Site Infrastructure:
Cloud Computing in Action

Dr. Silvio Pardi
INFN-NAPOLI

SCGCCW 2014 –TBILISI
Third ATLAS South Caucasus Grid & Cloud Computing Workshop

The NIST Definition of Cloud Computing

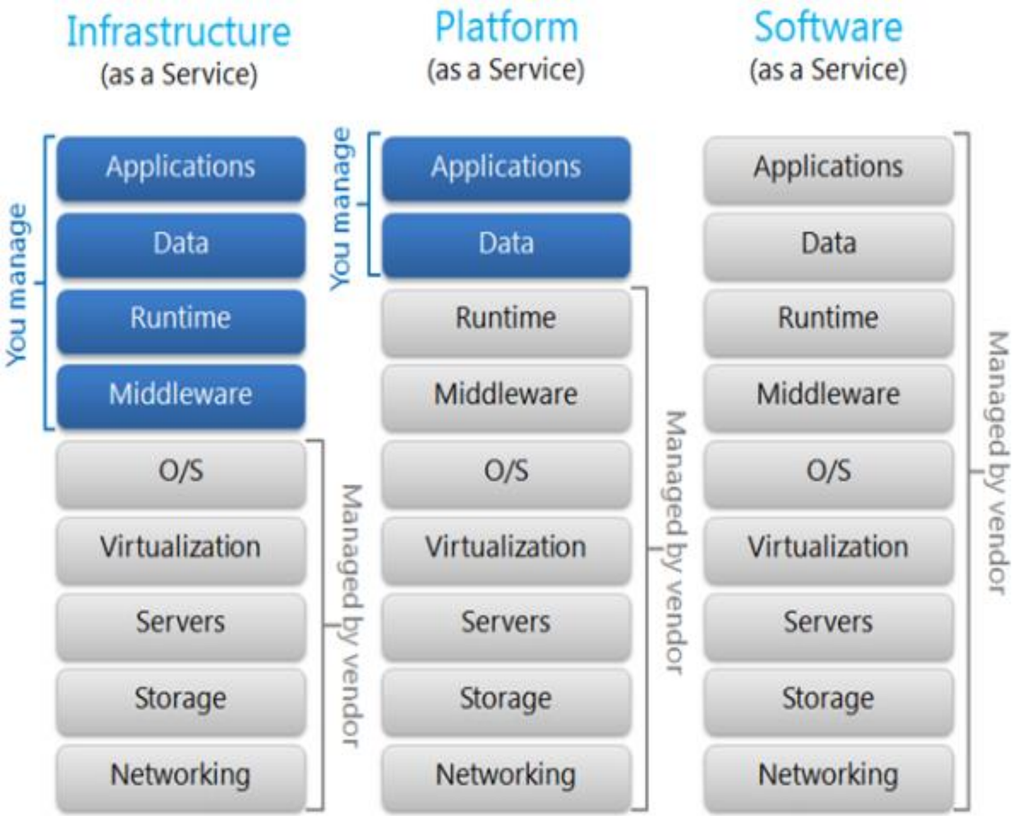
Cloud computing is a paradigm of resource provisioning that stress the concept of resource virtualization.

Following the NIST definition a Cloud Infrastructure must have five essential characteristics:

- On-demand self-service.
- Broad network access.
- Resource pooling.
- Rapid elasticity.
- Measured service.

Cloud Service and Deployement Models

Three service models: Infrastructure as a Service (Iaas), Platform as a Service (Paas), Software as a Service (SaaS)



http://blogs.msdn.com/b/johnalioto/archive/2010/08/16/10050822.aspx

Four Deployment models

Private Cloud: The cloud infrastructure is provisioned for exclusive use by a single organization/institution

Community Cloud: The cloud infrastructure is provisioned for exclusive use by a specific community.

Public Cloud: The cloud infrastructure is provisioned for open use by the general public (AMAZON, Google Compute Engine, Microsoft)

Hybrid Cloud: The cloud infrastructure is a composition of two or more distinct cloud infrastructures (private, community, or public)

Attractive features of Cloud Computing:

- Virtualization increase the application portability.

- The cloud allows any resource centers to contribute to experiments distributed computing without a specific know-how on the applications.

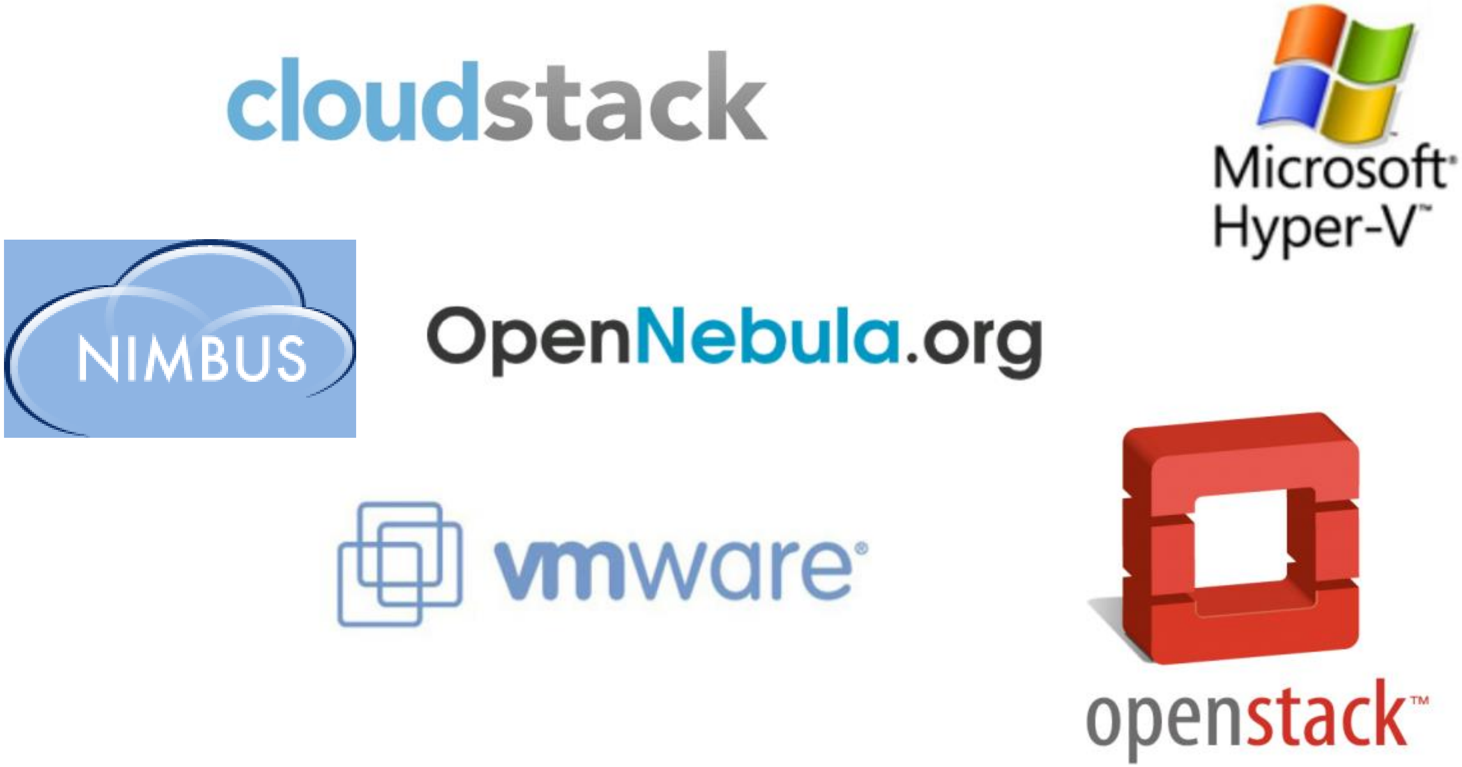
- Infrastructure as a Service (IaaS) clouds provide a simple way to dynamically manage the load between multiple projects within a single center.

- A distributed cloud aggregates heterogeneous clouds into a unified resource with a single entry point for users (namely, a PanDA queue).

Use-Cases

- PanDA queues in the cloud
- Use Opportunistic resources
- Analysis Clusters in the Cloud (Tier 3)
- High Availability Services

Common Tools



- OpenStack: Currently the most used cloud solution
- Puppet as preferred configuration management tool
- Most deployments depend on CVMFS

Cern Tools

To accelerate the adoption of Cloud Technologies in HEP experiments, CERN lab has developed a set of tools that included:

CernVM is a baseline Virtual Software Appliance for the participants of CERN LHC experiments. The Appliance represents a complete, portable and easy to configure user environment for developing and running LHC data analysis locally and on institutional and commercial clouds.

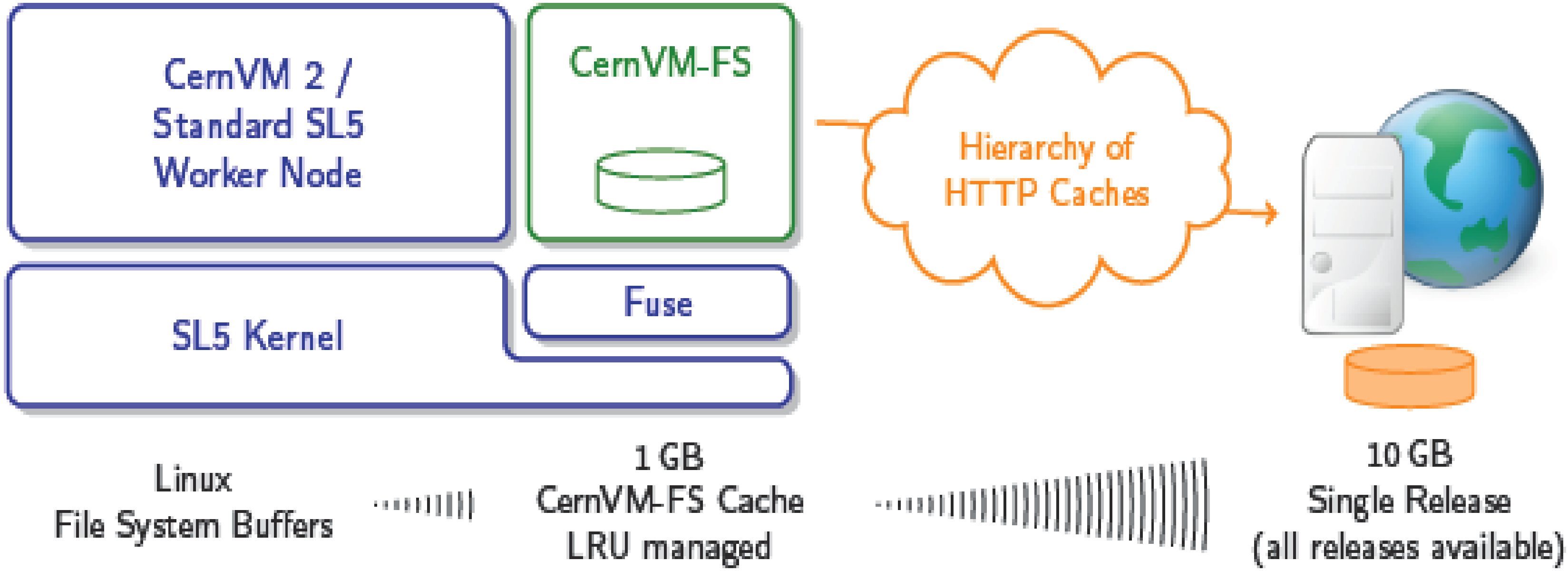
CernVM Co-Pilot is a framework for instantiating an ad-hoc computing infrastructure on top of distributed computing resources. Such resources include commercial computing clouds (e.g. Amazon EC2), scientific computing clouds (e.g. CERN lxcloud), as well as the machines of users participating in volunteer computing projects (e.g. BOINC)

CernVM Online is a new mechanism that allows you to effortless contextualize your CernVM instances, both in your desktop and the cloud, providing a user-friendly interface.

CernVM File System (CernVM-FS)

CVMFS is a network file system based on HTTP and optimized to deliver experiment software in a fast, scalable, and reliable way.
CernVM-FS is included in CernVM, however it can also be used outside CernVM.

CVMFS plays a key role in deploying all the current Cloud Based solutions for HEP experiments


CernVM 2 / Standard SL5 Worker Node
CernVM-FS
Hierarchy of HTTP Caches
SL5 Kernel
Fuse
Linux File System Buffers
1 GB CernVM-FS Cache LRU managed
10 GB Single Release (all releases available)

ATLAS & Cloud Computing

The ATLAS experiment has promoted a large program of Cloud Computing R&D activities, finalized to study how to exploit the Cloud paradigm.

The international ATLAS community has been able to demonstrate the feasibility of integrate transparently various cloud resources into the PanDA workload management system.

In the next slide I will present an overview on the main technologies now in place and a set of activities carried on that include:

- Cloud Scheduler
- Overlay on ATLAS HLT nodes
- Experience with commercial cloud
- A Study for a distributed Tier2

Source: https://cds.cern.ch/record/1621892/files/ATL-SOFT-PROC-2013-034.pdf

Cloud Scheduler

Cloud Scheduler is a python package for managing VMs on IaaS clouds

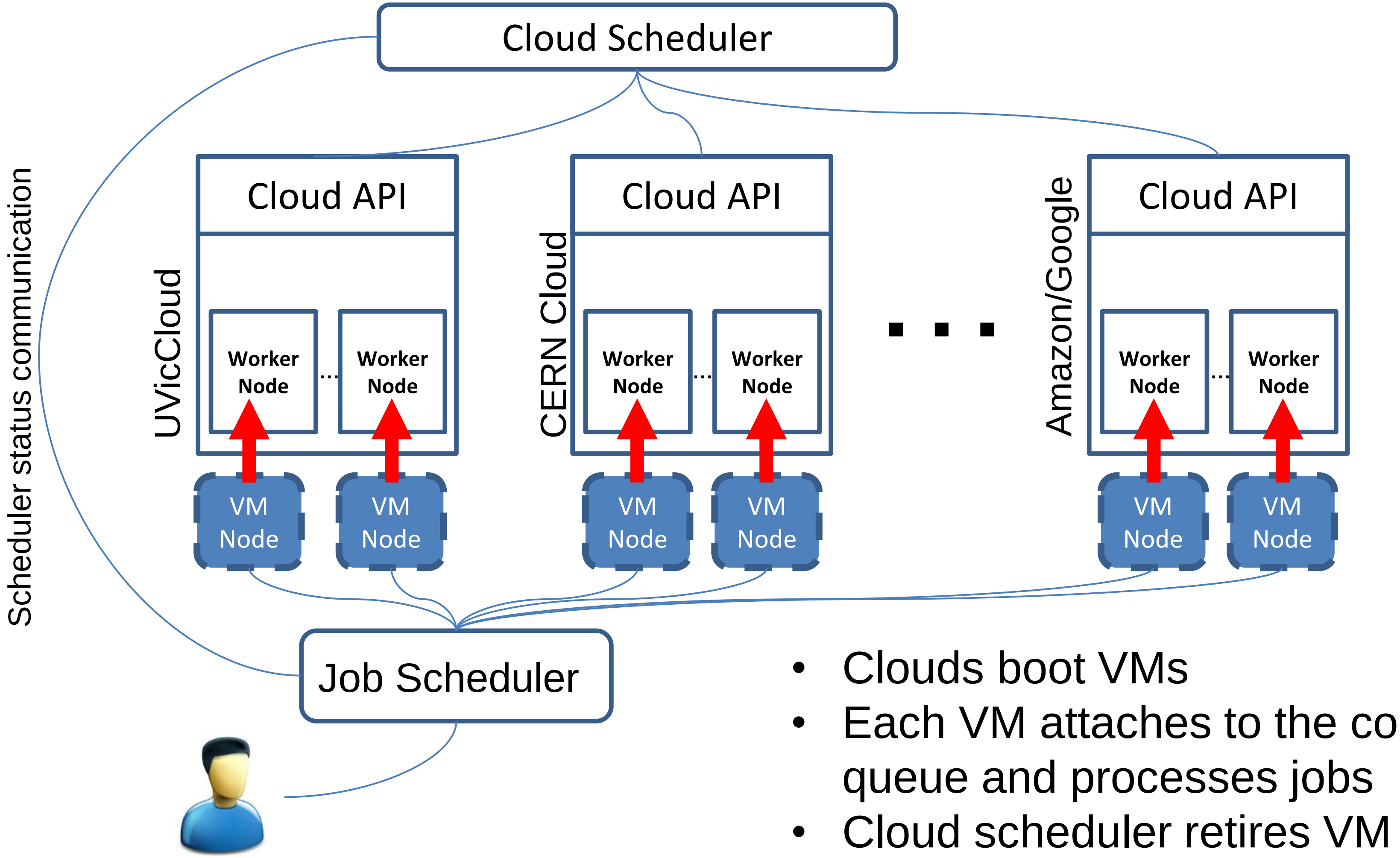
- Users submit HTCondor jobs
  - Optional attributes specify virtual machine properties
- Developed at UVic and NRC since 2009
- Used by ATLAS, CANFAR, and BaBar

Key Features of Cloud

- Dynamically manages quantity and type of VMs in response to user demand
- Easily connects to many IaaS clouds, and aggregates their resources
- Provides IaaS resources in the form of an ordinary HTCondor batch system
- CernVM images are used (both Xen and KVM flavors), and Puppet is used to manage the system configuration of the VMs

Source: F. Berghaus (University of Victoria) – "An Overview of Cloud Scheduler"
http://heprc.phys.uvic.ca/sites/heprc.phys.uvic.ca/files/belle2-cloudscheduler.pdf

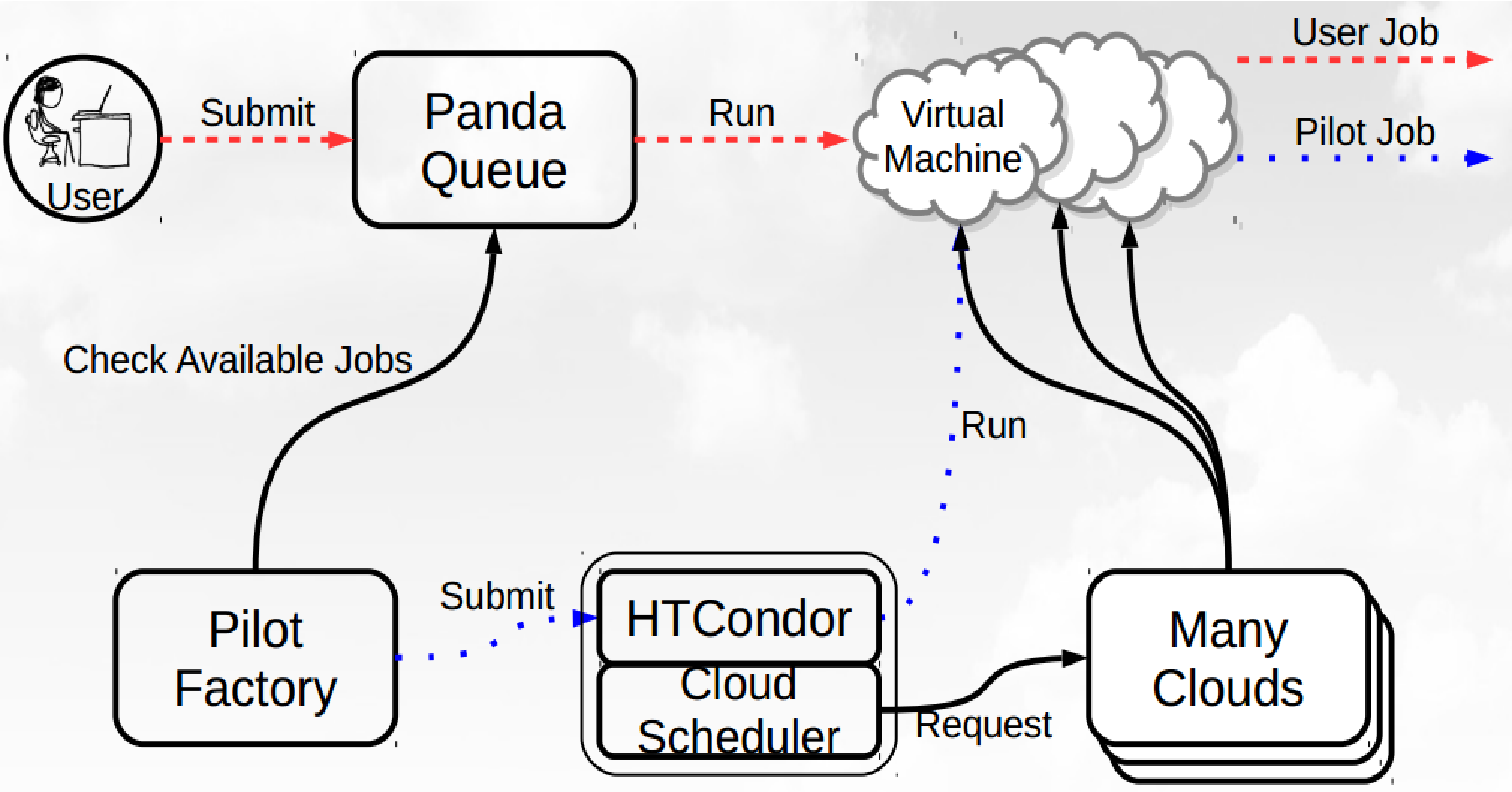
How Cloud Scheduler works



- Clouds boot VMs
- Each VM attaches to the condor queue and processes jobs
- Cloud scheduler retires VM when no jobs require that VM

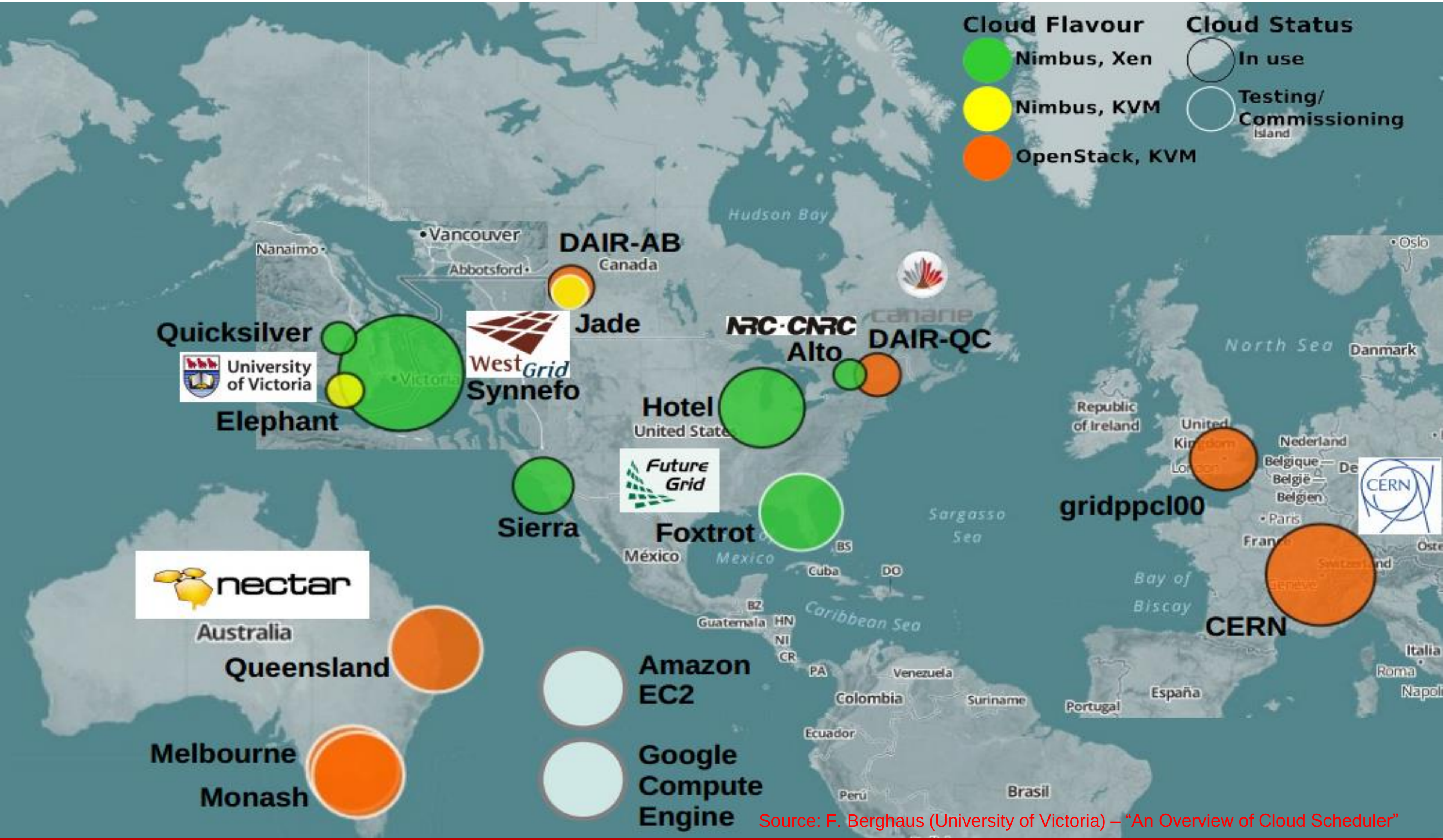
Source: F. Berghaus (University of Victoria) – “An Overview of Cloud Scheduler”

Cloud Job Flow (on the Grid)



Source: F. Berghaus (University of Victoria) – “An Overview of Cloud Scheduler”

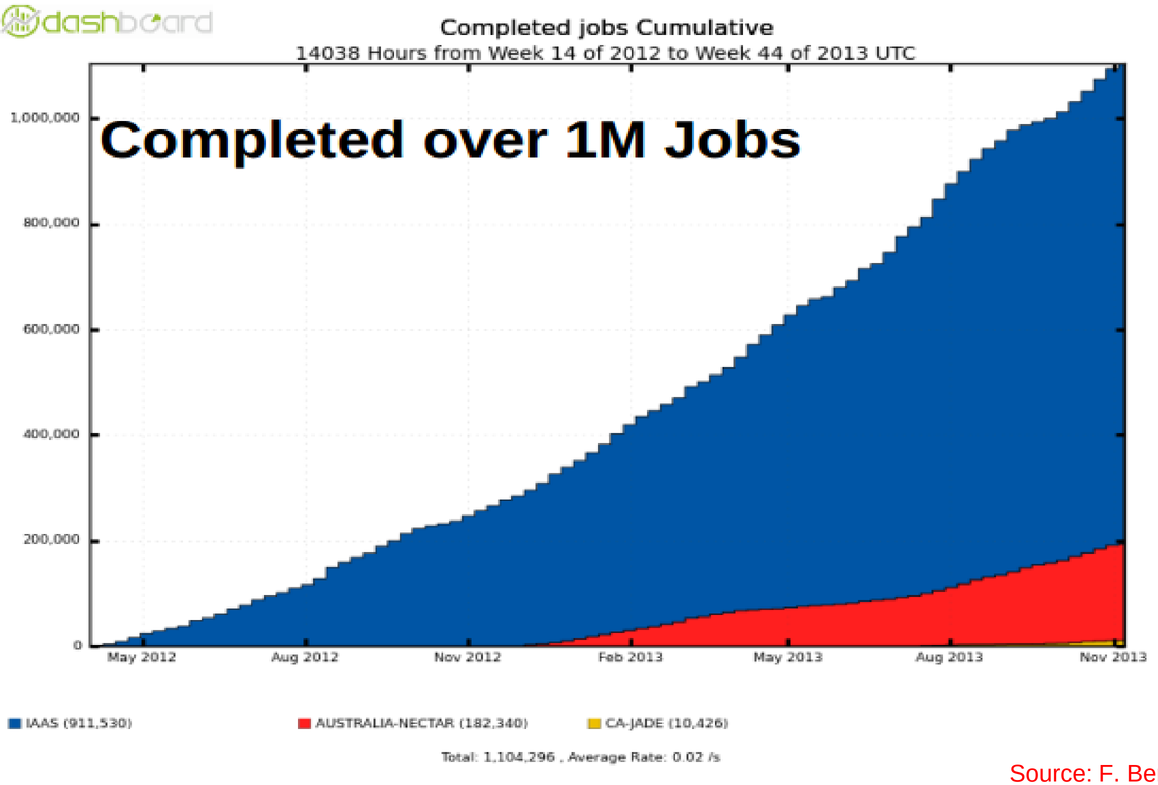
The ATLAS Grid of Clouds



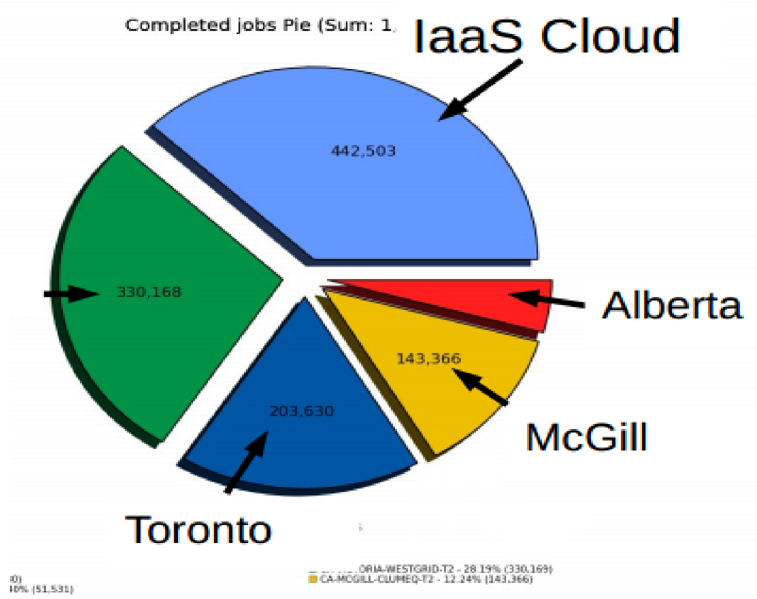
Source: F. Berghaus (University of Victoria) – “An Overview of Cloud Scheduler”

Cloud Production Experience

- Started operation April 2012



Completed over 1M Jobs



IaaS Cloud

Victoria

Alberta

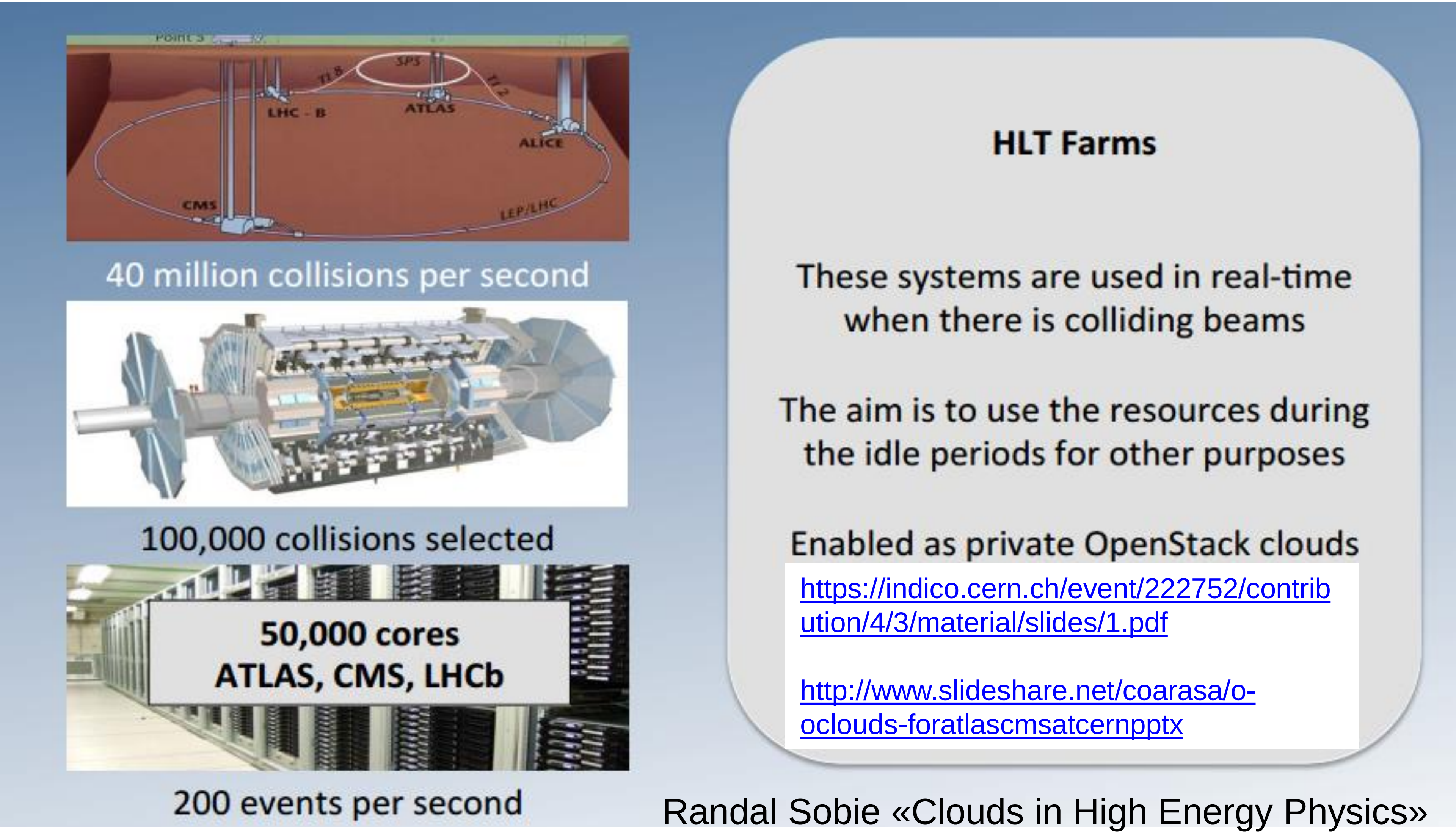
McGill

Toronto

- Similar performance to dedicated facilities at
  - University of Victoria
  - McGill University
  - University of Alberta
  - University of Toronto

Source: F. Berghaus (University of Victoria) – “An Overview of Cloud Scheduler”

Overlay on ATLAS HLT nodes



40 million collisions per second

100,000 collisions selected

50,000 cores
ATLAS, CMS, LHCb

200 events per second

HLT Farms

These systems are used in real-time when there is colliding beams

The aim is to use the resources during the idle periods for other purposes

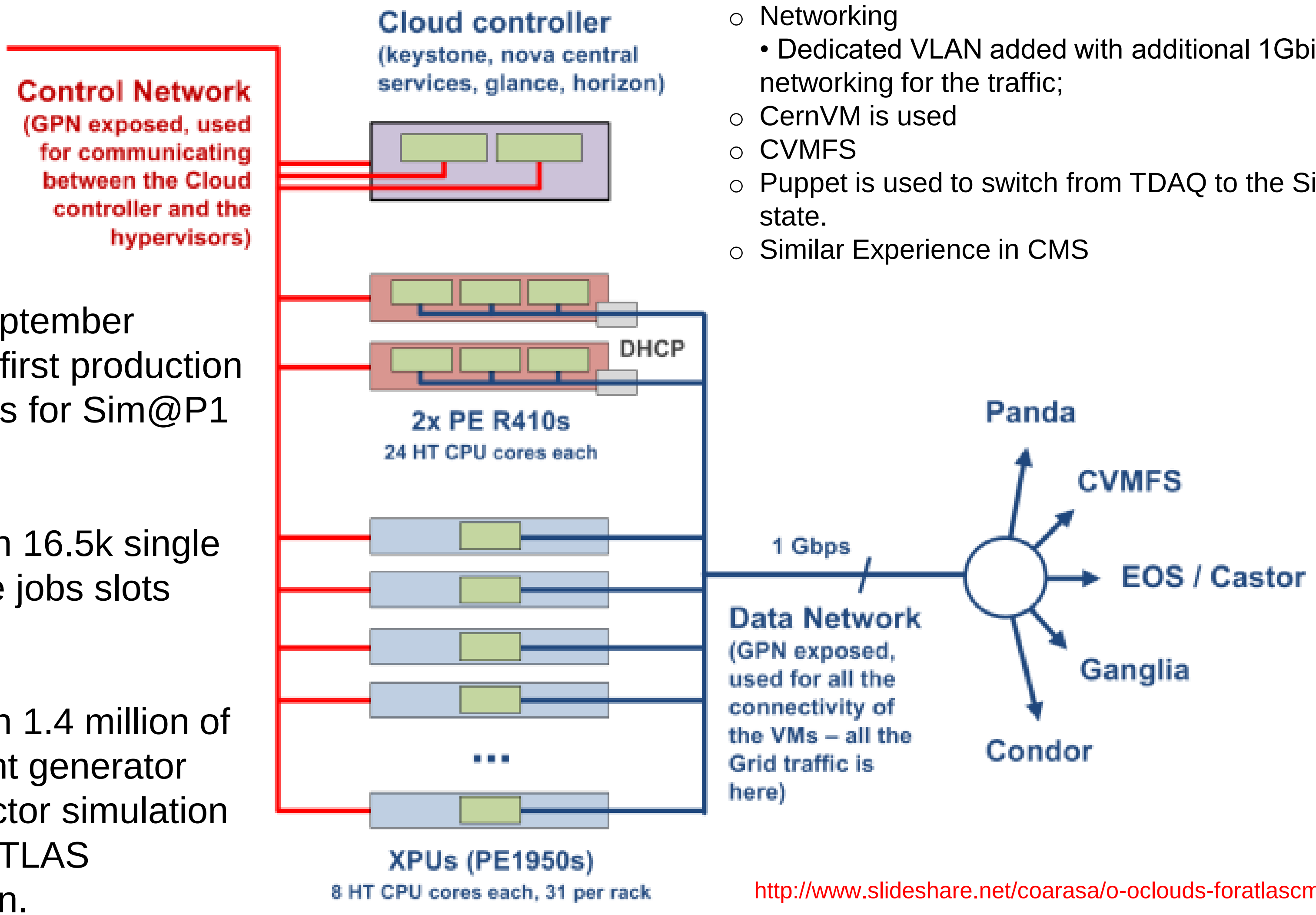
Enabled as private OpenStack clouds

https://indico.cern.ch/event/222752/contribution/4/3/material/slides/1.pdf

http://www.slideshare.net/coarasa/o-oclouds-foratlascmsatcernpptx

Randal Sobie «Clouds in High Energy Physics»

Overlay on ATLAS HLT nodes


Cloud controller
(keystone, nova central services, glance, horizon)
Control Network
(GPN exposed, used for communicating between the Cloud controller and the hypervisors)
DHCP
2x PE R410s
24 HT CPU cores each
...
XPUs (PE1950s)
8 HT CPU cores each, 31 per rack
1 Gbps
Data Network
(GPN exposed, used for all the connectivity of the VMs – all the Grid traffic is here)
Panda
CVMFS
EOS / Castor
Ganglia
Condor

- July − September 2013 the first production operations for Sim@P1 project
- More than 16.5k single CPU core jobs slots used.
- More than 1.4 million of both event generator and detector simulation jobs for ATLAS production.

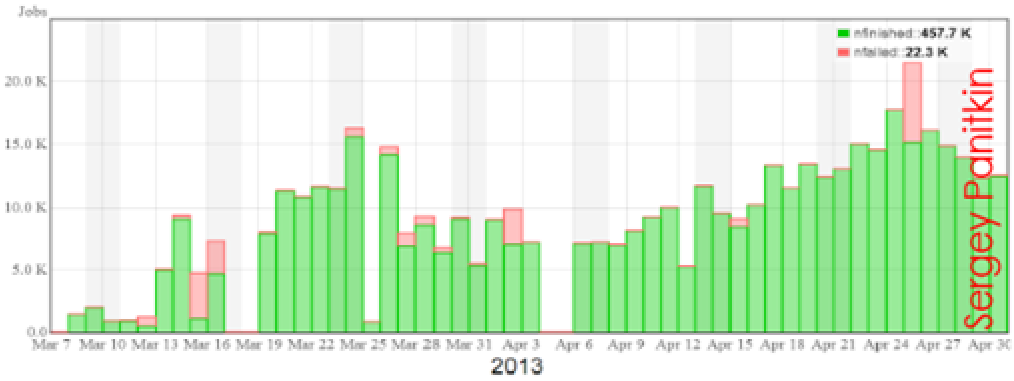
- ○ Networking
  - Dedicated VLAN added with additional 1Gbit/s networking for the traffic;
- ○ CernVM is used
- ○ CVMFS
- ○ Puppet is used to switch from TDAQ to the Sim@P1 state.
- ○ Similar Experience in CMS

http://www.slideshare.net/coarasa/o-oclouds-foratlascmsatcernpptx

J.A. Coarasa (CERN) Overlay opportunistic clouds in CMS/ATLAS at CERN OpenStack Summit, 15-18 April 2013, Portland, USA

Google Compute Engine Project

- ATLAS was invited to participate in GCE closed preview, August 2012
- Google allocated additional resources after initial period
  - 5M core-hours, 4k cores for 2 months (original allocation: 1k cores)
- Resources organized as HTCondor based PanDA queue
  - Transparently included into ATLAS computational grid
- Project idea: test long term stability while running a cloud cluster similar in size to a T-2 site in ATLAS
- 8 weeks of running, computationally intensive workloads, Physics Event generators, fast and full detection simulation
- Very stable running on GCE side, most problems on ATLAS side non cloud related, overall failure rate about 6% mostly during start-up
- 458K Jobs completed, 214Mevent generated and processed



Sour

Supercomputers” - https://cds.cern.ch/record/1669859/files/ATL-SOFT-SLIDE-2014-115.pdf

Amazon Elastic Compute Cloud (EC2)

- RACF BNL group received grant allocation from Amazon EC2 in 2013
- Set up hybrid cloud using resources at BNL T-1 and “elastic” part of cloud on Amazon EC2 (i.e. spanning geographically distributed EC2 sites). HTCondor-G with cloud interface.
- Ran 5000 EC2 VMs for about 3 weeks, Executed ATLAS simulation production jobs (high CPU, low I/O)
- Reliable operations of EC2 platform but poor job efficiency due to long running jobs

Source: Dr. Paul Nilsson ”Extending ATLAS Computing to Commercial Clouds and Supercomputers” - https://cds.cern.ch/record/1669859/files/ATL-SOFT-SLIDE-2014-115.pdf

Study for a distribted DISTRIBUTED TIER2

Goal of the investigation

- Investigate the possibility of creating a unified, geographically distributed ,Tier2 class infrastructure through technologies of Cloud Computing

- Verify the impact of latency on the system stability and performance

- Test the resilience of distributed file systems on geographical setup

- Investigate the possibility of implementig T2 service machines such as Grid service SE, CE, Squid.

People

INFN-ROMA

Cristina Bulfon
Alessandro De Salvo
Carlo Graziosi
Daniela Anzellotti
Danilo D' Angelo
Enrico Pasqualucci
Alessandro Spanu
Marco Esposito

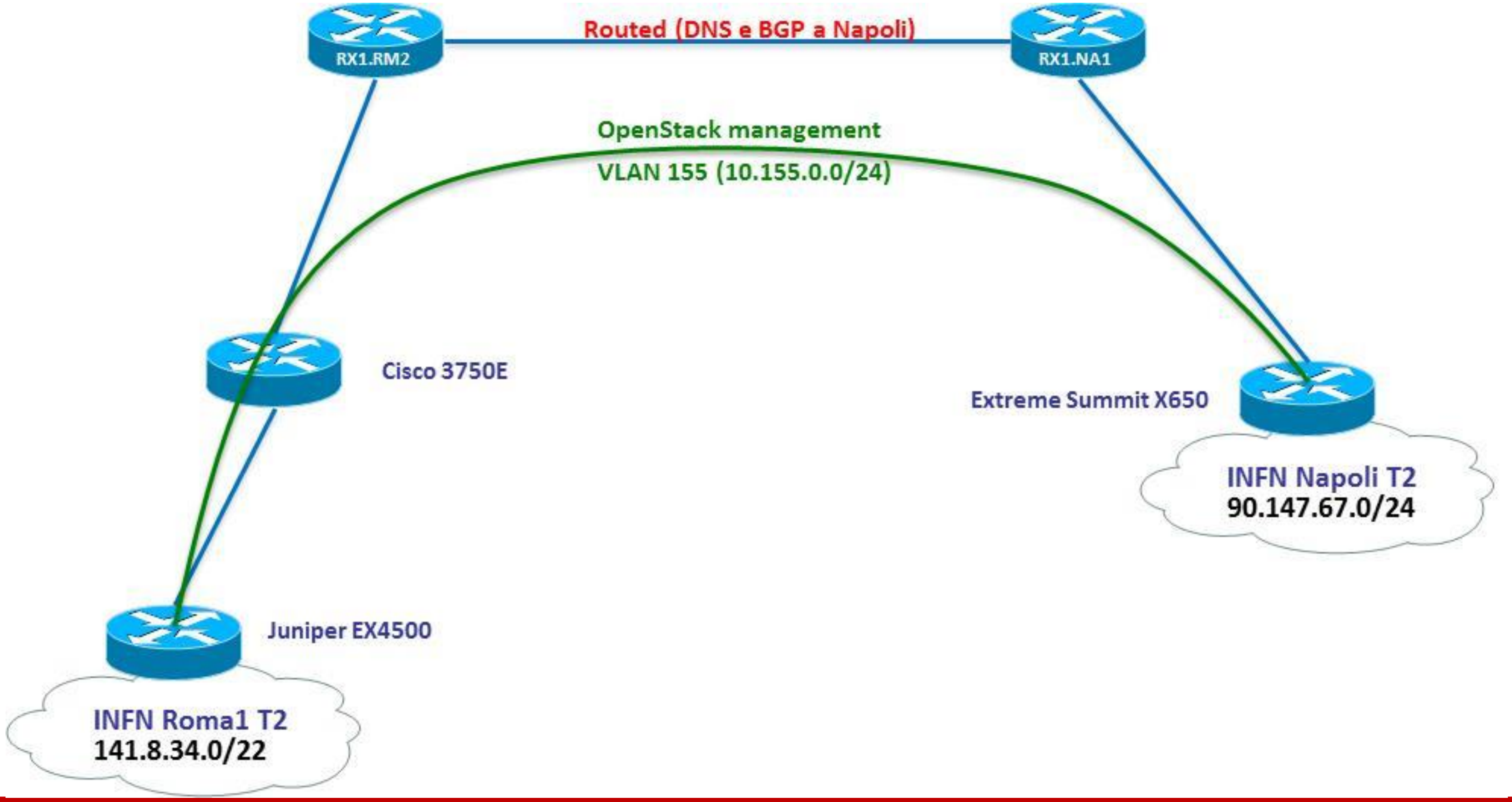
INFN-NAPOLI

Enzo Capone
Gianpaolo Carlino
Alessandra Doria
Silvio Pardi

GARR
Massimo Carboni (INFN-LNF)
Paolo Bolletta
Lorenzo Puccio

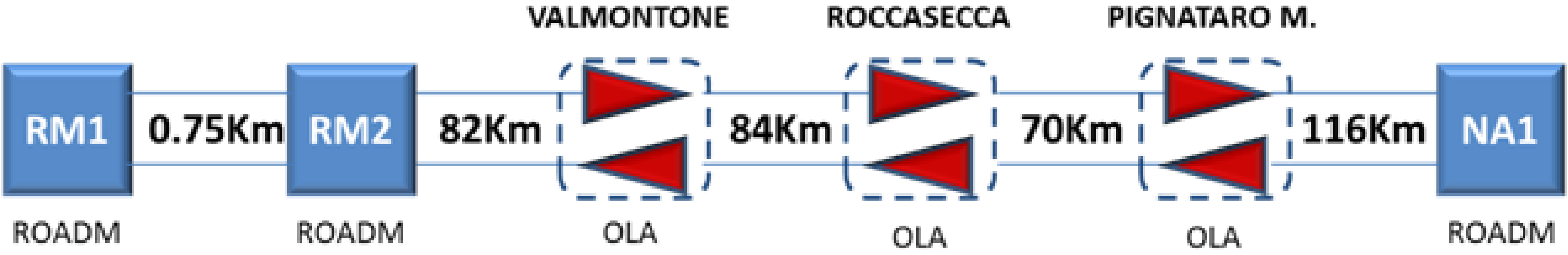
NETWORK LINK

The realized end-to-end service transports transparently Ethernet frames between the Client interfaces of the two INFN Tier2 sites (Napoli and Roma). The link offer a 1Gbps guaranteed bandwidth.

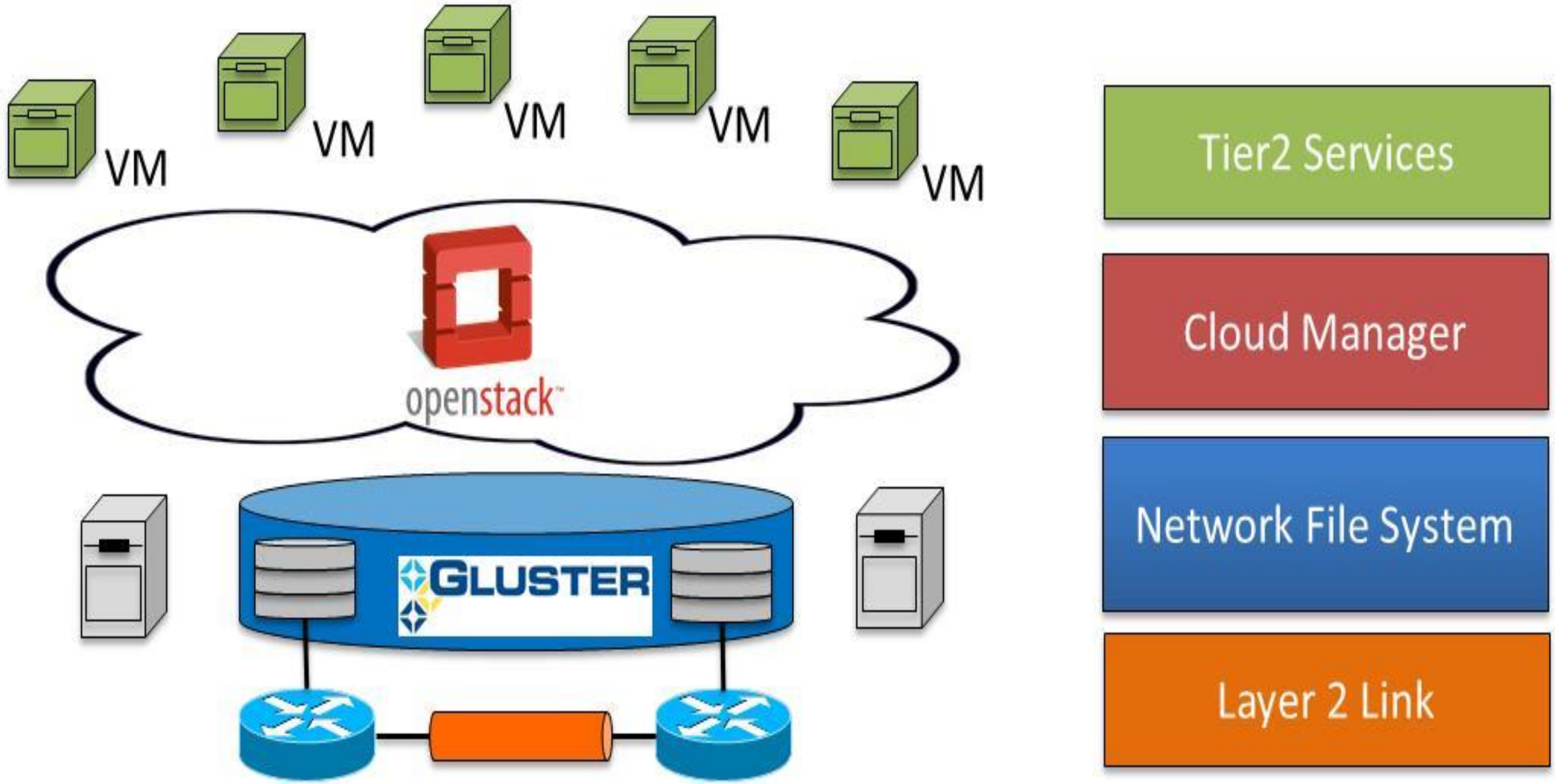


NETWORK LINK

- RTT : 5 ms
- JITTER (measured with IPERF in UDP) : 0.08 ms
- Throughput measured with IPERF between two servers: 938Mbit/s
- Packet Loss 0% on1x10^6 trials (measured with flooding ping)

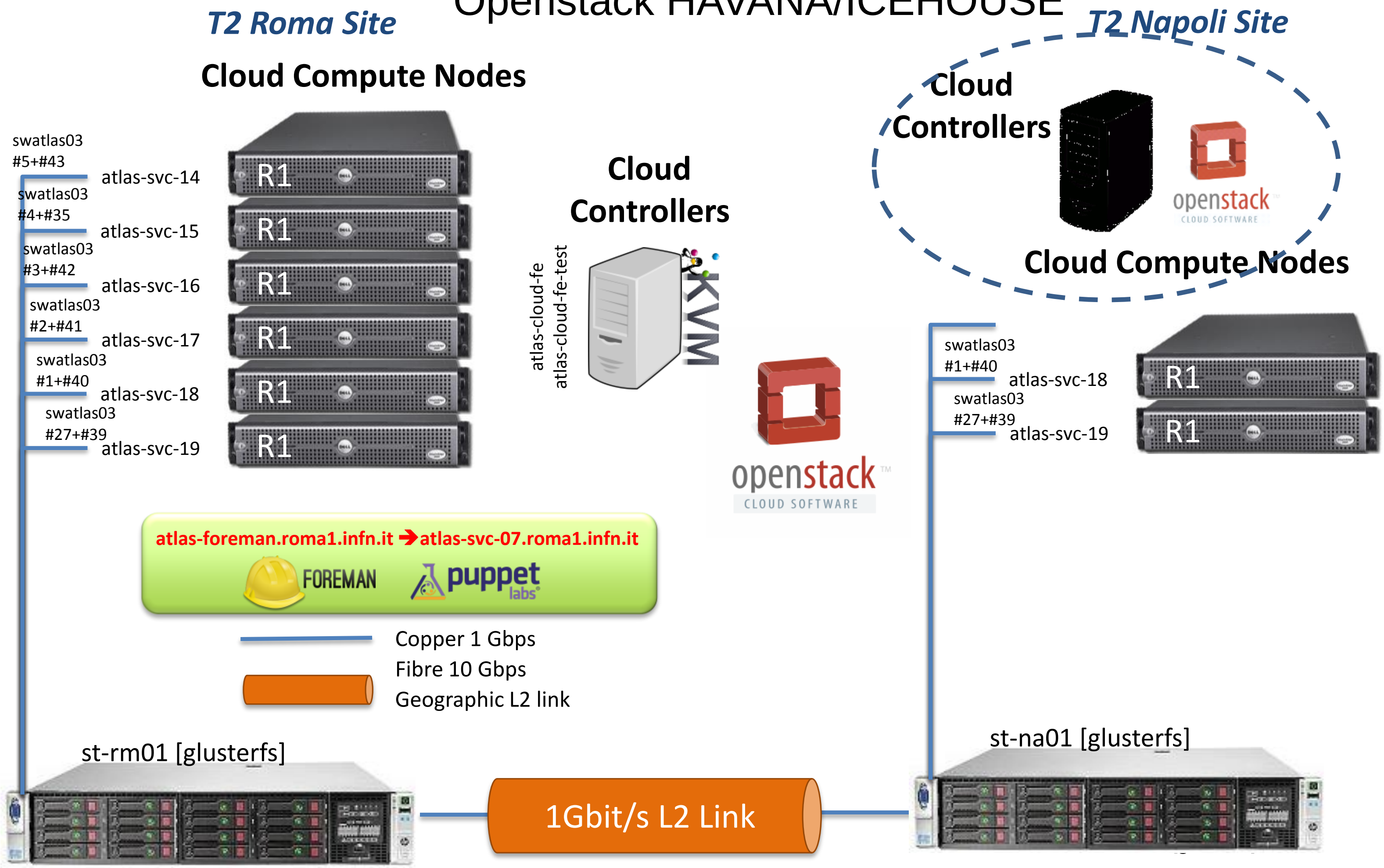


ARCHITECTURE



IMPLEMENTATION

Gluster 3.4.3
Openstack HAVANA/ICEHOUSE



T2 Roma Site

Cloud Compute Nodes

swatlas03 #5+#43 atlas-svc-14

swatlas03 #4+#35 atlas-svc-15

swatlas03 #3+#42 atlas-svc-16

swatlas03 #2+#41 atlas-svc-17

swatlas03 #1+#40 atlas-svc-18

swatlas03 #27+#39 atlas-svc-19

Cloud Controllers

atlas-cloud-fe
atlas-cloud-fe-test

atlas-foreman.roma1.infn.it →atlas-svc-07.roma1.infn.it

Copper 1 Gbps

Fibre 10 Gbps
Geographic L2 link

st-rm01 [glusterfs]

1Gbit/s L2 Link

T2 Napoli Site

Cloud Controllers

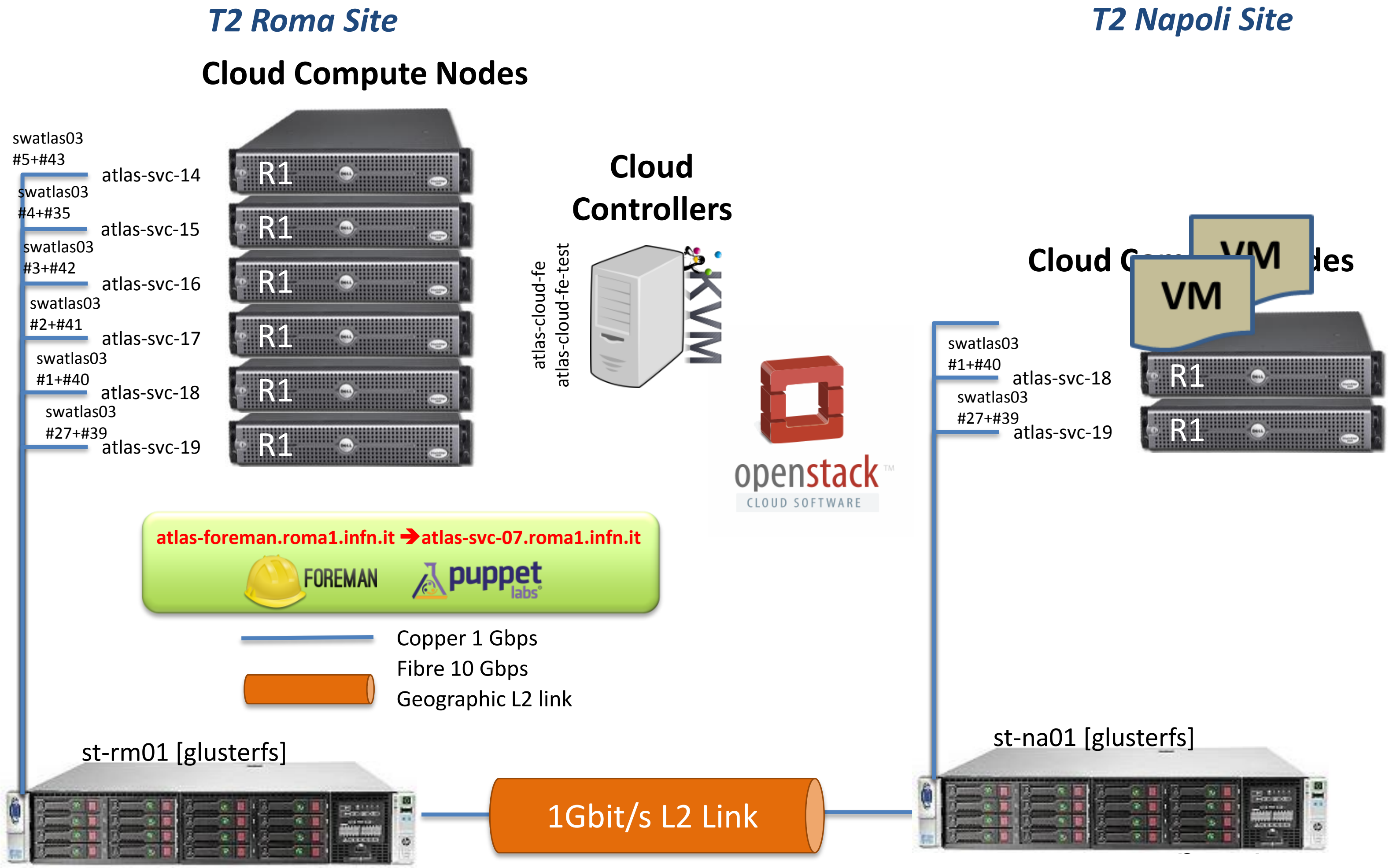
Cloud Compute Nodes

swatlas03 #1+#40 atlas-svc-18

swatlas03 #27+#39 atlas-svc-19

st-na01 [glusterfs]

RESILIENCE TEST



T2 Roma Site

T2 Napoli Site

Cloud Compute Nodes

swatlas03 #5+#43 atlas-svc-14

swatlas03 #4+#35 atlas-svc-15

swatlas03 #3+#42 atlas-svc-16

swatlas03 #2+#41 atlas-svc-17

swatlas03 #1+#40 atlas-svc-18

swatlas03 #27+#39 atlas-svc-19

Cloud Controllers

atlas-cloud-fe

atlas-cloud-fe-test

KVM

openstack CLOUD SOFTWARE

Cloud C **VM** **VM** **des**

swatlas03 #1+#40 atlas-svc-18

swatlas03 #27+#39 atlas-svc-19

atlas-foreman.roma1.infn.it →atlas-svc-07.roma1.infn.it

FOREMAN puppet labs

Copper 1 Gbps

Fibre 10 Gbps

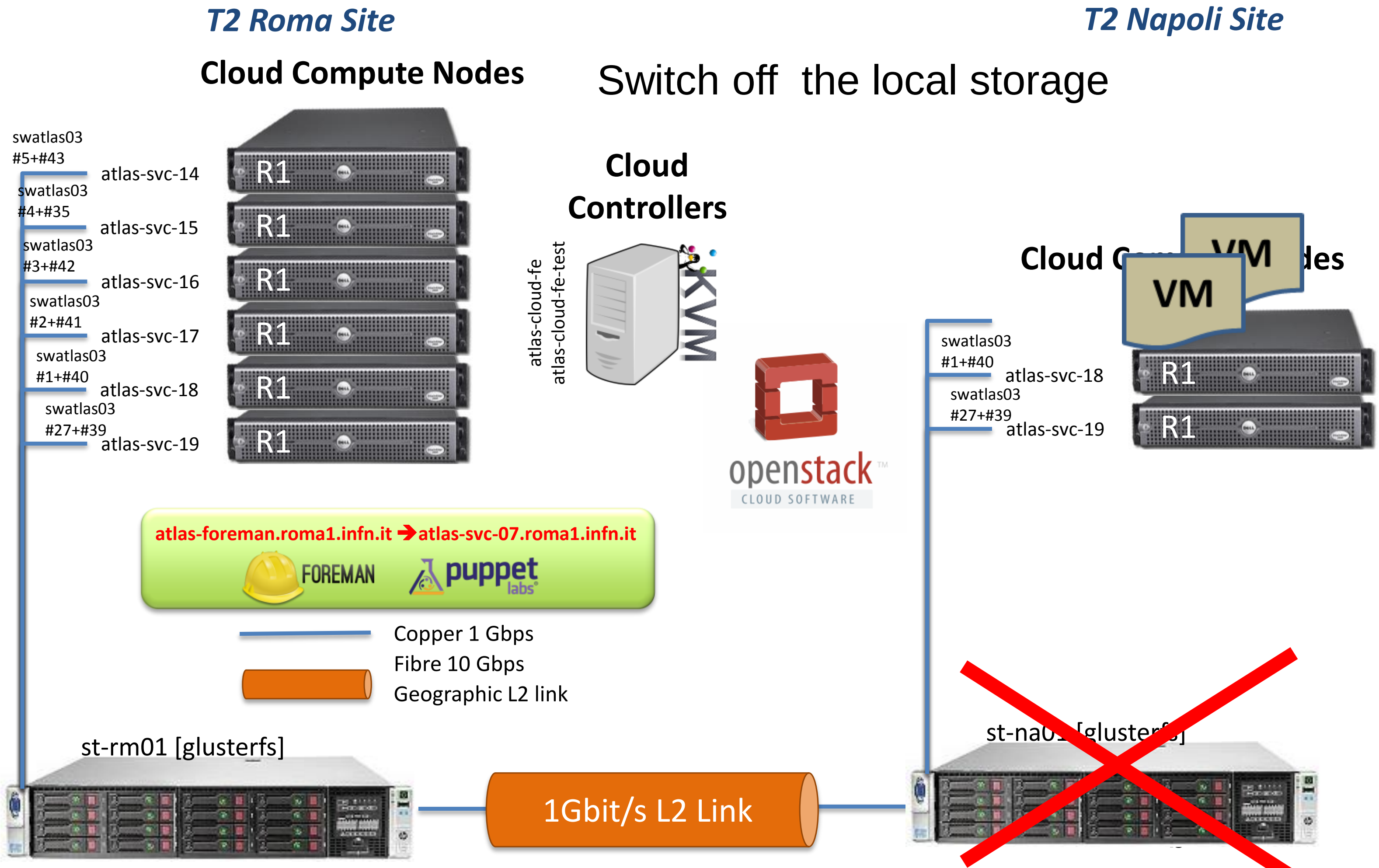
Geographic L2 link

st-rm01 [glusterfs]

1Gbit/s L2 Link

st-na01 [glusterfs]

RESILIENCE TEST



T2 Roma Site

Cloud Compute Nodes

Switch off the local storage

swatlas03 #5+#43 atlas-svc-14

swatlas03 #4+#35 atlas-svc-15

swatlas03 #3+#42 atlas-svc-16

swatlas03 #2+#41 atlas-svc-17

swatlas03 #1+#40 atlas-svc-18

swatlas03 #27+#39 atlas-svc-19

Cloud Controllers

atlas-cloud-fe

atlas-cloud-fe-test

KVM

openstack™ CLOUD SOFTWARE

T2 Napoli Site

Cloud C **des**

VM

VM

swatlas03 #1+#40 atlas-svc-18

swatlas03 #27+#39 atlas-svc-19

atlas-foreman.roma1.infn.it → atlas-svc-07.roma1.infn.it

FOREMAN puppet labs

Copper 1 Gbps

Fibre 10 Gbps

Geographic L2 link

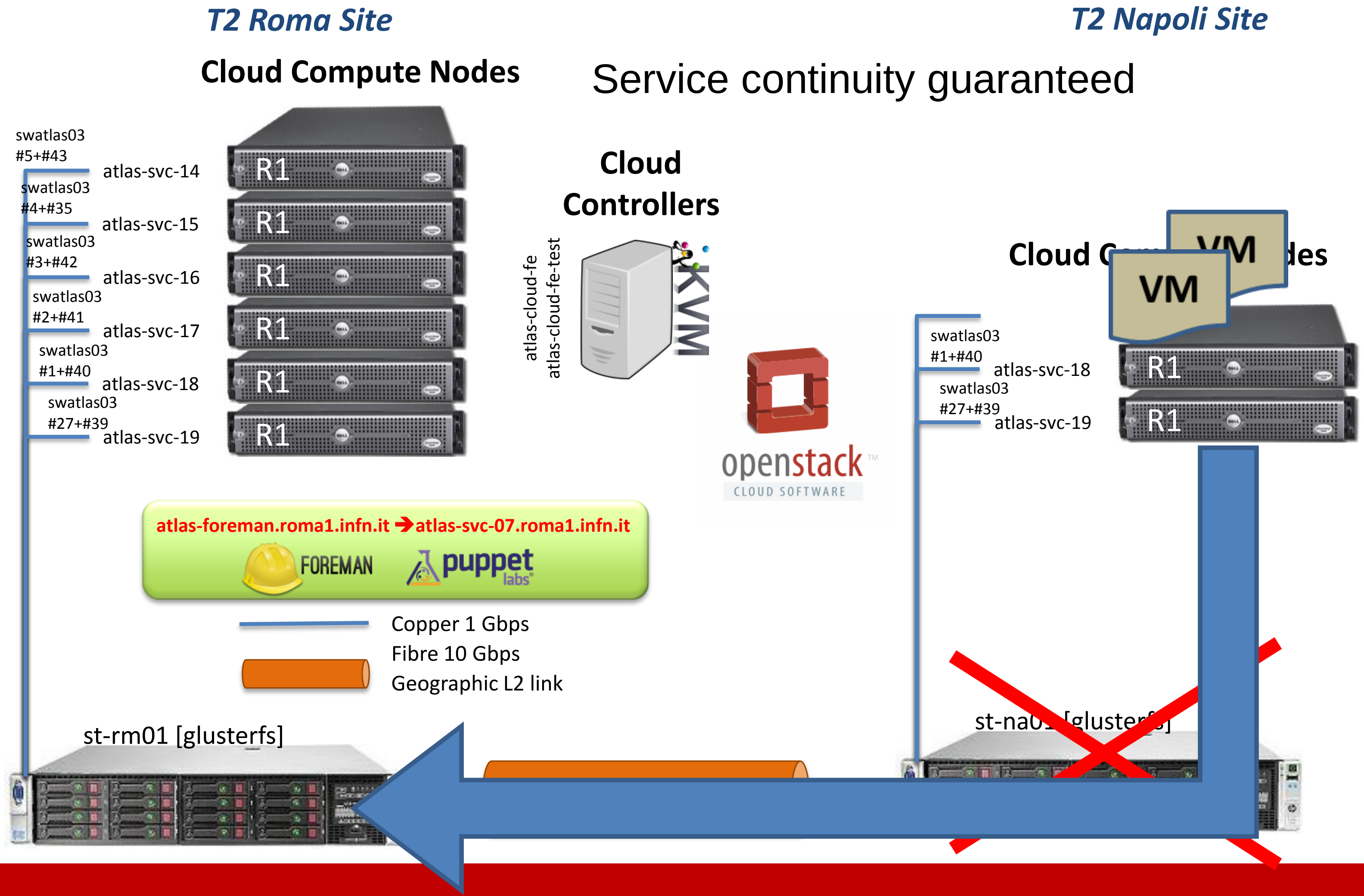
st-rm01 [glusterfs]

1Gbit/s L2 Link

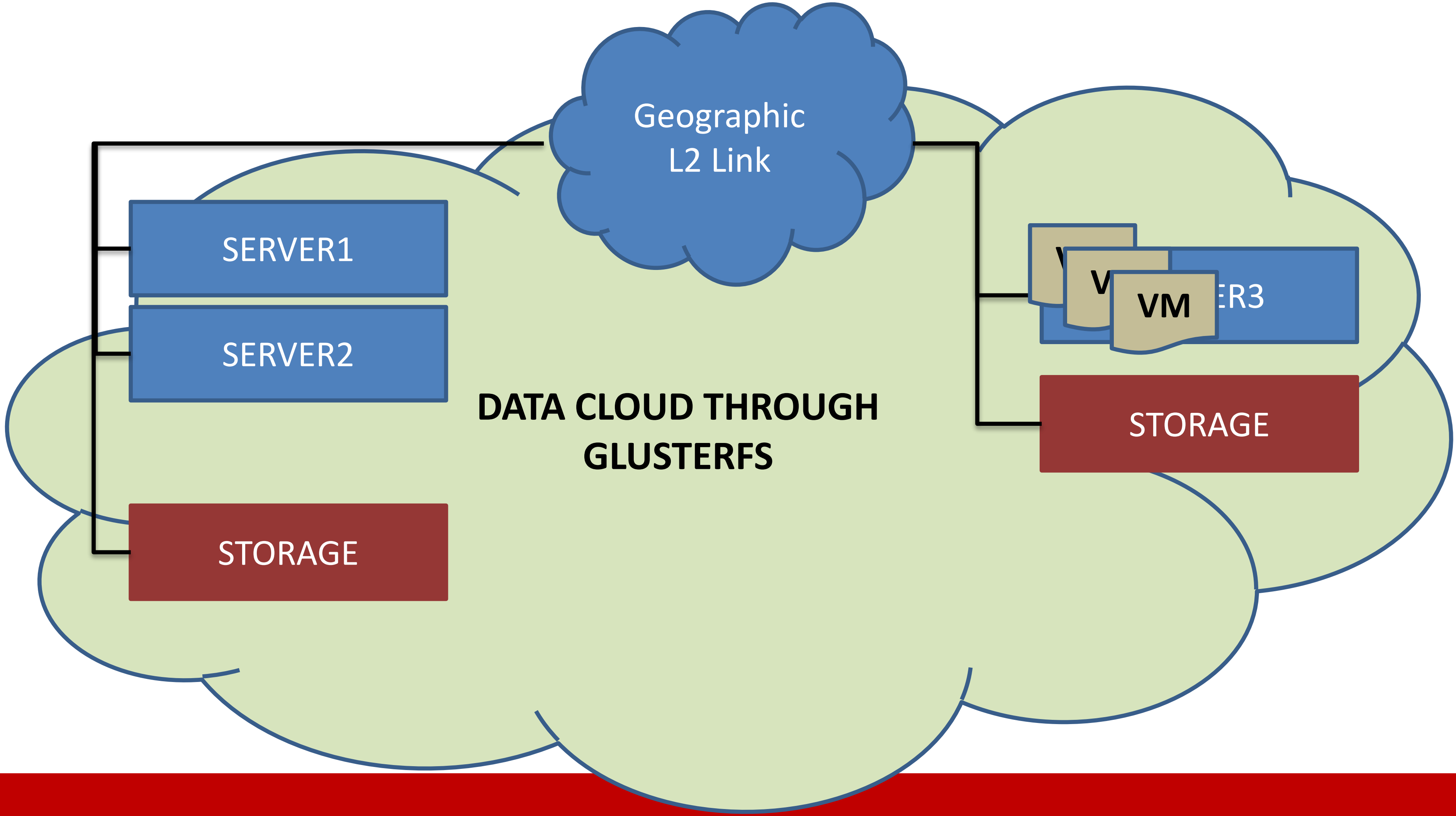
st-na01 [glusterfs]

RESILIENCE TEST

Service continuity guaranteed



INFN-ROMA INFN-NAPOLI

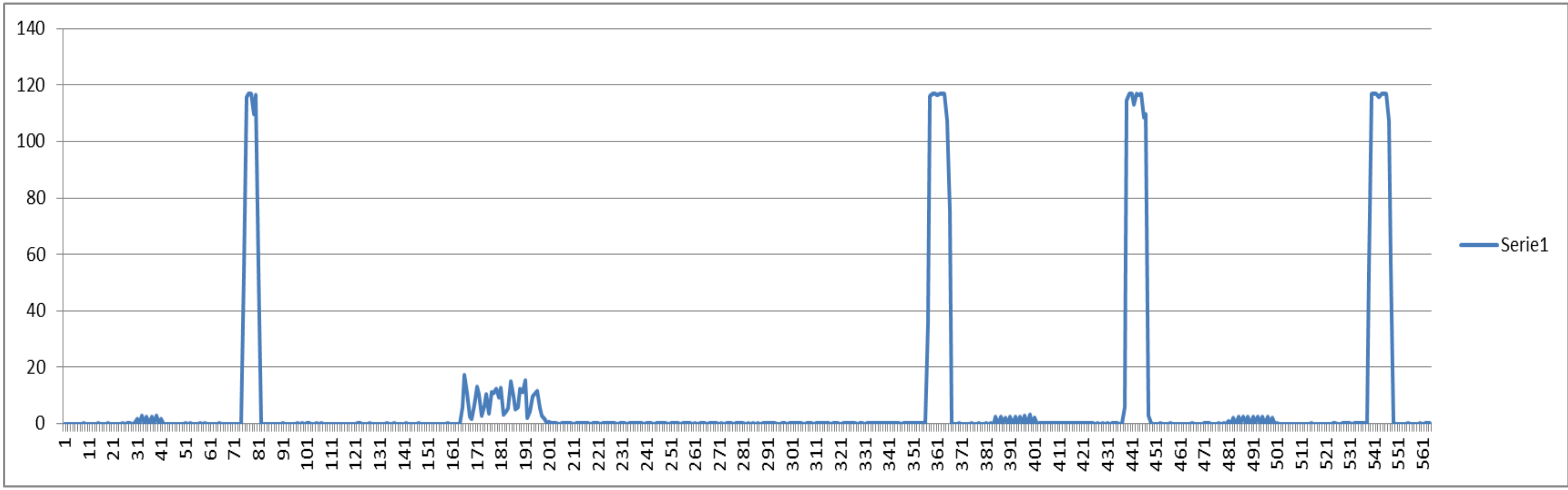


LIVE MIGRATION TEST

Live Migration done with virsh migrate.

Migration time of 10 seconds with Virtual Machine active

Bandwidth saturation during a single machine migration.



SMALL FILE BENCHMARK

In order to test the incidence of latency on the files system we increased the latency between the two sites up to 37ms. It was possible by changing the topology of the end-to-end circuit over the GARR-X infrastructure.

For the resilience test we used smallfile as benchmark, executed in three different configurations

- Local File System
- Gluster file system 5ms
- Gluster file system 37ms

SMALL FILE BENCHMARK

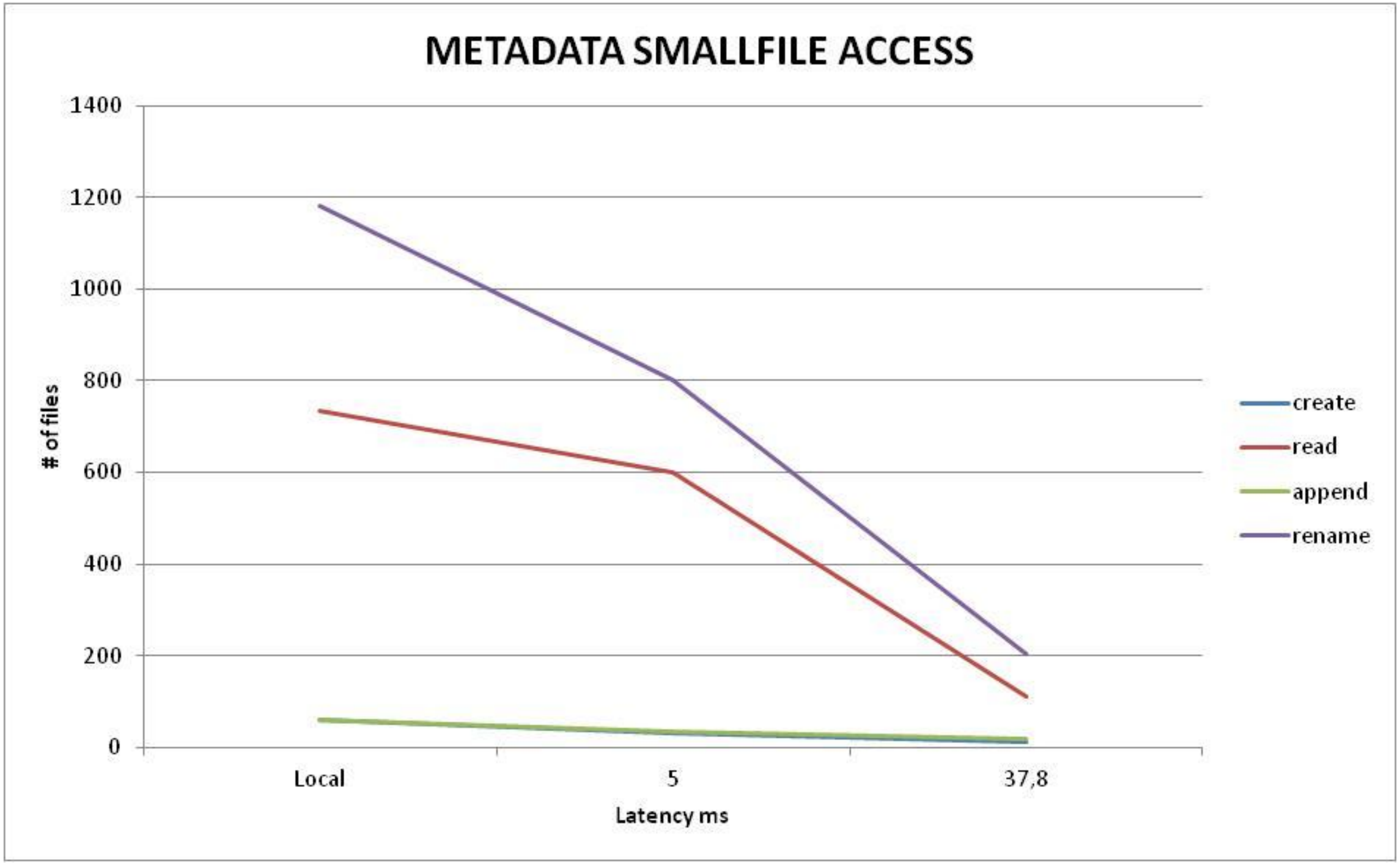
Small file is a benchmark stressing the metadata performance rather than the IO performance of a file system.

It creates large set of smallfiles (few kb each one) and measures the performance on

- Create
- Read
- Rename
- Append

For additional information see:
https://github.com/bengland2/smallfile

SMALL FILE BENCHMARK



FUTURE TEST

- Test the full infrastructure varying the latency and individuating the critical thresholds for the service point of view.
- Test and comparison of performances with additional distributed file-systems (CEPH, Gluster 3.5)
- Implementation of a federated Cloud infrastructure through the feature of Openstack Cells
- Implement a set of real services in HA configuration over the distributed infrastructure.
- Expand the experience with a multisite setup, involving additional INFN sites (MPLS)

CONCLUSION

- The ATLAS Cloud Computing R&D has been able to demonstrate the feasibility to integrate transparently various cloud resources into the PanDA workload management system.

- Experiences with commercial Clouds have also demonstrated the feasibility to implement hybrid cloud with “elastic” part in outsourcing.

- Tests on Cloud infrastructure deployed over a geographic layer 2 link, have demonstrated the possibility to expand Cloud in a high latency environment without impact on stability or resilience. It allows to design new scenarios of high availability services moving Virtual Machines from a site to an other in the same Cloud Infrastructure.

- The Cloud Computing in action has demonstrated to open a range of new opportunities for the High Energy Physics experiments.