

# ***Current Cloud infrastructure at CESGA***

**Esteban Freire, Iván Díaz, , Alejandro Feijoo, Javier López Cacheiro,  
Pablo Rey and Carlos Fernández**

**Big Data Platforms meeting, 15/01/2015**

# Outline



- Introduction
  - CESGA Cloud Infrastructure
- EGI FedCloud Infrastructure
- Hadoop
- New Big Data Platform at CESGA

# Introduction



- CESGA site must adapt and react to very complicated requirements.
- Necessary to economize on energy consumption.
- Our collaboration on different projects (EGI FedCloud, Hadoop...) requires new testing and verification resources, which must be assigned flexibly depending on demand.

# Cloud Infrastructure



## 36 HP ProLiant SL2x170z G6 (Virtualization Nodes)

Processor	2 x INTEL Processor E5520 (2.26 GHz, 8MB Cache, 4 cores per processor)
Memory	16 GB
Storage	470GB
Networking	Embedded 2 Ports Intel 1Gb

## 6 HP ProLiant DL180 G6 (Server & Storage)

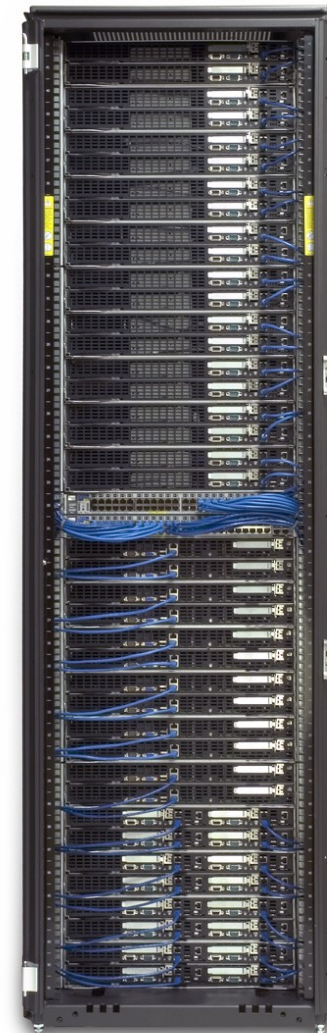
Processor	2 x INTEL Processor E5520 (2.26 GHz, 8MB Cache, 4 cores per processor)
Memory	16 GB
Storage	5700GB (5 nodes) / 2000GB (1 nodes)
Networking	Embedded 2 Ports Intel 1Gb (6 nodes) / 2 Ports NetXen 10G (4 nodes)

## HP ProLiant SL2x160 G6 (Server)

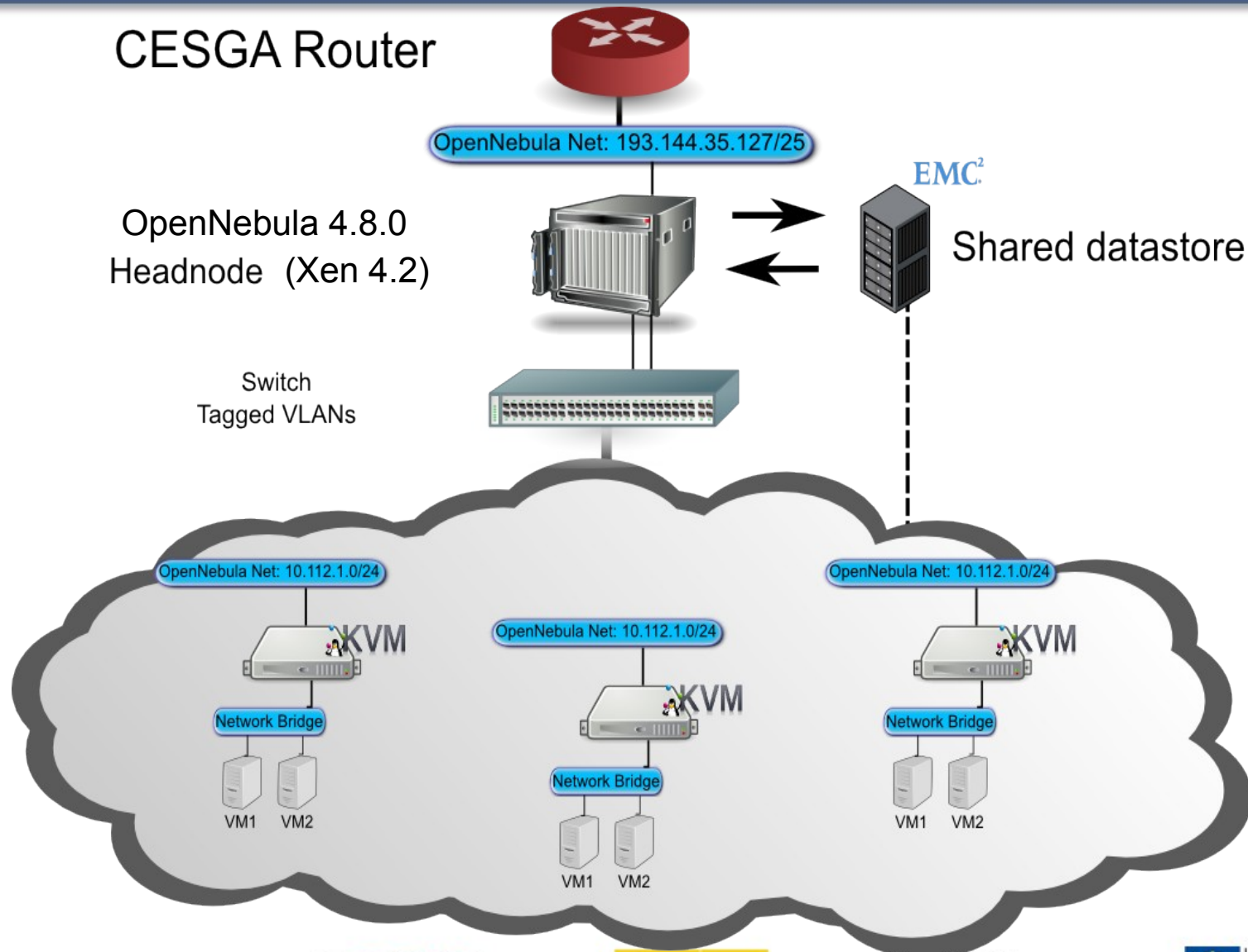
Processor	2 x INTEL Processor E5504 (2 GHz, 4MB Cache, 4 cores per processor)
Memory	32 GB
Storage	1200 GB
Networking	Embedded 2 Ports Intel 1Gb

## HP ProLiant SL2x165 G6 (Server)

Processor	2 x AMD Processor 2435 (2.6 GHz, 6 cores per processor)
Memory	32 GB
Storage	1200 GB
Networking	Embedded 2 Ports Intel 1Gb



# Cloud Infrastructure



# EGI FedCloud Infrastructure



- CESGA is collaborating as resource provider and developer within EGI FedCloud Task Force.
- EGI FedCloud auth its based on x509 authentication and OGF OCCl 1.1.5-1 (rOCCl server)
  - ON OCCl service is not OGF compliant...
- Virtual Organizations and users management is based on PERUN service developed by CESNET.

# Hadoop



- CESGA is providing a testing Hadoop infrastructure for local users thanks to ON framework.
- Hadoop is an emerging technology that could benefit from existing developments in cloud.
- It represents a real challenge to the Fedcloud infrastructure due to its heavy usage of disk and network I/O.
- Running Big Data calculations over federated clouds is attracting the interest of the research community.

# Hadoop



- This service was already tested using Fedcloud infrastructure. The most recent publications about it can be found on the following links:

## **“Hadoop analytics provisioning based on a virtual infrastructure”,**

<https://indico.egi.eu/indico/contributionDisplay.py?contribId=1&sessionId=13&confId=2160>

## **“Leveraging EGI Federated Cloud Infrastructure for Hadoop analytics”,**

<http://ocs.editorial.upv.es/index.php/IBERGRID/IBERGRID2014/paper/viewFile/192/102>



# Hadoop



- FedCloud is especially suitable for:
  - small and medium-size Hadoop jobs.
  - where the data set is already pre-deployed in HDFS.
- Optimized startup times are very close to those obtained in Amazon EC2.
- The initial data set deployment (put time) imposes a large overhead.
- Scalability: we were able to run the TeraGen benchmark up to 2TB and the TeraSort benchmark up to 500GB.

# New Big Data Platform



- We will acquire a new big data platform during this year. It should be installed and configured by the end of July 2015.

# New Big Data Platform



- We will acquire a new big data platform during this year. It should be installed and configured by the end of July 2015.

# Questions?



## Thank you for your attention!

# Questions?

