



SCIENTIFIC CLOUD COMPUTING IN TORINO

Stefano Bagnasco, Stefano Lusso, Sara Vallero *et al.*

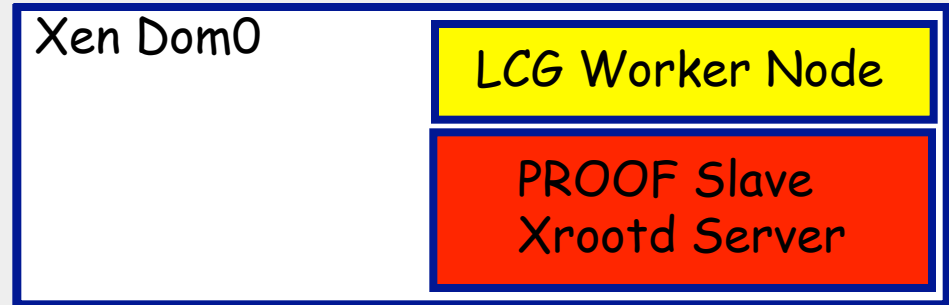
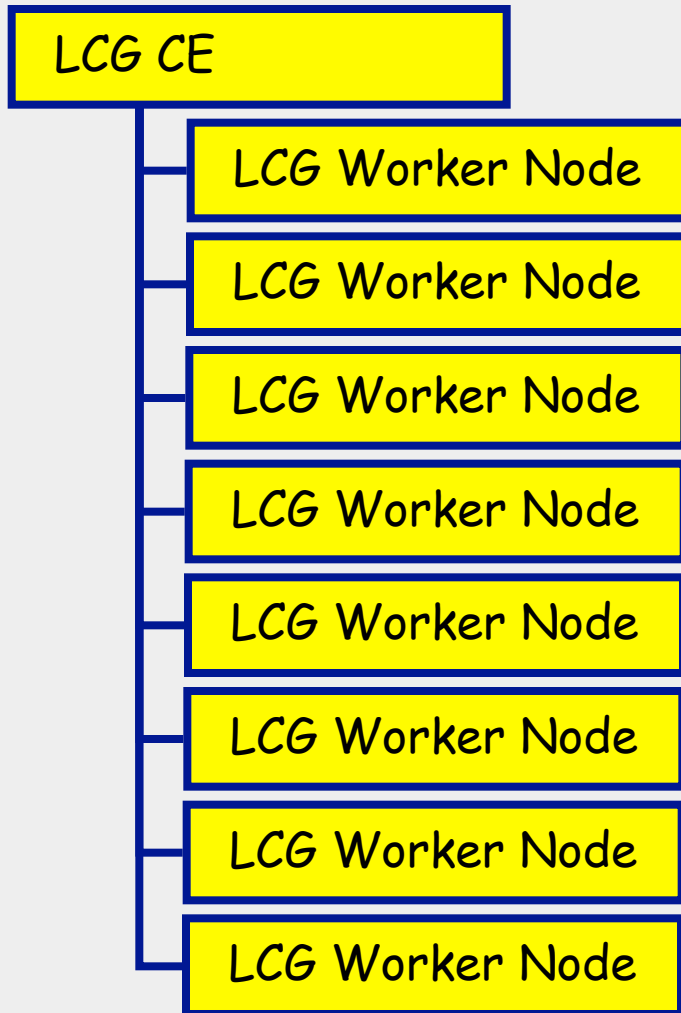
ALICE Tier-1/Tier-2 Workshop | Torino, Feb 23-25, 2015

- About 2K CPU cores, up to ~1800 job slots
 - (This includes servers)
 - Of which 200 “reserved” for BES-III on a separate instance
- 10Gb/1Gb internal network
- 10 Gb/s connectivity to GARR network
 - Can be easily upgraded if needed
- About 1.8 PB overall of disk space
- What is sadly missing is manpower...
 - Very dedicated, very competent and very small team...

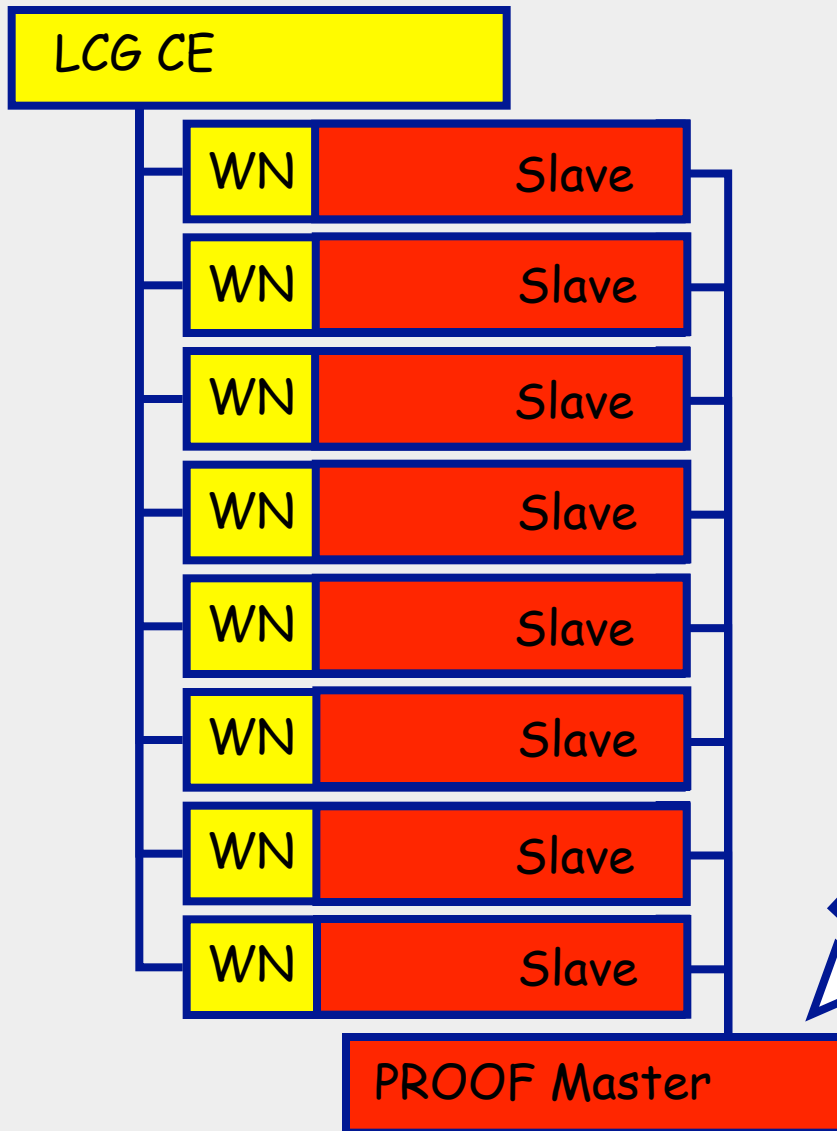


(Circa 2008)

2007: VIRTUAL PROOF CLUSTER



- **Xen can dynamically allocate resources to either machine**
 - Both memory and CPU scheduling priority!
 - Memory is the issue, CPU priority limit is enough
 - Normal operation: PROOF slaves are “dormant” (minimal memory allocation, very low CPU priority)
 - Interactive access: dynamically increase resources to the PROOF instances, job on WN slows down
 - Alternatively, “wake up” more slaves



Advantages

- Grid batch job on the WN ideally never completely stops, only slows down
- Non-CPU-intensive I/O operations can go on and do not timeout
- Both environments are sandboxed and independent, no interference



VAF prototype

S. Bagnasco
D. Berzano

Introduction

The Xen
approach

Xen feasibility
benchmarks

The Prototype

Prototype
benchmarks

Outlook

Thanks!

A prototype of a dynamically expandable Virtual Analysis Facility

S. Bagnasco, D. Berzano *et al.*
INFN Torino

ACAT, Erice – November 5th, 2008



Prototype benchmarks

Grid jobs CPU usage with different resources

VAF prototype

S. Bagnasco
D. Berzano

Introduction

The Xen
approach

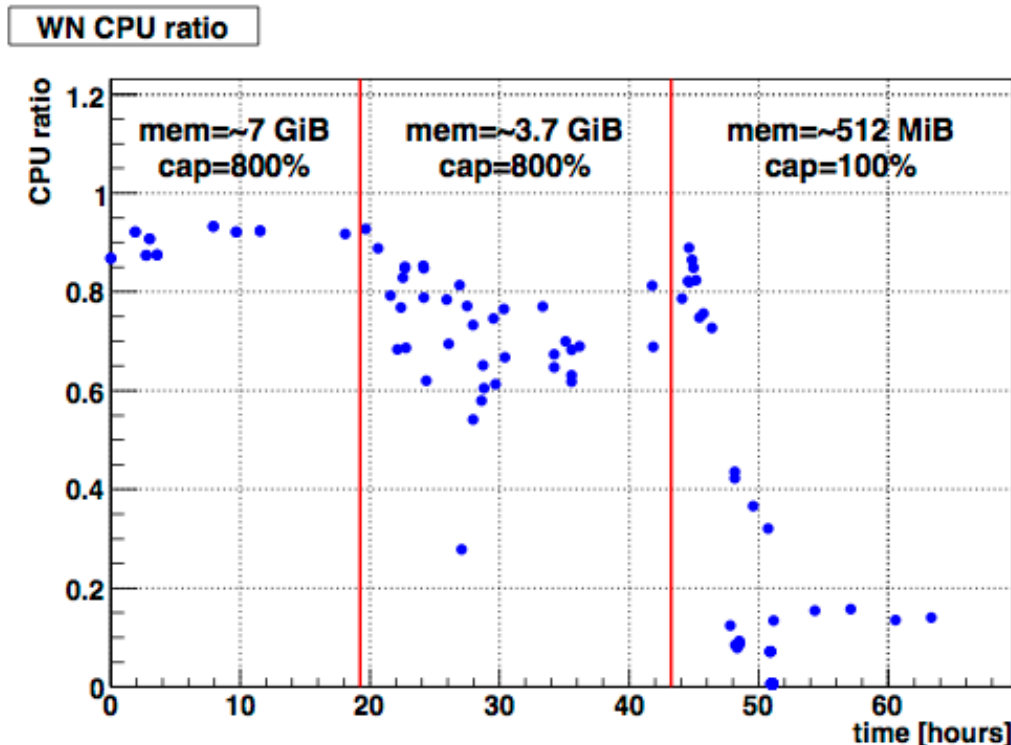
Xen feasibility
benchmarks

The Prototype

Prototype
benchmarks

Outlook

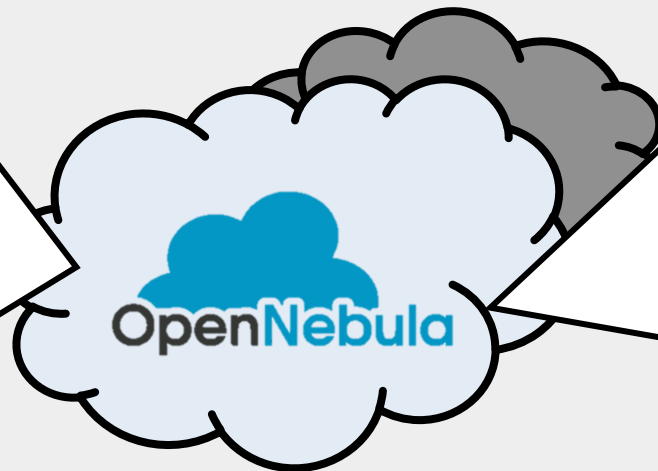
Thanks!



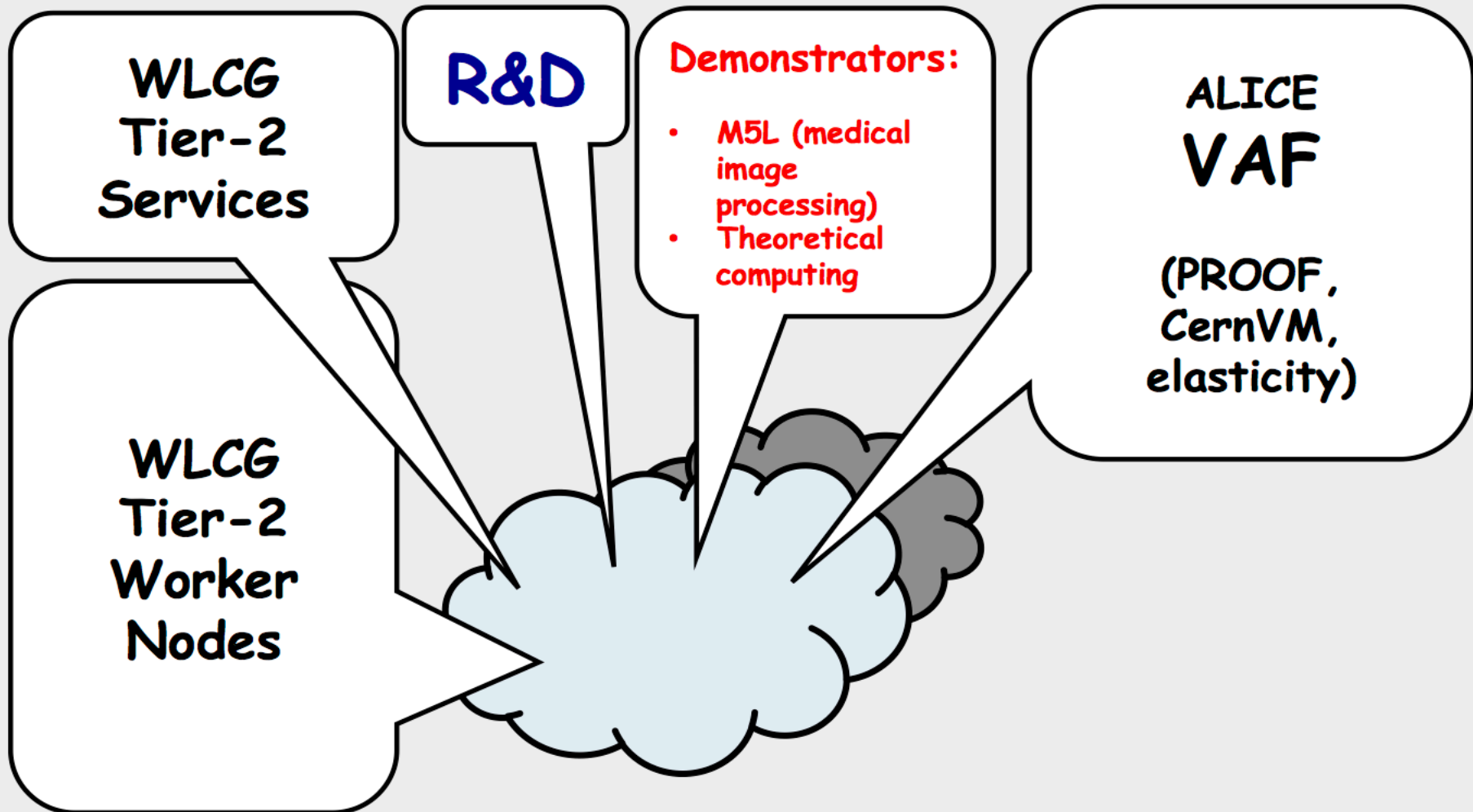
- As swap usage increases, jobs turn from CPU-bound into I/O-bound
- No more job failures wrt average \Rightarrow they slow down but don't crash

13

WLCG
Tier-2
Worker
Nodes



ALICE
Analysis
Facility
(PROOF)





VMs providing **critical services**:

- Run on a cluster of server-class redundant hypervisors
- Public & private IP
- Shared system disks on resilient storage allowing live migration (Services need to run continuously)

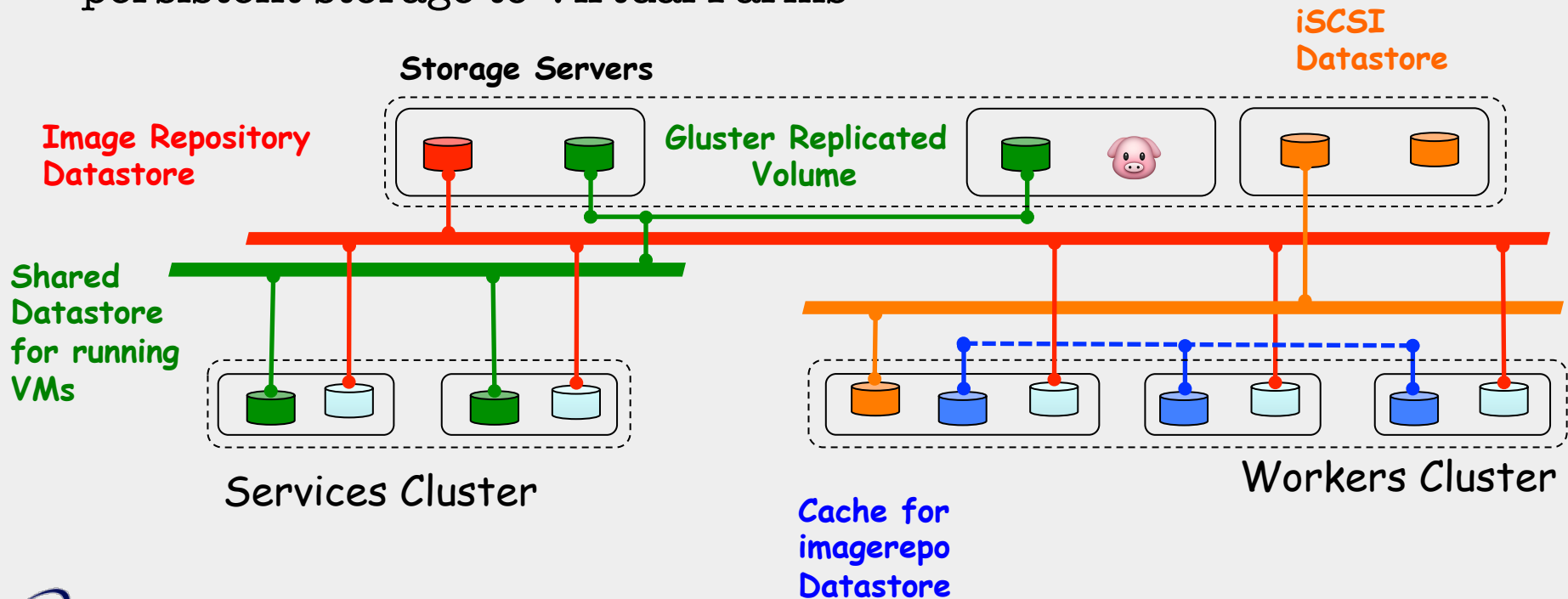


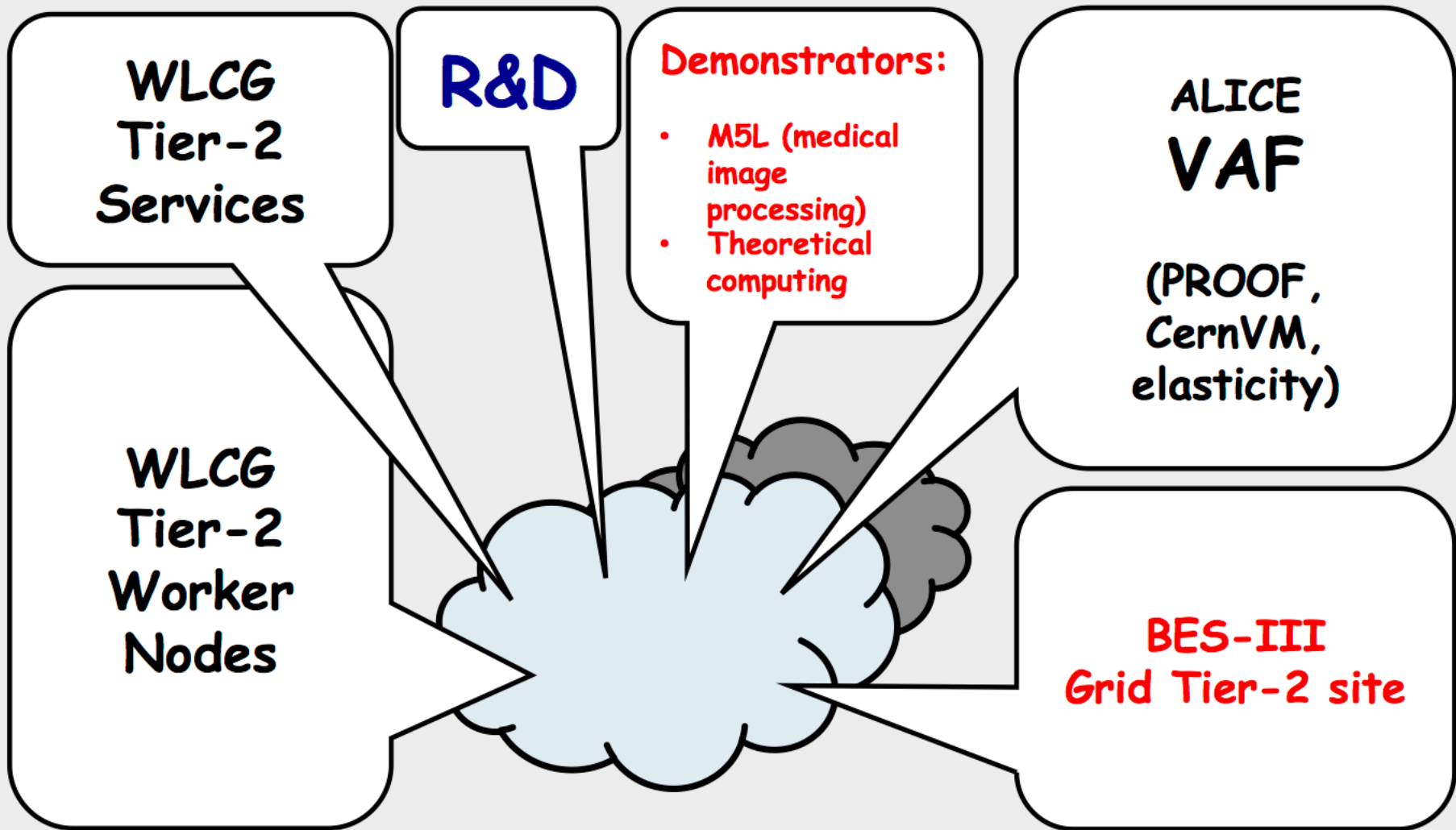
VMs providing **computing workforce**:

- Run on a cluster of compute-optimized, less expensive hypervisors
- Locally cached image repository for fast startup (Workers are often reallocated)
- Access to fast storage for data
- Private network only

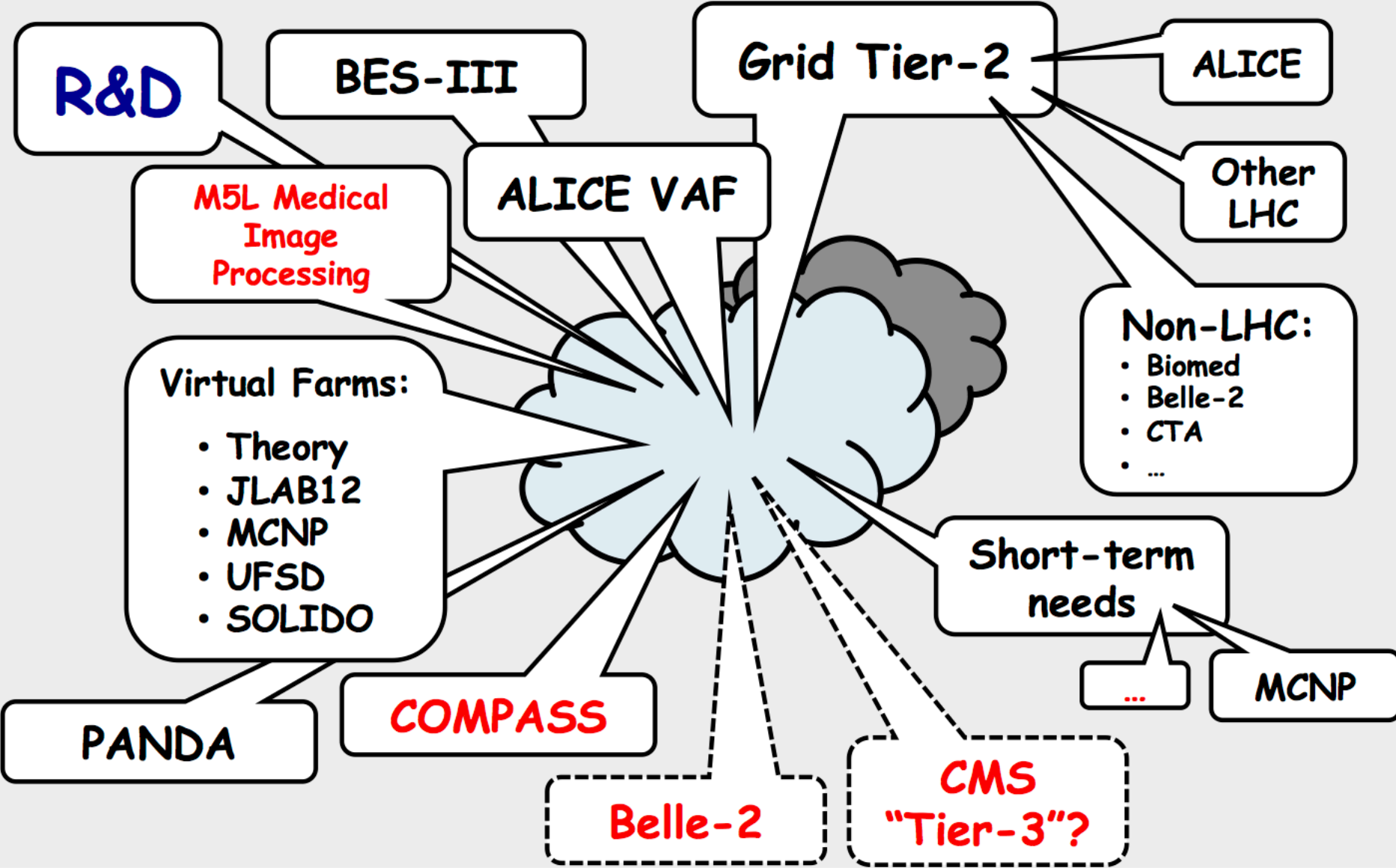
UNDER THE HOOD 2

- **Services System Datastore** is **shared** to allow live migration
- **Workers System Datastore** is **local** to the hypervisors to increase I/O capacity. Images repository is locally cached on each hypervisor to reduce startup time
- **Persistent Space Datastore** is mounted on the relevant hypervisors using the **iSCSI** Transfer Manager to provide persistent storage to Virtual Farms

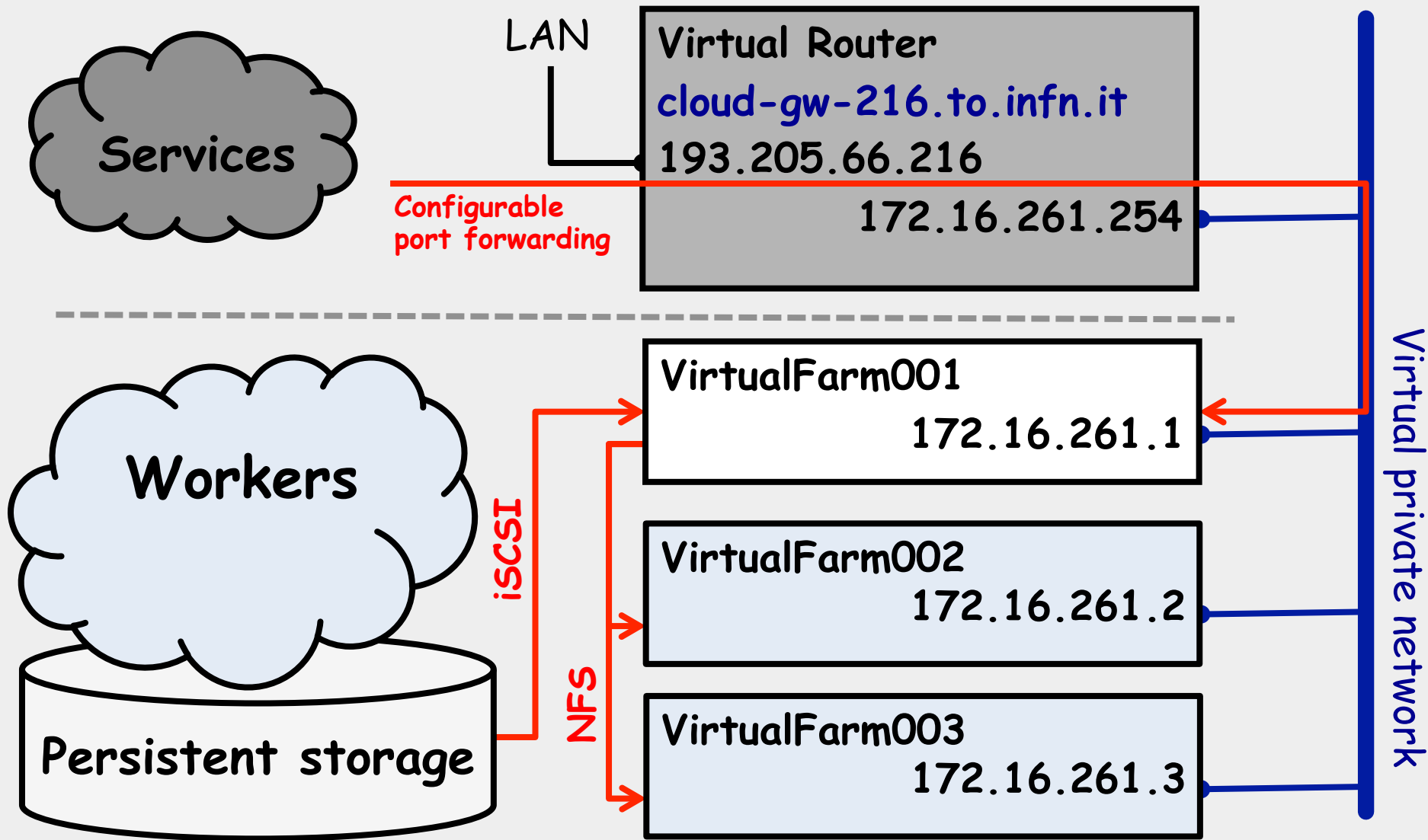




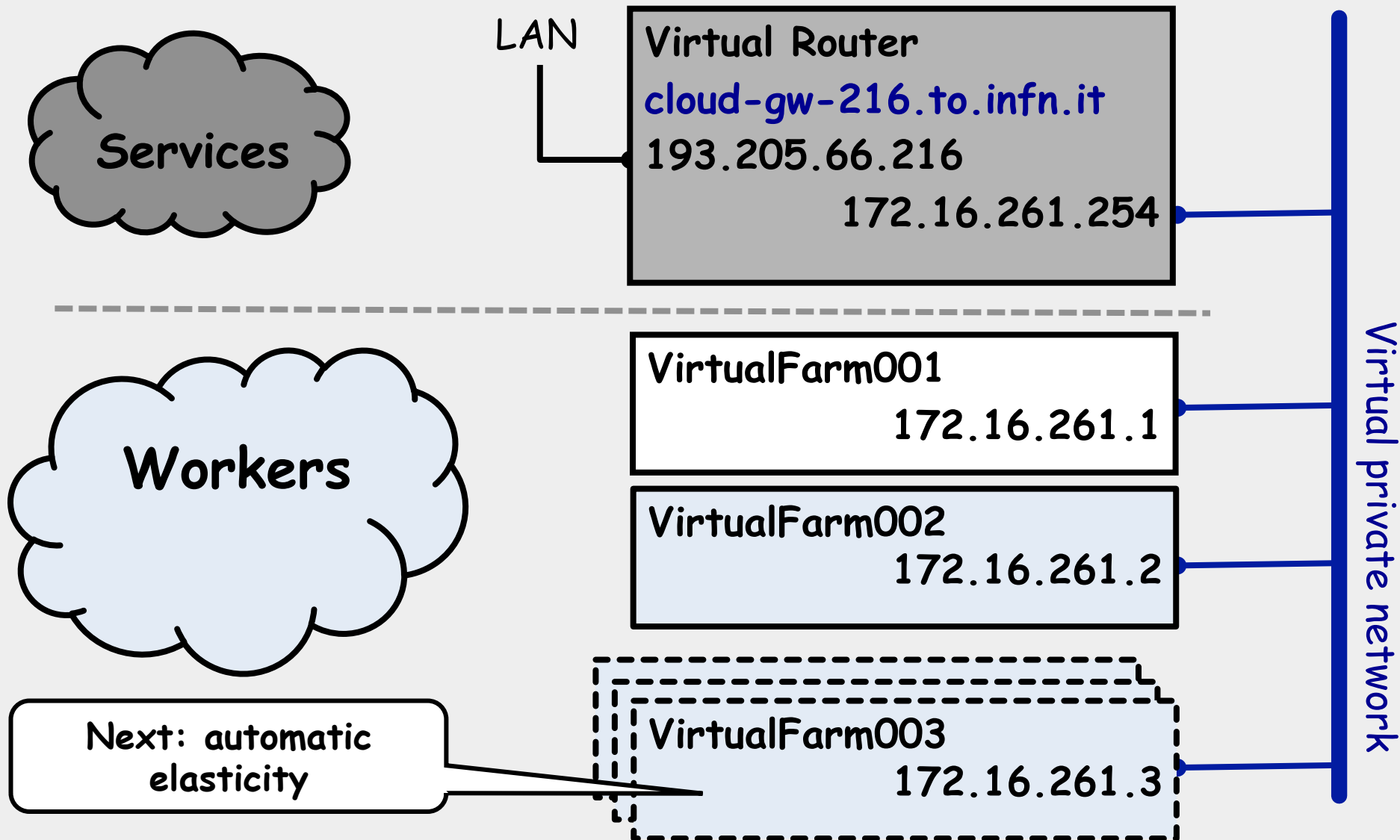
...AND NOW!



VIRTUAL FARM PROVISIONING MODEL



VIRTUAL FARM PROVISIONING MODEL



- In the context of the **STOA-LHC PRIN** project:
 - Monitoring & accounting of heterogeneous resources
 - Automatic elasticity of virtual clusters
 - More: see Sara's talk
- In the contest of the upcoming **INDIGO-DataCloud** EU project:
 - Advanced cloud scheduling
 - PaaS tools for scientific computing
- Other projects
 - Open City Platform

Major overhaul and redesign of the infrastructure this year

- This was one of the very first Cloud infrastructures to reach production maturity in our community
 - Still suffers from some design limitations
- After a number of years we learned a lot...
 - Users will never relinquish a resource by themselves even if they're not using it. Never.
 - OpenNebula core DB is touchy and need a robust infrastructure
- ...and also the tools evolved
 - e.g. for virtual networking
 - but we will stick to OpenNebula!

Extend and improve the Virtual Farm provisioning model

- Storage space provisioning still rudimentary
- Monitoring and accounting
 - ...which may translate into billing at some point!
 - see also Sara's talk
- Support more use cases

C3S: THE NEXT CHALLENGE

- The University of Torino won a 900 kEUR grant by Compagnia di San Paolo to build an inter-departmental scientific computing facility
- INFN-Torino is a partner in the project and will host the facility
- The design of the facility is a common project between INFN, and the Departments of Physics and Computer Science

C3S: THE NEXT CHALLENGE

- The “Centro di Competenza per il Calcolo Scientifico” (C3S) will serve both as a production and an R&D facility
 - Diverse and manifold infrastructure: conventional nodes, “fat” nodes with 4-way servers and huge memory, GPU nodes
 - High-bandwidth, low-latency interconnection
 - Two-tier storage
 - Will share the physical infrastructure with the existing cloud facility
- Will also help building a forum for scientific computing-related activities, and be a platform on which to host further projects
 - Platform for exploring new technologies (e.g. low power CPUs, non-GPU many-core,...)

Our ideas:

- Manage the “conventional” part of the facility with a cloud infrastructure like the existing one
 - But this is a true HPC cluster!
- Explore the opportunity of managing heterogeneous resources with a single tool
 - e.g.: GPU virtualization? Can containers help?
- Experiment with interoperability and cloudbursting
- ...of course, use the resources for ALICE whenever possible according to shares
 - For example, to test ITS reconstruction software for Run 3

- The Torino site is stable and growing
 - The overhaul of the Cloud infrastructure is to be funded (among others) by INFN's Computing Committee
- We have several R&D activities ongoing
 - External projects to fund the manpower
 - Coordination with other efforts is mandatory
- The C3S project is both a challenge and an opportunity
 - Will help to fund much needed infrastructural work
 - Will help (hopefully) to attract students and people from other communities

- Questions?