MB)		Allocated VCPUs	
				Virtual	On bare	Virtual	On bare
cld-ctrl-01.pd.infn.it (IP: 192.168.60.40")	4 minutes ago	down / disabled	0	Virtual: 512 1,06% Virtualizable: 48088.5	Virtual: 512 1,60% Total Bare: 32059	Virtual: 0 0,00% Virtualizable: 32	Virtual: 0 0,00% Total Bare: 8
cld-nl-01.cloud.pd.infn.it (IP: 192.168.60.56")	4 minutes ago	up / enabled	10	Virtual: 78848 54,49% Virtualizable: 144705	Virtual: 78848 81,73% Total Bare: 96470	Virtual: 59 46,09% Virtualizable: 128	Virtual: 59 184,38% Total Bare: 32
cld-nl-02.cloud.pd.infn.it (IP: 192.168.60.57")	4 minutes ago	up / enabled	11	Virtual: 113152 78,19% Virtualizable: 144705	Virtual: 113152 117,29% Total Bare: 96470	Virtual: 87 67,97% Virtualizable: 128	Virtual: 87 271,88% Total Bare: 32
cld-nl-03.cloud.pd.infn.it (IP: 192.168.60.58")	4 minutes ago	up / enabled	10	Virtual: 100864 69,70% Virtualizable: 144705	Virtual: 100864 104,55% Total Bare: 96470	Virtual: 81 63,28% Virtualizable: 128	Virtual: 81 253,13% Total Bare: 32
cld-nl-04.cloud.pd.infn.it (IP: 192.168.60.59")	4 minutes ago	up / enabled	10	Virtual: 107008 73,95% Virtualizable: 144703.5	Virtual: 107008 110,92% Total Bare: 96469	Virtual: 90 70,31% Virtualizable: 128	Virtual: 90 281,25% Total Bare: 32
cld-nl-05.cloud.pd.infn.it (IP: 192.168.60.60")	4 minutes ago	up / enabled	14	Virtual: 112640 77,84% Virtualizable: 144703.5	Virtual: 112640 116,76% Total Bare: 96469	Virtual: 69 53,91% Virtualizable: 128	Virtual: 69 215,63% Total Bare: 32
cld-nl-06.cloud.pd.infn.it (IP: 192.168.60.61")	4 minutes ago	up / enabled	14	Virtual: 82944 57,32% Virtualizable: 144703.5	Virtual: 82944 85,98% Total Bare: 96469	Virtual: 61 47,66% Virtualizable: 128	Virtual: 61 190,63% Total Bare: 32
cld-nl-07.cloud.pd.infn.it (IP: 192.168.60.62")	4 minutes ago	up / enabled	11	Virtual: 104960 54,44% Virtualizable: 192799.5	Virtual: 104960 81,66% Total Bare: 128533	Virtual: 83 64,84% Virtualizable: 128	Virtual: 83 259,38% Total Bare: 32
cld-nl-08.cloud.pd.infn.it (IP: 192.168.60.63")	4 minutes ago	up / enabled	16	Virtual: 121344 62,94% Virtualizable: 192799.5	Virtual: 121344 94,41% Total Bare: 128533	Virtual: 91 71,09% Virtualizable: 128	Virtual: 91 284,38% Total Bare: 32

CAOS architecture



collector:

- gathers data at regular intervals
- workaround to some Ceilometer issues
  - directly query its backend (MongoDB)
  - query the OpenStack API (e.g. for quotas)
  - correction of wrong data (e.g. negative cpu time)
- data pre-aggregation at coarser granularity (hourly, daily)
- computations across different metrics (cpu efficiency)

backend:

- provides a time series framework for writing and reading metrics
- data aggregation and downsampling at given resolution/time range

dashboard:

- frontend to easily get resource usage information
- setting of time ranges/resolution parameters
- display data in graphs and tabular forms

metrics aggregated per project and per compute node

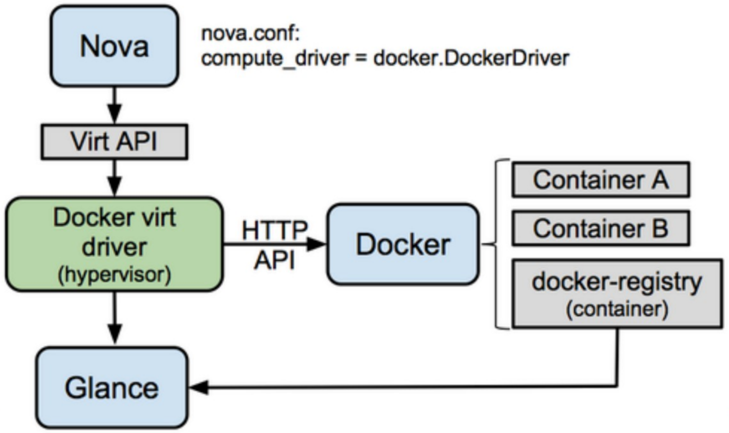
- cpu time, wall clock time, cpu efficiency
- projects quotas and their usages (RAM, DISK, number of VMs, ...)
- current and past usages easy to determine
- fast aggregation over selected time range and granularity

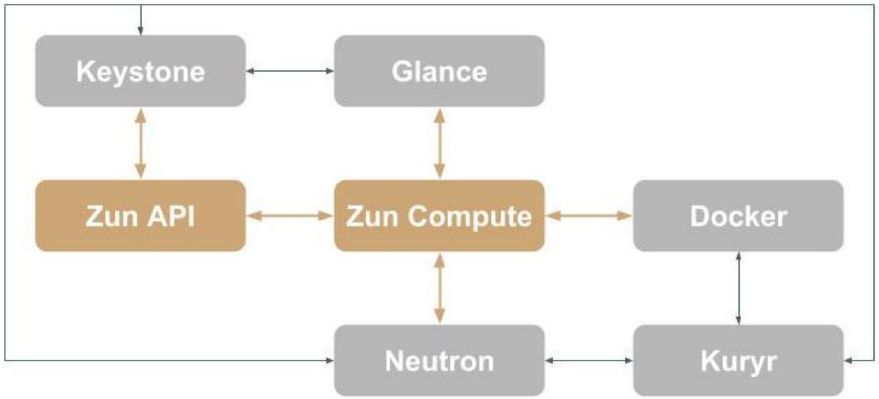
Support for containers

- One compute node instrumented with nova-docker
  - To instantiate docker containers instead of Virtual Machines
  - Docker images in Glance are tagged so that they match the compute nodes running nova-docker

```
> docker save <image-name> | glance image-create –container-format=docker \
--disk-format=raw --name <image-name>
> glance image-update --property hypervisor_type=docker <image-id>
```

- nova-docker not supported anymore by his developer
  - Maintained by Indigo-DataCloud (up to OpenStack Newton)
  - Replaced by ZUN (released in OpenStack Pike for the first time)
    - ZUN evaluation report in Indigo-DataCloud WP4 deliverable





Resource allocation

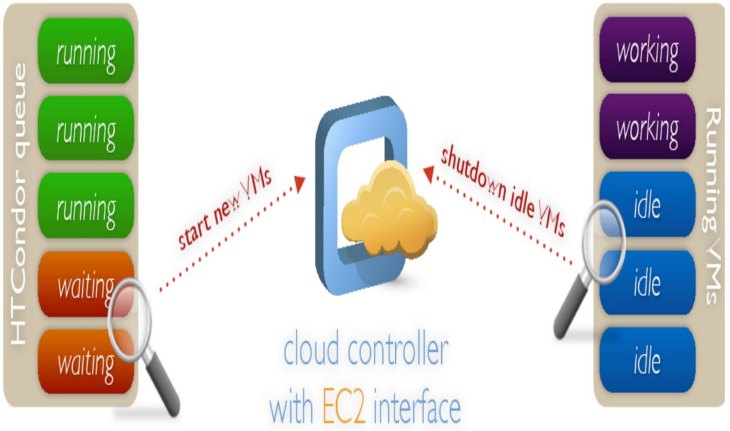
- By default every compute node usable by any project. On these nodes:
  - CPU overcommittment factor: 4.0
  - RAM overcommitment factor: 1.5
- Single pool of resources exposed to users (i.e. no availability zones, etc.)
  - We used to expose (through an availability zone) the compute nodes with shared file system, but this resulted in a bad resource usage

Resource allocation (cont.ed)

- Hardware owners can request specific policies for their hypervisors
- E.g. the MuTom group doesn't want to use overcommitment for its hypervisor
- ➡ Implemented through host aggregates
- We discourage the use of “big” VMs
  - Largest public flavour has 8 VCPU - 16GB
  - Larger flavours created only if really needed and visible only to the relevant project(s)
- Flavors have max 25GB of ephemeral disk
  - To limit snapshot size
  - To manage the limited disk space on the hypervisors

Batch cluster on the Cloud

- We provide users with a simple tool which allows the instantiation of a dynamic HTCondor batch cluster
- Based on the elastiq software
  - User specifies the image to be used, the flavor, the min and max number of slots
  - New VMs are automatically created (via EC2) and configured as HTCondor worker nodes when there are jobs waiting in the queue
  - VMs are automatically destroyed when idle (i.e. queue empty)



Some policies

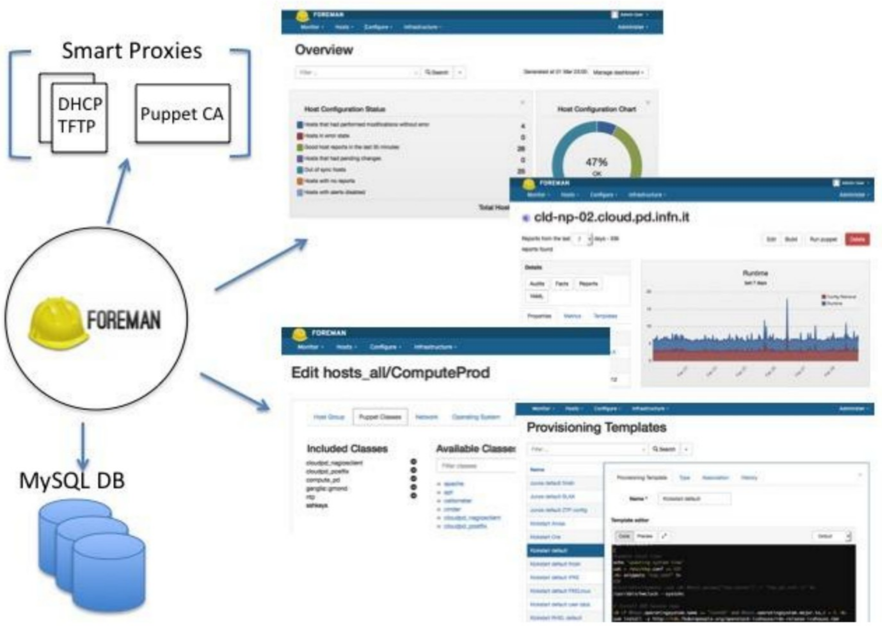
- Modified the default OpenStack behavior so that users can't stop/delete VMs belonging to other users of the same project
  - ➡ Customization of Nova policy.json
  - ➡ Use of Nova 2.0 since in Mitaka not supported in Nova 2.1
    - Support (temporarily ?) reintroduced in nova 2.1 in Ocata
- Users can not delete volumes created by other users (even if they belong to the same project)
  - ➡ Customization of Cinder policy.json

Images

- Users can create and publish their own images
  - Max 25 GB size
  - Can not be public, but can be shared with other projects
- We (admins) provide and maintain a few public images
  - SL6 / CentOS 7 with mostly used packages + CVMFS (configured to use the proper squids) + LDAP setting (for user authentication)
  - Sysadmins accounts enabled
  - Users can ask sysadmin support only for instances built using these images
- We do backup on images and snapshots

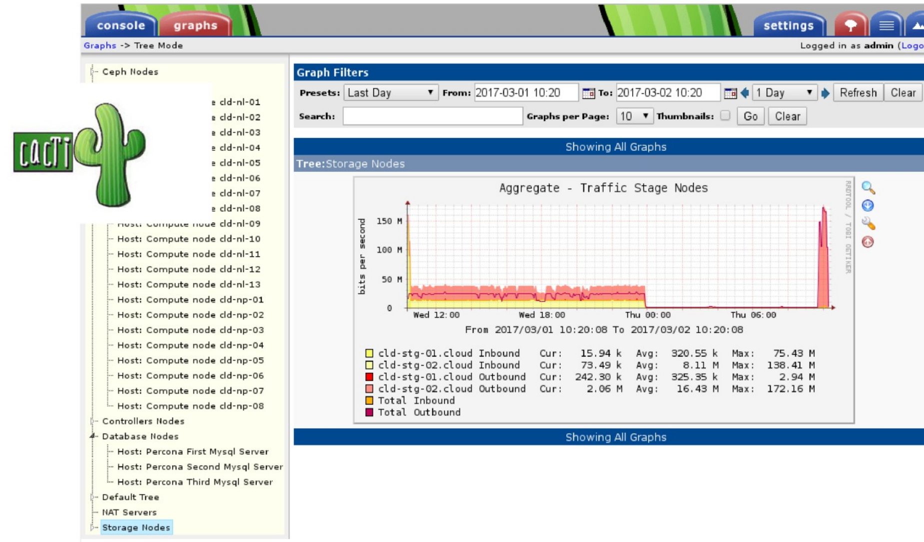
Provisioning and configuration

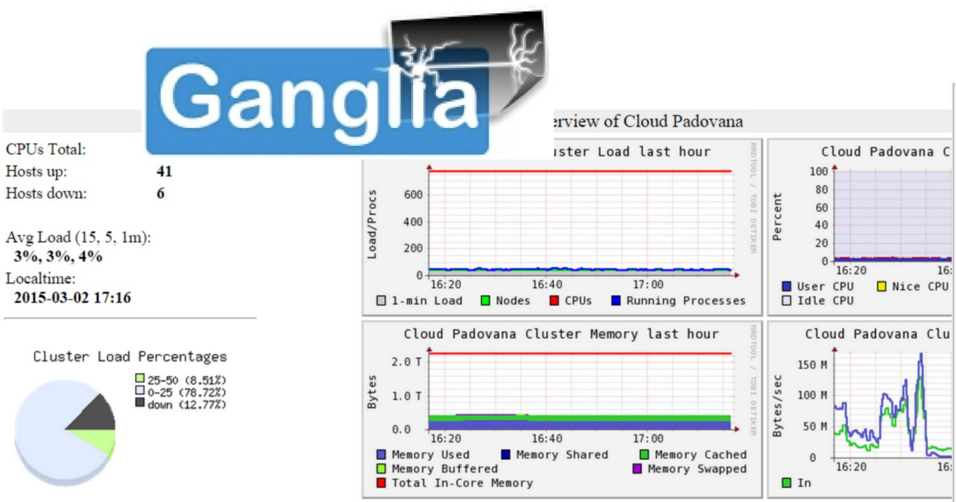
- Foreman/Puppet used as provisioning and configuration systems
- Puppet used to configu OpenStack and the ancillary services (nagic ganglia, etc)
- We implement most of the used puppet classes

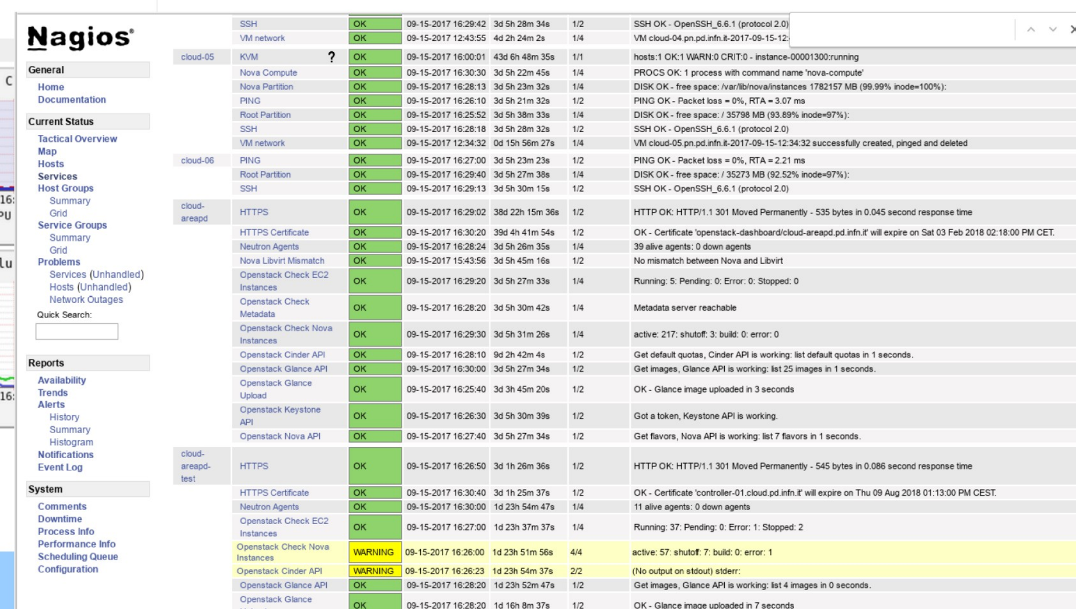


Monitoring

Based on the same tools used in the Padova-Legnaro Tier-2: Ganglia, Nagios, Cacti







Auditing

Monitor of floating IP allocations

```
[root@cld-ctrl-01 ~]# . admin-openrc.sh
[root@cld-ctrl-01 ~]# os-ip-trace 90.147.77.136
+--------------------------------------+------------------+----------------------+---------------------+
|              device id               |    user name     |   associating date   |
+--------------------------------------+------------------+----------------------
| 478cd520-d44e-4936-bb59-fd0211dc2ac4 | sgaravat@infn.it | 2016-12-16 09:50:07
| b1d76ac2-fec9-45e0-8795-dfcf6e6eff79 | sgaravat@infn.it | 2016-12-16 09:52:11
+--------------------------------------+------------------+----------------------
```

Use of ulogd to monitor Cloud instances network activities

```
grep 10.63.15.25 /var/log/ulogd.log
Nov 29 11:59:15 cld-ctrl-01 ulogd[11498]: [DESTROY] ORIG: SRC=10.63.15.25 DST=64.90.42.85 PROTO=TCP SPT=42455
DPT=80 PKTS=0 BYTES=0 , REPLY: SRC=64.90.42.85 DST=90.147.77.147 PROTO=TCP SPT=80 DPT=42455 PKTS=0 BYTES=0
Nov 29 12:01:04 cld-ctrl-01 ulogd[11498]: [DESTROY] ORIG: SRC=10.63.15.25 DST=169.254.169.254 PROTO=TCP SPT=38
693 DPT=80 PKTS=0 BYTES=0 , REPLY: SRC=10.63.15.1 DST=10.63.15.25 PROTO=TCP SPT=9697 DPT=38693 PKTS=0 BYTES=0
Nov 29 12:01:04 cld-ctrl-01 ulogd[11498]: [DESTROY] ORIG: SRC=10.63.15.25 DST=169.254.169.254 PROTO=TCP SPT=38
694 DPT=80 PKTS=0 BYTES=0 , REPLY: SRC=10.63.15.1 DST=10.63.15.25 PROTO=TCP SPT=9697 DPT=38694 PKTS=0 BYTES=0
Nov 29 12:01:04 cld-ctrl-01 ulogd[11498]: [DESTROY] ORIG: SRC=10.63.15.25 DST=169.254.169.254 PROTO=TCP SPT=38
695 DPT=80 PKTS=0 BYTES=0 , REPLY: SRC=10.63.15.1 DST=10.63.15.25 PROTO=TCP SPT=9697 DPT=38695 PKTS=0 BYTES=0
...
```

Centralization and archival of all relevant log files

Archival (on shadow tables) of records concerning deleted instances

Updates

- We perform one OpenStack update per year (i.e. skipping one OpenStack release)
  - as balance between having latest features and fixes and the need of limiting the manpower
- Every change (in particular the updates) are prepared and tested on a testbed environment before being deployed in production
  - small infrastucture which however simulates the production one (e.g. services deployed in High Availability)
- We are now running OpenStack Mitaka but almost ready with the update to Ocata

Other clouds @ Padova

- INFN-PADOVA-STACK instance of the EGI-FedCloud
- CloudVeneto.it

Many services shared (monitoring, foreman/puppet,...) shared among all Cloud instances

CloudVeneto.it

- Partners: UniPd (10 depts), INFN-Padova, INFN-Legnaro
- 480 core (in HT), 90 TB
- ~ 40 projects, ~ 90 users
- Preparing integration with Cloud Area Padovana
  - Single OpenStack Cloud exposed with 2 “entry points” (Unipd and INFN)

EGI FedCloud

- 144 slot, ~ 1300 HS06
- Keep separated with Cloud Area Padovana to avoid problems integrating EGI specific services
  - E.g. keystone-voms
- Keystone-voms planned to be replaced with keystone v3 federation + gridsite
  - ➡ integration with Cloud Area Padovana can then be done

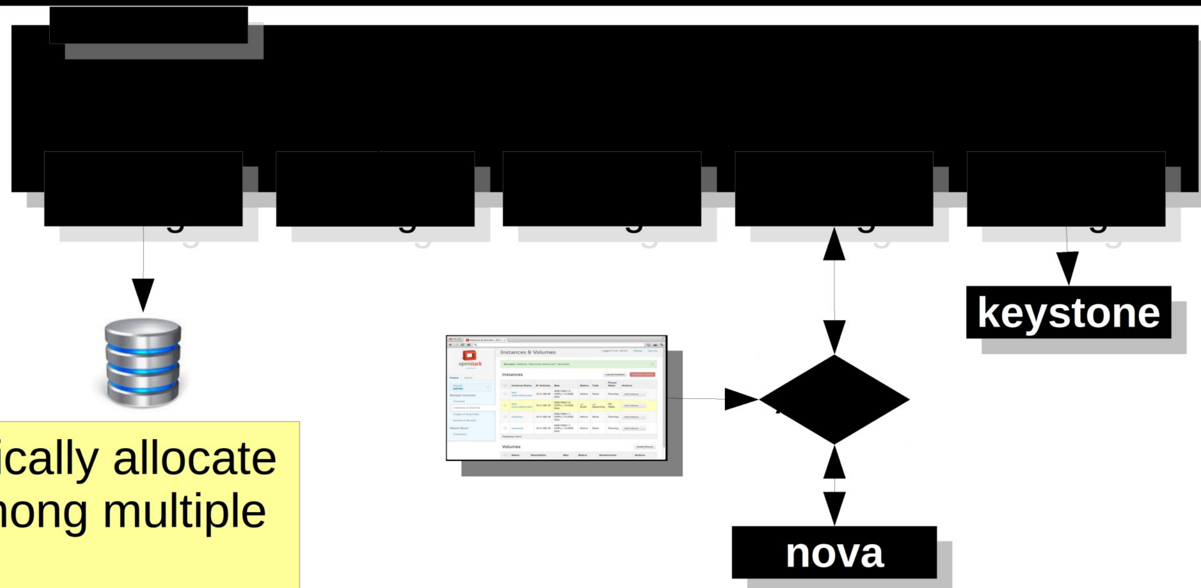
THE END

Questions?

Backup slides

Synergy

[redacted]

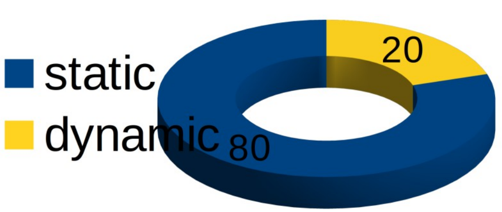


Possibility to dynamically allocate part of resources among multiple projects (fair-share)

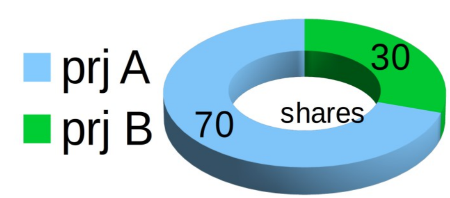
Instances are killed after a certain time (to enforce fair share)

Requests for new instances are queued if no resources are available

total resources



dynamic resources

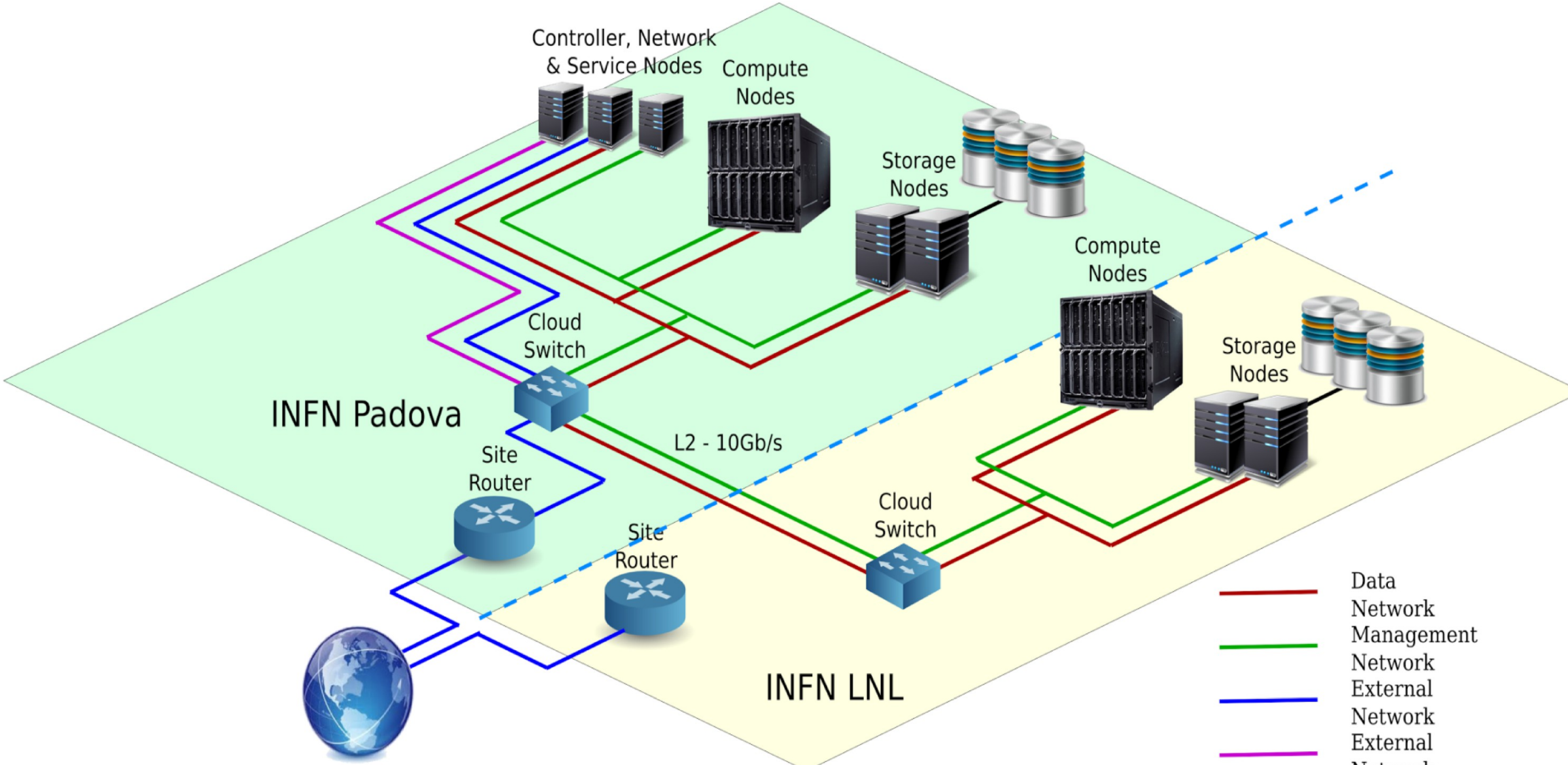


Synergy integration

- Tried integration between Synergy and batch applications using elastiq
- Services (synergy, elastiq, HTCondor) configurations need to be properly set to reduce
  - the number of jobs killed because VMs deleted by Synergy
  - the number of instances (in scheduling state) killed by elastiq because they didn't join in time the HTCondor cluster

Networking

LOGO



Some use cases: CMT

CMT: Cosmic Muon Tomography

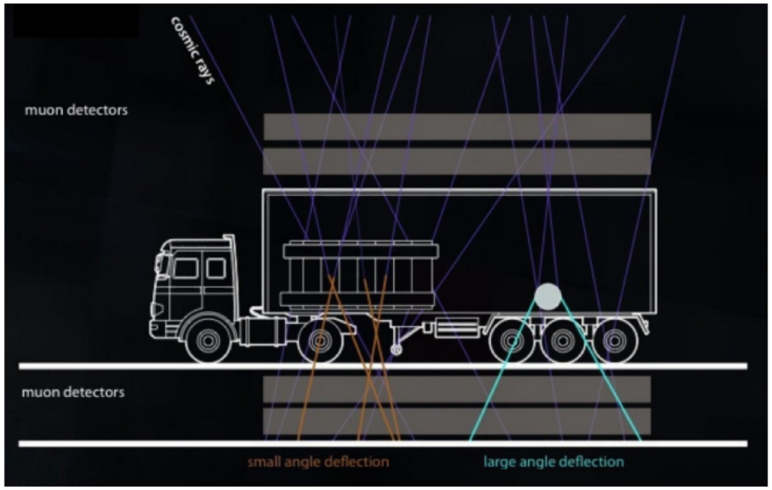
Imaging system to scan inaccessible volumes of materials with high atomic number…

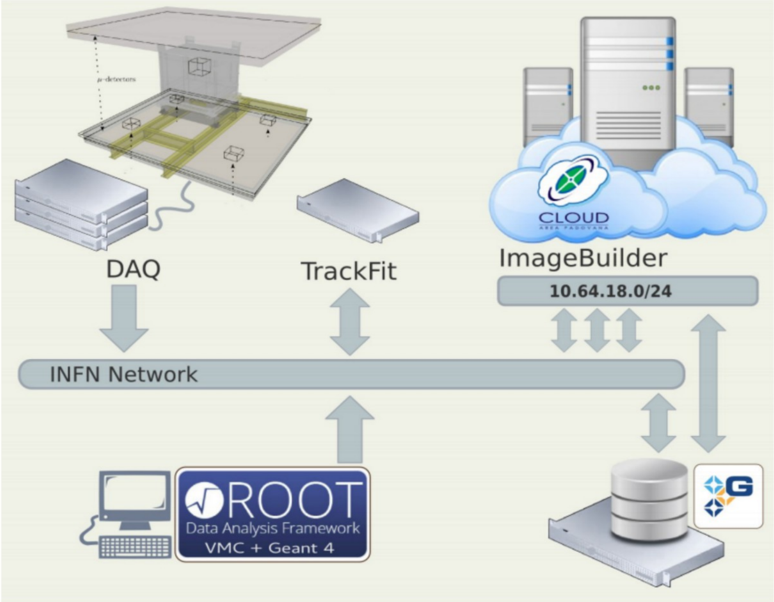
… by studying the scattering of muons after matter collisions and applying statistical and iterative algorithms

… to find the optimal density of the target material that fits the muon trajectories and then its volume and shape

The algorithms for data analysis require a huge computing power

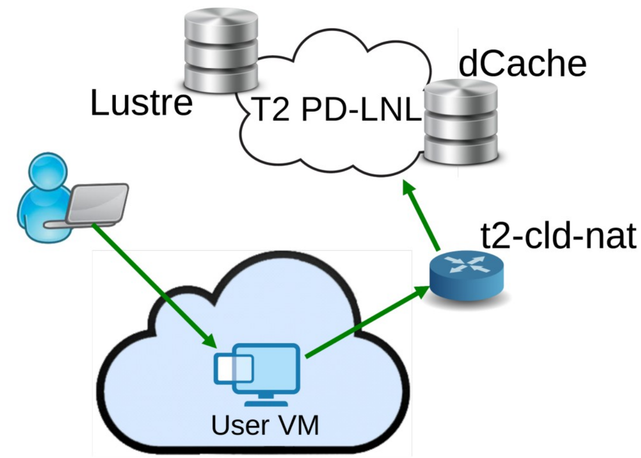
Using ImageBuilder as reconstruction software

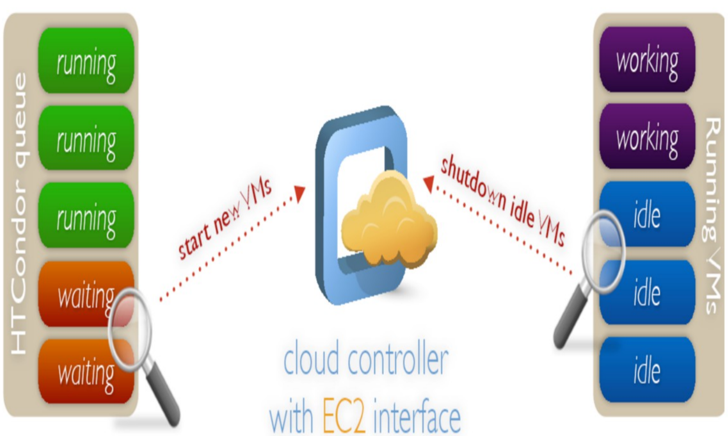




Some use cases: CMS

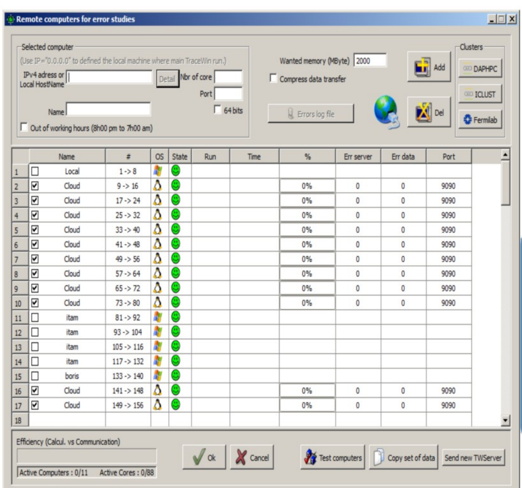
- Interactive usage
  - code development and build, ntuple productions, end-user analysis, Grid User Interface
    - each user instantiates his own VM and destroys it when not needed anymore
  - Cloud is integrated with the local Tier-2
    - access to dCache and Lustre storage
- Batch usage
  - elastic batch cluster (HTCondor) managed by elastiq
    - new VMs are created when there are jobs waiting in the queue
    - VMs are destroyed when idle (i.e. queue empty)

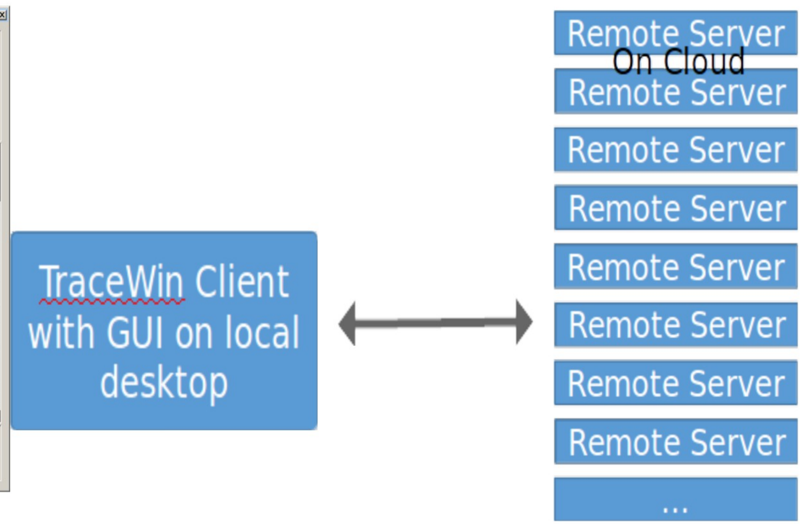




Some use cases: SPES

- Simulations for the SPES experiment
  - to tune the “perturbated” parameters of the accelerator
  - necessary to have many simulations in a short time
- This is done on the client using a client-server framework called TraceWin





Issues (1/n)

- inefficient storage of information
  - each measurement is accompanied by redundant metadata
- trade-off between history retention and storage
  - ~150-200 GB in 3 months, depending on cloud activity
  - might be solved in the future using gnocchi db


```
{
  "_id": "577e73252960a744a92f6c8f",
  "counter_name": "cpu",
  "counter_volume": "26910000000",
  "timestamp": "2016-07-07T15:20:05Z",
  "project_id": "b38a0dab349e42bdbb469274b20a91b4",
  "user_id": "5213bbc53b0f45e3ae9ec945970cc230",
  "resource_id": "87b1c046-9393-4974-814a-d2471a028714",
  "recorded_at": "2016-07-07T15:20:05.654Z",
  "message_signature":
"fcce5b42dc52b8f23600cce99b2dba6c9bb7d438687969696a2c93b42997f1c3",
  "message_id": "488a05be-4456-11e6-b25f-44a84225d06f",
  "source": "openstack",
  "counter_unit": "ns",
  "resource_metadata": {
    "status": "active",
    "cpu_number": 1,
    "ephemeral_gb": 0,
    "display_name": "segatta-test1",
    "name": "instance-000749a3",
    "disk_gb": 20,
    "kernel_id": null,
    "image": {
      "id": "d67f6702-9c70-4972-9bc5-2adea2e2ca85",
      "links": [{
          "href":
"http://192.168.60.40:8774/1b2caeedb3e2497b935723dc6e142ec9/images/d67f6702-9c70-
4972-9bc5-2adea2e2ca85",
          "rel": "bookmark"
        }],
      "name": "CentOS7"
    },
    "ramdisk_id": null,
    "vcpus": 1,
    "memory_mb": 2048,
    "instance_type": "12",
    "host": "fcaf16d02b798416d8f2a0c382589dc01c8d395ddc9af3bcc04bc894",
    "root_gb": 20,
    "image_ref": "d67f6702-9c70-4972-9bc5-2adea2e2ca85",
    "flavor": {
      "name": "cldareapd.small",
      "links": [{
          "href":
"http://192.168.60.40:8774/1b2caeedb3e2497b935723dc6e142ec9/flavors/12",
          "rel": "bookmark"
        }],
      "ram": 2048,
      "ephemeral": 0,
      "vcpus": 1,
      "disk": 20,
      "id": "12"
    },
    "OS-EXT-AZ:availability_zone": "nova",
    "image_ref_url":
"http://192.168.60.40:8774/1b2caeedb3e2497b935723dc6e142ec9/images/d67f6702-9c70-
4972-9bc5-2adea2e2ca85"
  },
  "counter_type": "cumulative"
}
```


What we need

```
“counter_name”: “cpu”,
“timestamp”: “2016-07-07T15:20:05Z”,
“counter_volume”: “26910000000”,
“project_id”: “b38a0dab349e42bdbb469274b20a91b4”
```

Issues (2/n)

- metrics refer to single resources (e.g. single VM)
- no data pre-aggregation
  - has to be done on the fly
  - too slow due to inefficiencies
- real case: cpu time of a project in a particular day:

(OpenStack Kilo)

```
# time ceilometer sample-list -m cpu -q 'project=e9ddc1e36c1348fa812c56f1922371b5;
timestamp>=2016-10-06;timestamp<=2016-10-07'
...
real    2m12.483s
```

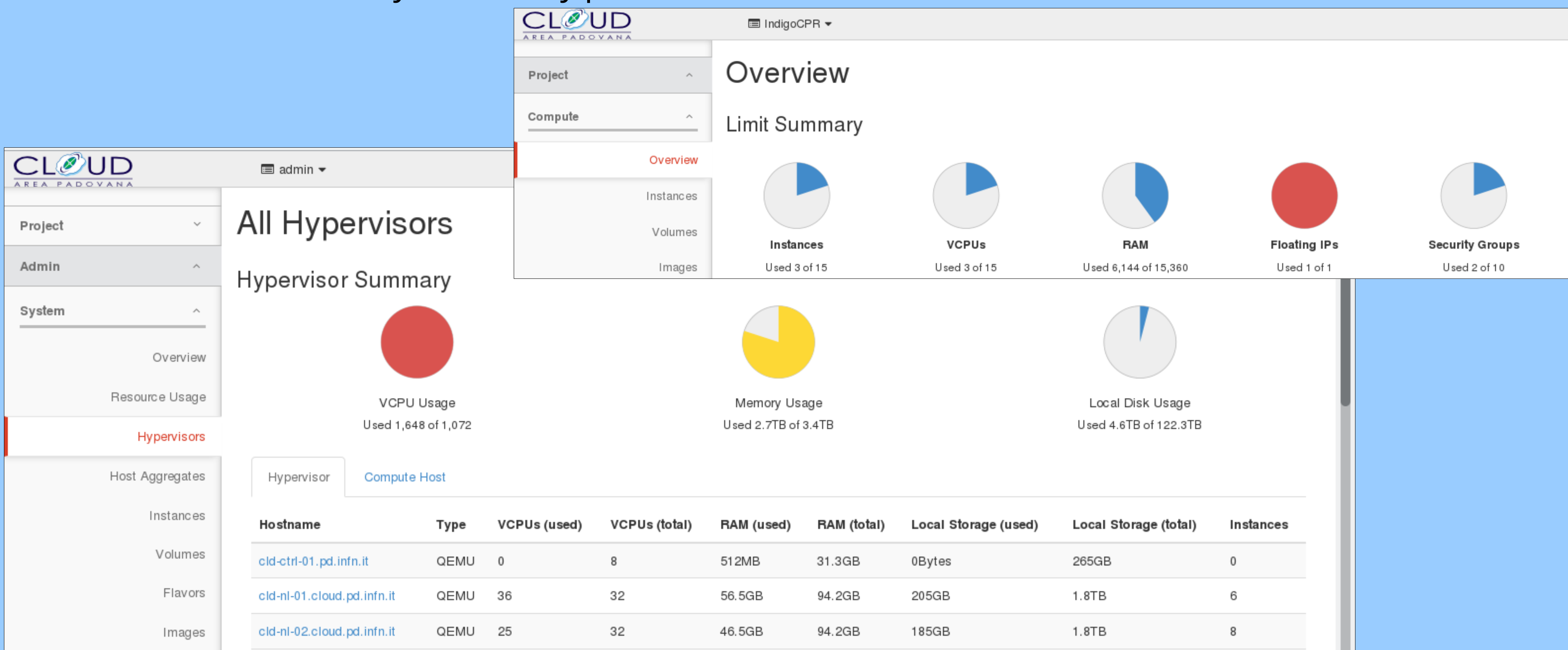
(OpenStack Mitaka)

```
# time ceilometer sample-list -m cpu -q 'project=e9ddc1e36c1348fa812c56f1922371b5;
timestamp>=2017-04-15;timestamp<=2017-04-16'
...
real    0m57.153s


# time ceilometer statistics -m cpu -q 'project=e9ddc1e36c1348fa812c56f1922371b5;
timestamp>=2017-04-15;timestamp<=2017-04-16' -p 3600 -a avg
...
real    2m49.561s
```

Issues (3/n)

- important metrics are not tracked by ceilometer
  - quotas, compute node usages, ...
  - available via OpenStack APIs
    - actual value only, no history preservation



Hw della Cloud Area Padovana

- A Padova:
  - Sistema Blade
    - 2 Enclosure
    - 4 lame DELL M620 ciascuna con 1 processore E52609 (8 core in HT), 32 GB RAM (per servizi)
    - 3 lame DELL M630 ciascuna con 2 processori E5-2650 v3 (40 core in HT), 96 GB
    - 5 lame DELL M620 ciascuna con 2 processori E5- 2670 v2 (40 core in HT), 96 GB RAM
    - 1 lama DELL M630 ciascuna con 2 processori E5-2670 v3 (48 core in HT), 512 GB
    - 6 lame DELL M630, ciascuna con 2 processori E5-2680 v3 (48 core in HT), 128 GB
  - Storage
    - Server iSCSI DELL MD3620i, con 23 dischi SAS da 900 GB
    - Espansione Dell MD1200 con 12 dischi da 4 TB
    - Espansione Dell MD1200 con 12 dischi da 10 TB
  - Altre risorse per altri servizi
    - Controller e Network Node, Foreman/Puppet, mysql, mongodb, HAProxy/KeepAlived, Ganglia, Nagios, NAT
- A Legnaro
  - 6 Fujitsu Primergy RX300S8 con 2 processori XEON E5 2650v2 (32 core in HT), 96 GB RAM
  - 7 DELL PowerEdge R430 con 2 processori E5-2640 v3 (32 core in HT), 128 GB RAM