Northeastern University
College of Computer and Information Science

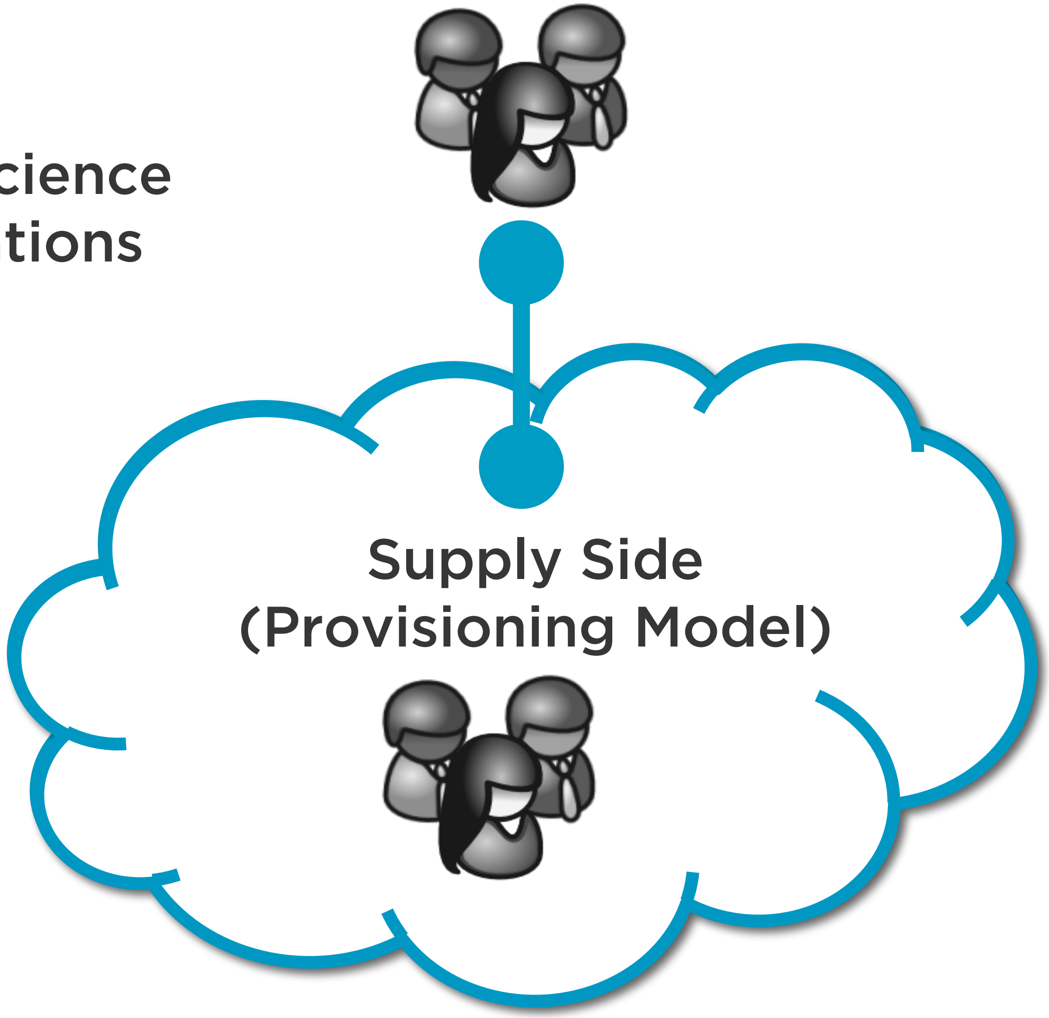
Bringing Private Cloud Computing to HPC and Science

January 31st, 2017

Dr. Ignacio M. Llorente
UCM Professor, Harvard Visiting Scholar, and OpenNebula Director

Demand Side
(Consumption Model)

HPC & Science
Applications



Supply Side
(Provisioning Model)

Contents

Building Private Cloud Computing to HPC and Science

The Anatomy of the Cloud

The Private HPC Cloud Use Case

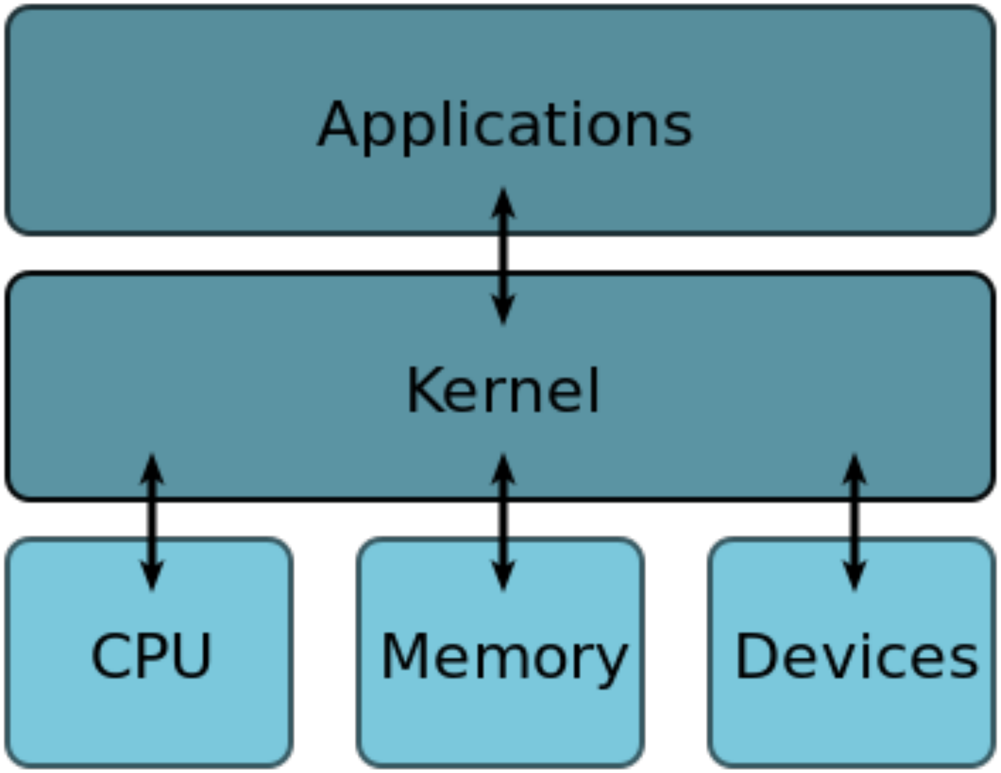
Main Challenges for Private HPC Cloud

Private HPC Cloud Case Studies

The Anatomy of the Cloud

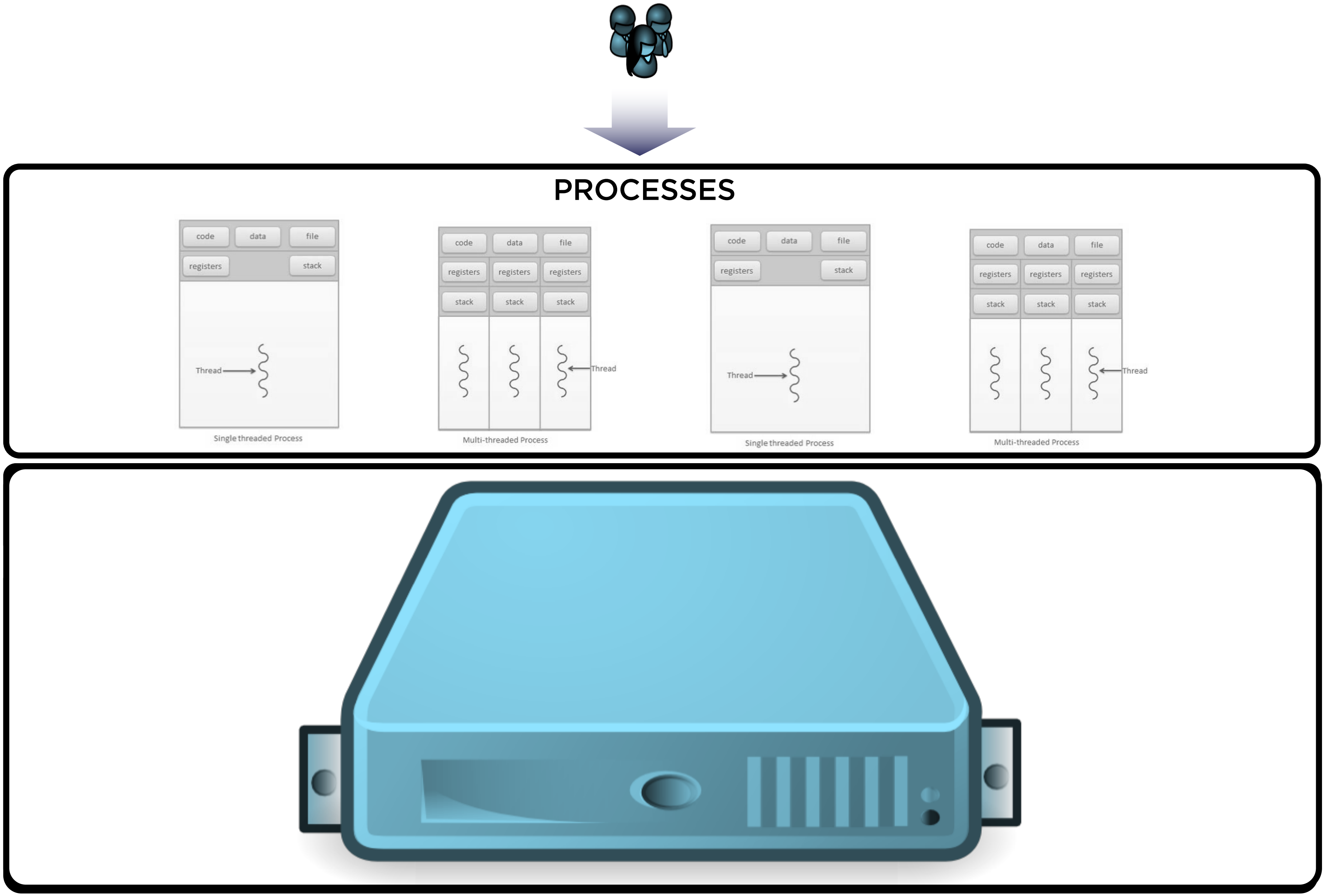
What is an Operating System?

“An operating system (OS) is system software that manages computer hardware and software resources and provides common services for computer programs” (source: Wikipedia)



The Anatomy of the Cloud

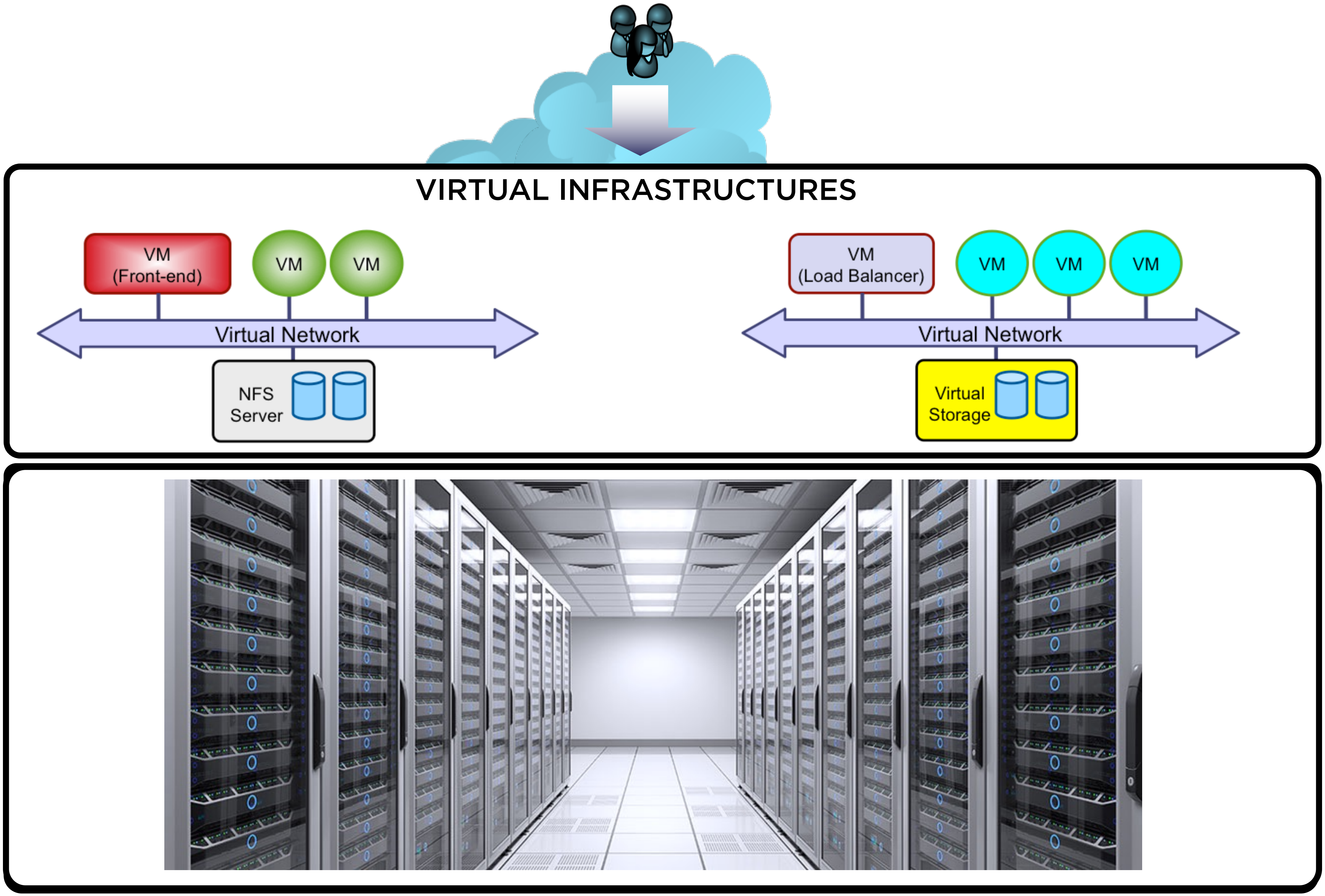
What is an Operating System?



PROCESSES

The Anatomy of the Cloud

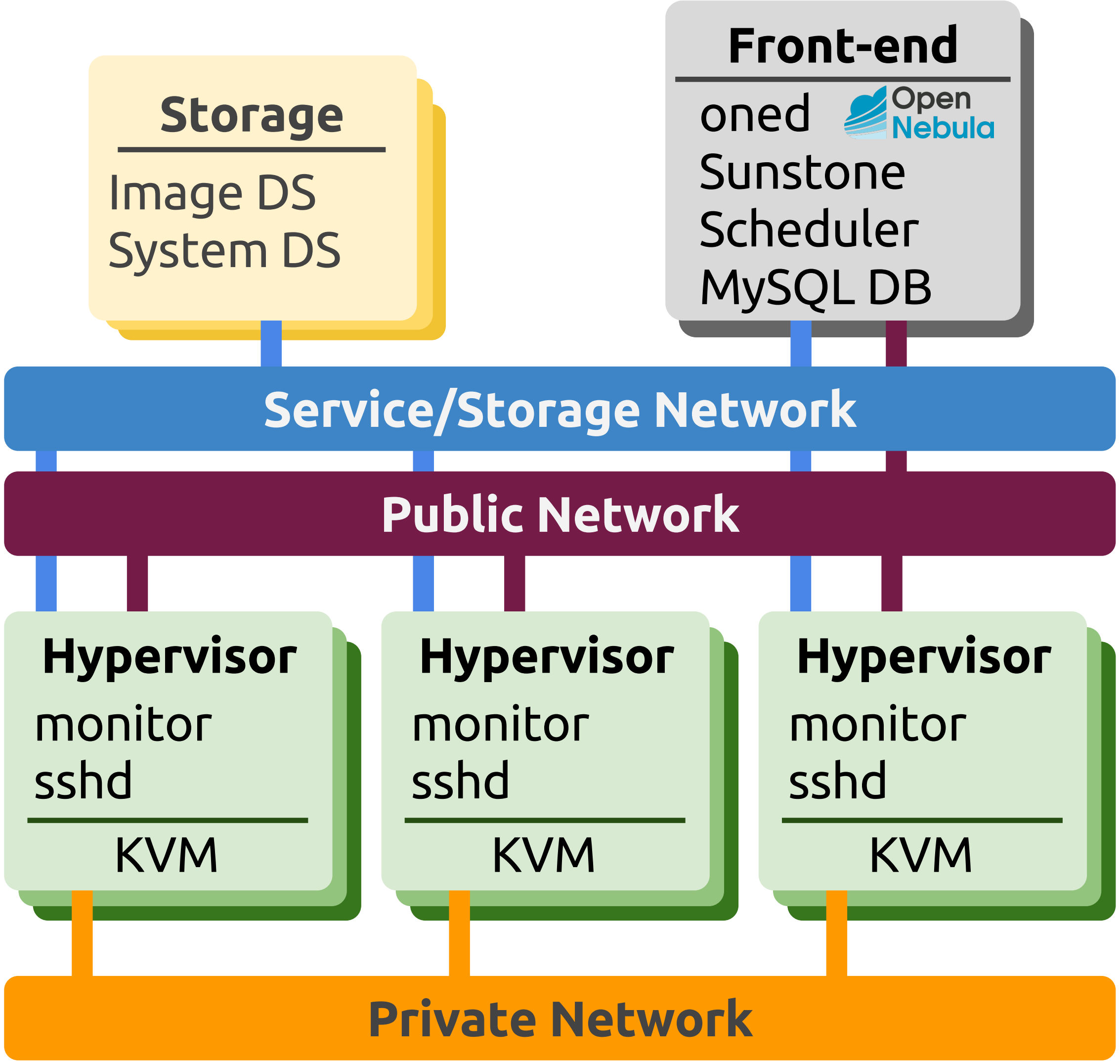
What is a Cloud Management Platform?



VIRTUAL INFRASTRUCTURES

The Anatomy of the Cloud

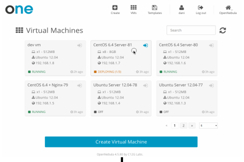
The Internals of a Cloud Instance



Dr. Ignacio M. Llorente

The Anatomy of the Cloud

Zone Federation



Virtual Machines

dev vm
x1 - 512MB
Ubuntu 12.04
192.168.1.8
RUNNING 2h ago

CentOS 6.4 Server-81
x8 - 8GB
Ubuntu 12.04
192.168.1.7
DEPLOYING (1/3) 3h ago

CentOS 6.4 Server-80
x1 - 512MB
Ubuntu 12.04
192.168.1.6
RUNNING 3h ago

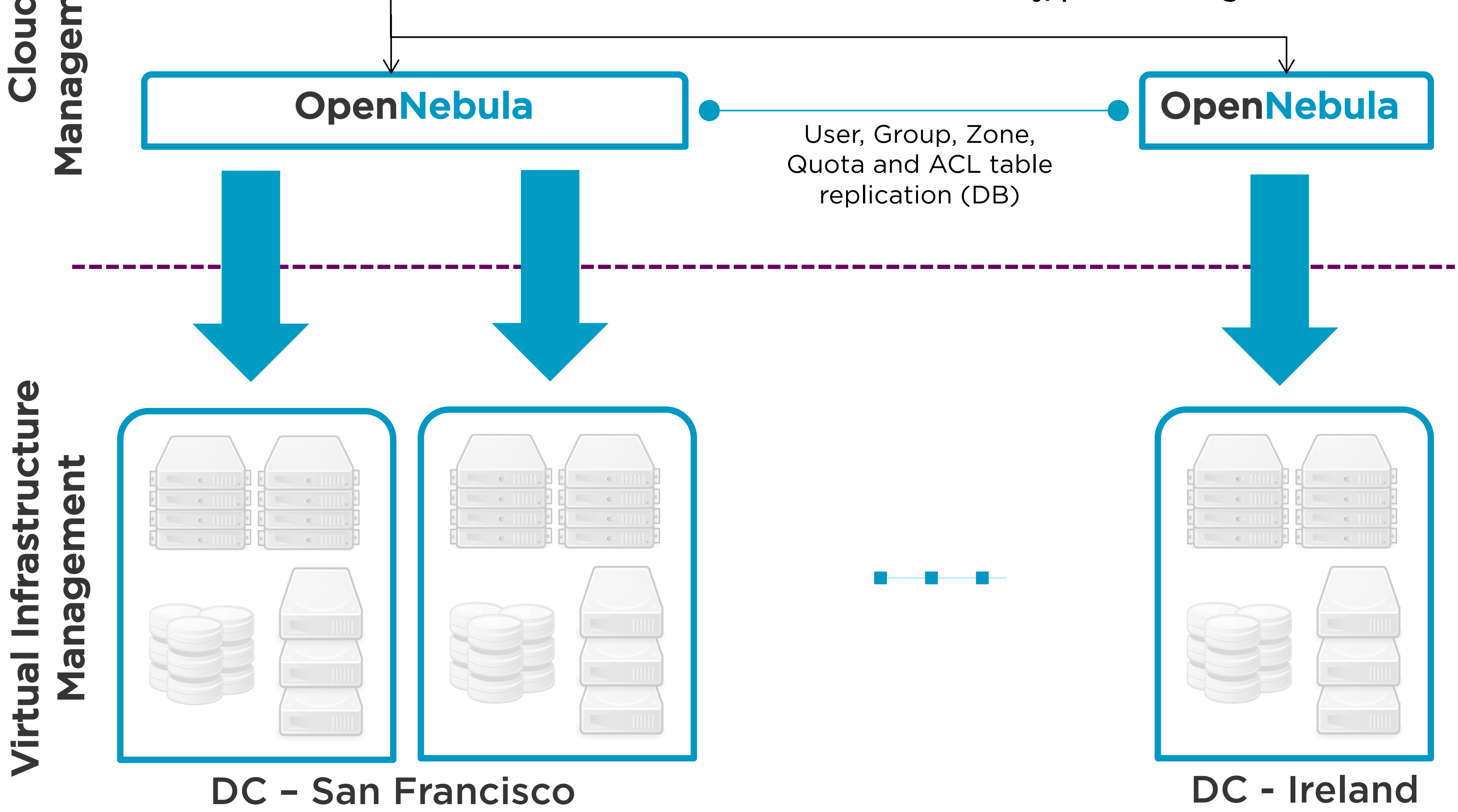
CentOS 6.4 + Nginx-79
x1 - 512MB
Ubuntu 12.04
192.168.1.5
RUNNING 3h ago

Ubuntu Server 12.04-78
x1 - 512MB
Ubuntu 12.04
192.168.1.4
OFF 3h ago

Ubuntu Server 12.04-77
x1 - 512MB
Ubuntu 12.04
192.168.1.3
OFF 3h ago

OpenNebula 4.5.80 by C12G Labs.

- Centralized multi-tenancy (quotas, groups…)
- Simple cloud GUIs and interfaces
- Service elasticity/provisioning

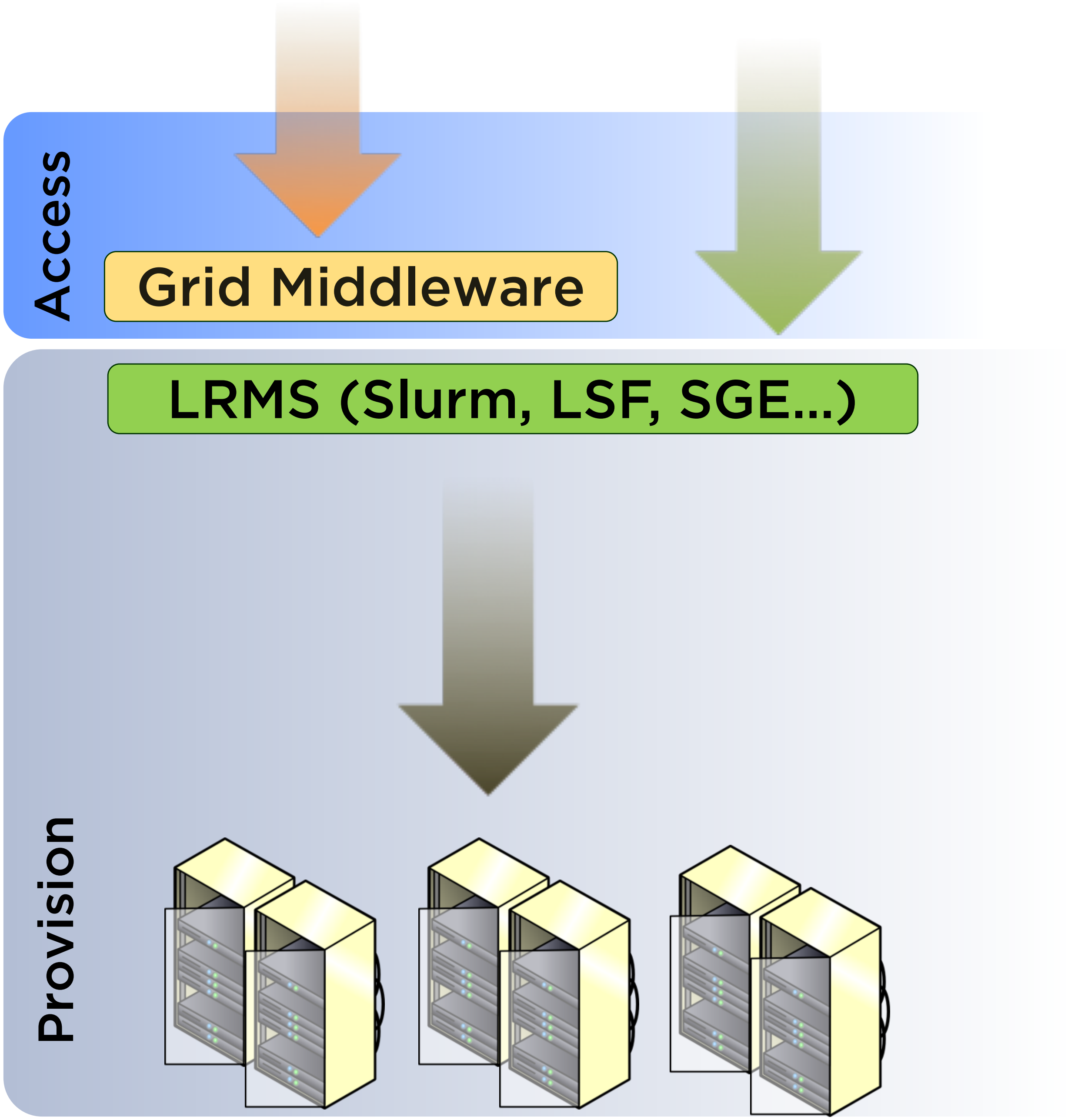


User, Group, Zone, Quota and ACL table replication (DB)

Dr. Ignacio M. Llorente

The Private HPC and Science Cloud Use Case

The Pre-cloud Era



Access

Grid Middleware

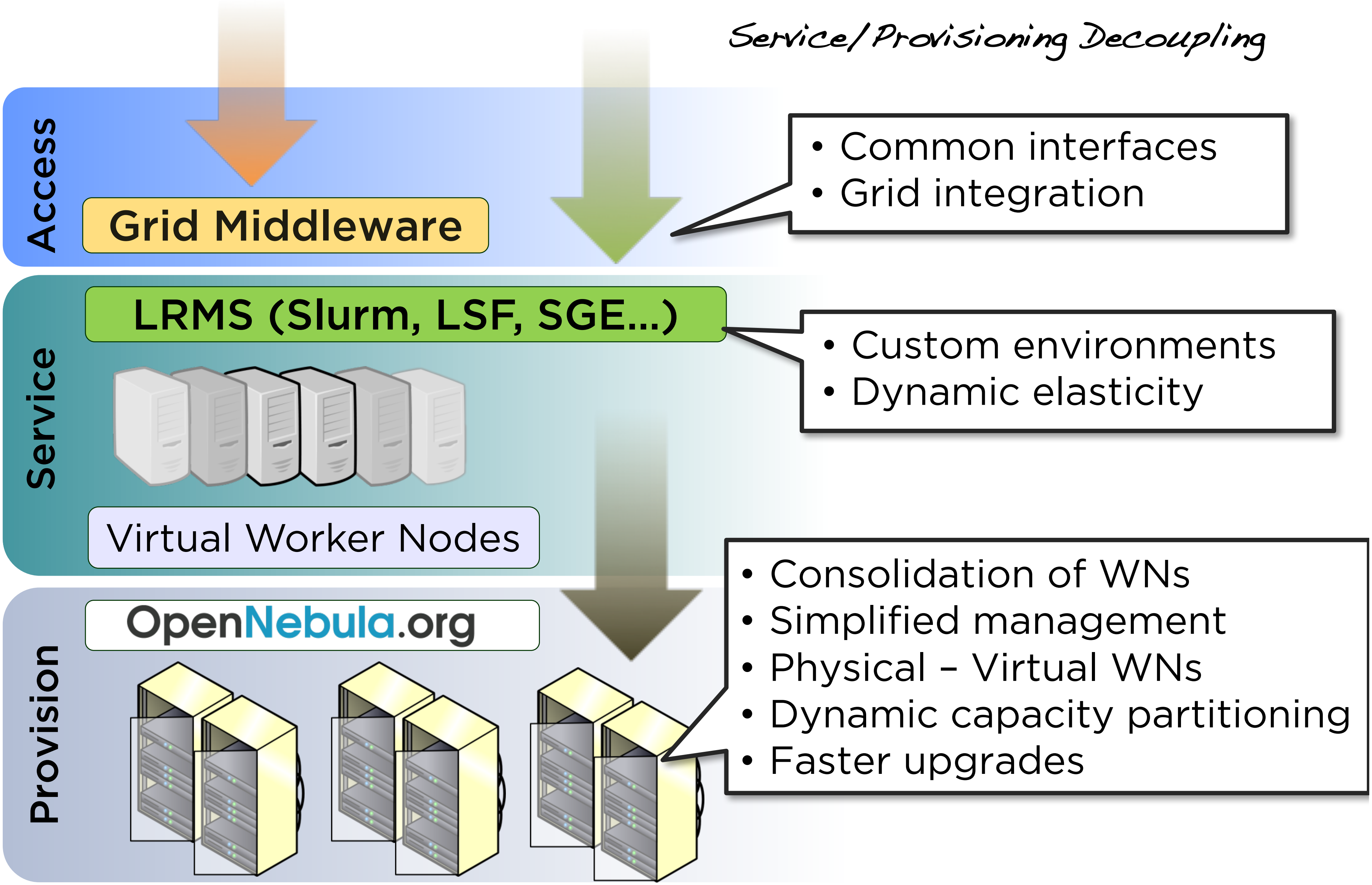
LRMS (Slurm, LSF, SGE…)

Provision

The Private HPC and Science Cloud Use Case

OpenNebula as an Infrastructure Tool – Enhanced Capabilities

Service/Provisioning Decoupling



Access

Grid Middleware

- Common interfaces
- Grid integration

Service

LRMS (Slurm, LSF, SGE...)

Virtual Worker Nodes

- Custom environments
- Dynamic elasticity

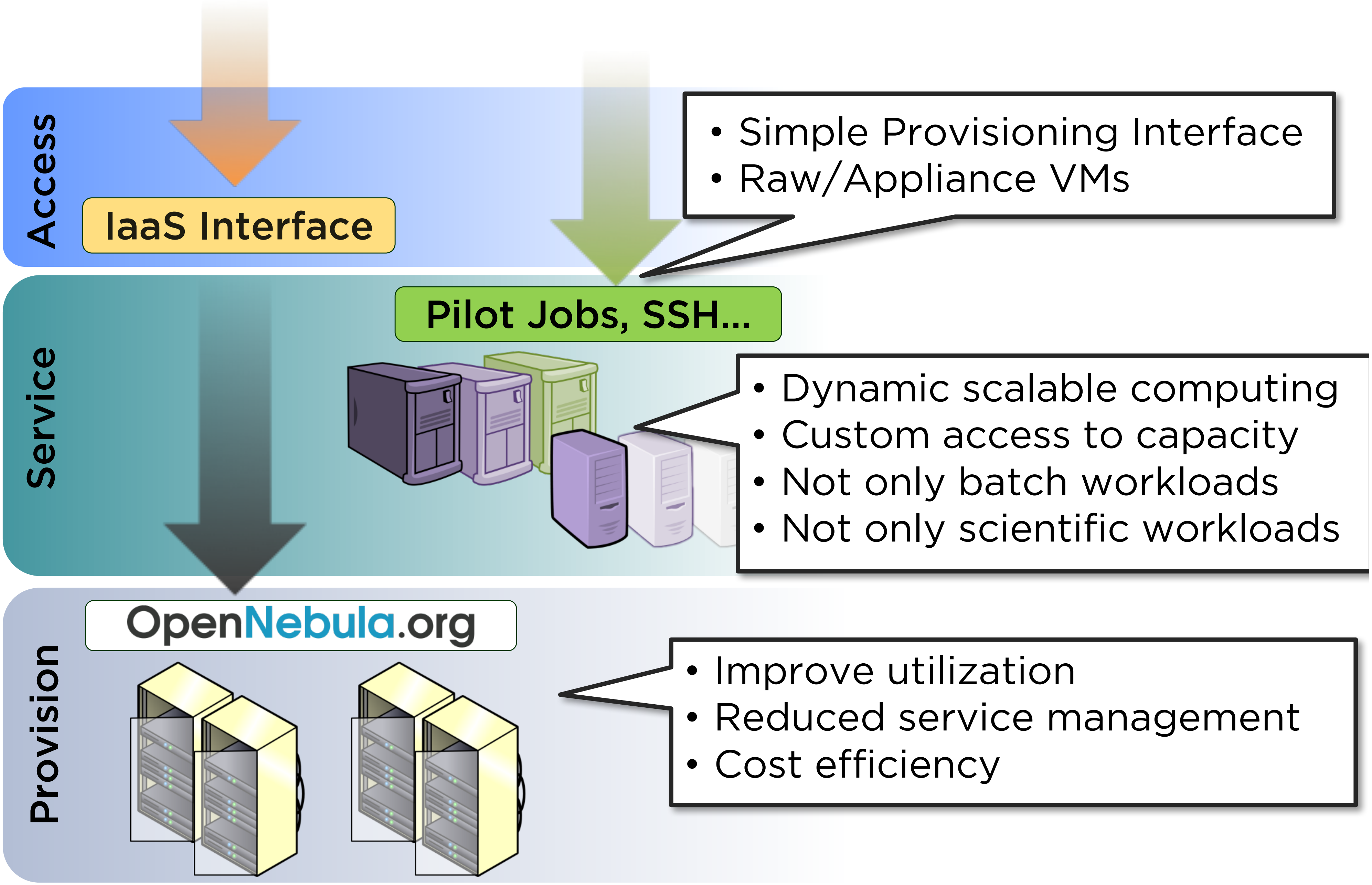
Provision

OpenNebula.org

- Consolidation of WNs
- Simplified management
- Physical – Virtual WNs
- Dynamic capacity partitioning
- Faster upgrades

The Private HPC and Science Cloud Use Case

OpenNebula as an Provisioning Tool – Enhanced Capabilities



- Simple Provisioning Interface
- Raw/Appliance VMs

- Dynamic scalable computing
- Custom access to capacity
- Not only batch workloads
- Not only scientific workloads

- Improve utilization
- Reduced service management
- Cost efficiency

Dr. Ignacio M. Llorente

The Private HPC and Science Cloud Use Case

Example: Research Computing at Harvard

FASRC

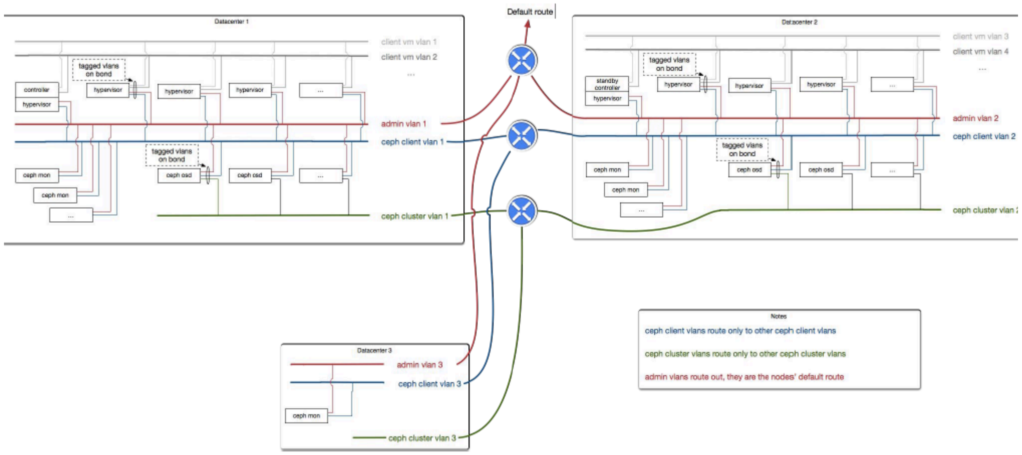
1. Core Cloud

- Production services in HA

Open Nebula

ceph

KVM



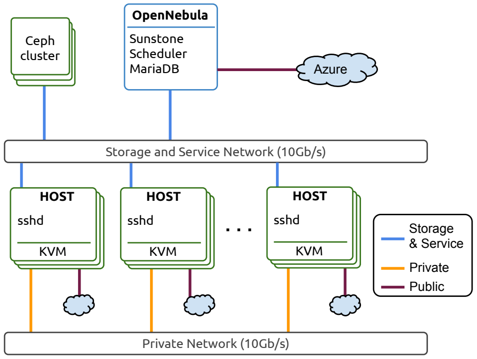
ceph client vlans route only to other ceph client vlans

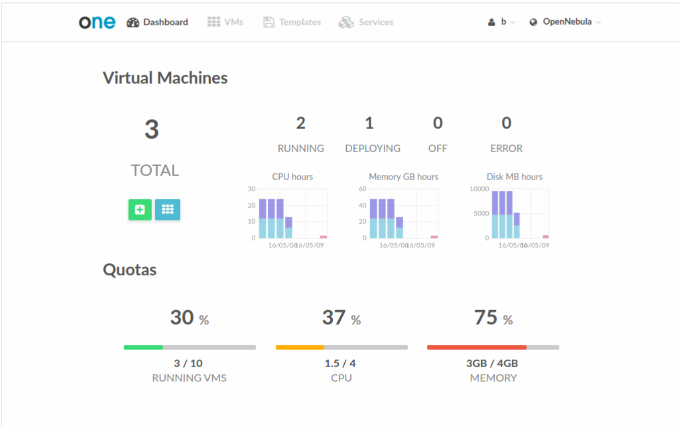
ceph cluster vlans route only to other ceph cluster vlans

admin vlans route out, they are the nodes' default route

2. Research Computing Cloud

- Self-service portal for science apps





Virtual Machines

3 TOTAL

2 RUNNING

1 DEPLOYING

0 OFF

0 ERROR

Quotas

30 % 3 / 10 RUNNING VMS

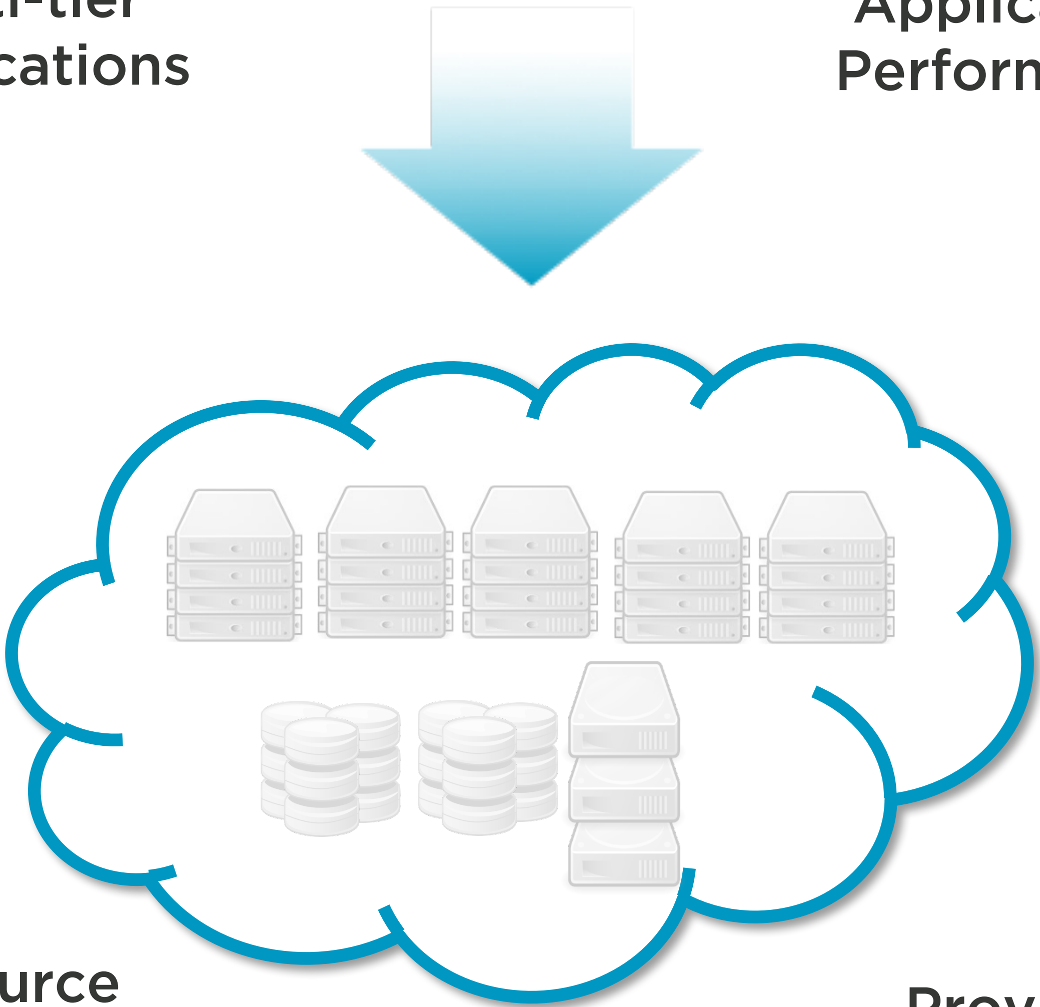
37 % 1.5 / 4 CPU

75 % 3GB / 4GB MEMORY

Dr. Ignacio M. Llorente

Main Challenges for Private HPC Cloud

Main Demands from Engineering, Research and Supercomputing



Multi-tier Applications

Application Performance

Resource Management

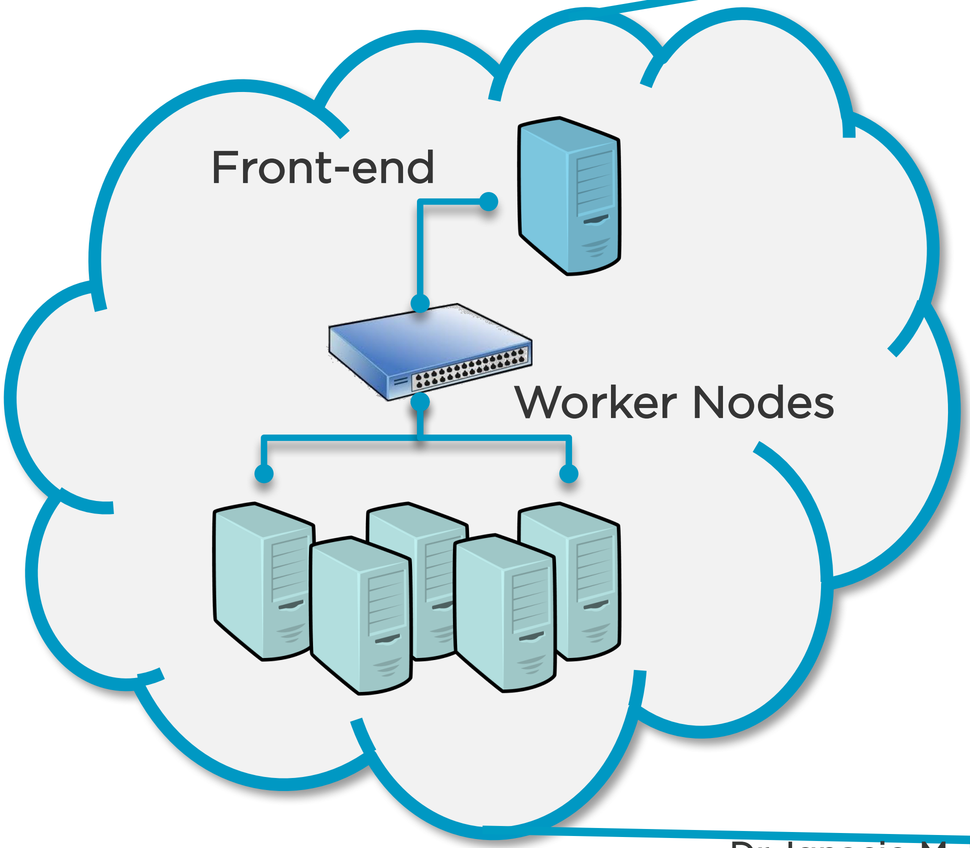
Provisioning Model

Main Challenges for Private HPC Cloud

Execution of Multi-tiered Applications

Requirements from Complex Applications

- Several tiers
- Deployment dependencies between components
- Each tier has its own cardinality and elasticity rules



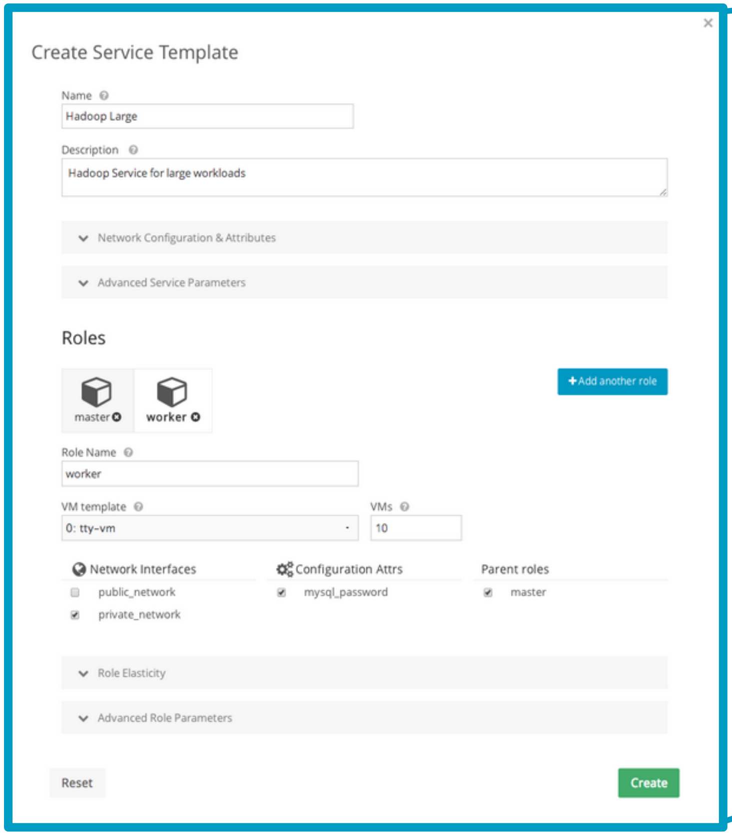
```
{ "name": ”Computing_Cluster",
  "deployment": "straight",
  "roles": [
    {
      "name": "frontend",
      "vm_template": 0
    }, {
      "name": "worker",
      "parents": frontend,
      "cardinality": 2,
      "vm_template": 3,
      "min_vms" : 1,
      "max_vms" : 5,
      "elasticity_policies" :  {
        ”expressions" : ”CPU> 90%”,
        "type" : "CHANGE",
        "adjust" : 2,
        "period_number" : 3,
        "period" : 10}, …
```

Main Challenges for Private HPC Cloud

Execution of Multi-tiered Applications

Functionality for management of interconnected multi-VM applications:

- Definition of **application flows**
- **Catalog** with pre-defined applications
- **Sharing** between users and groups
- Management of **persistent scientific data**
- Automatic **elasticity**



Create Service Template

Name: Hadoop Large

Description: Hadoop Service for large workloads

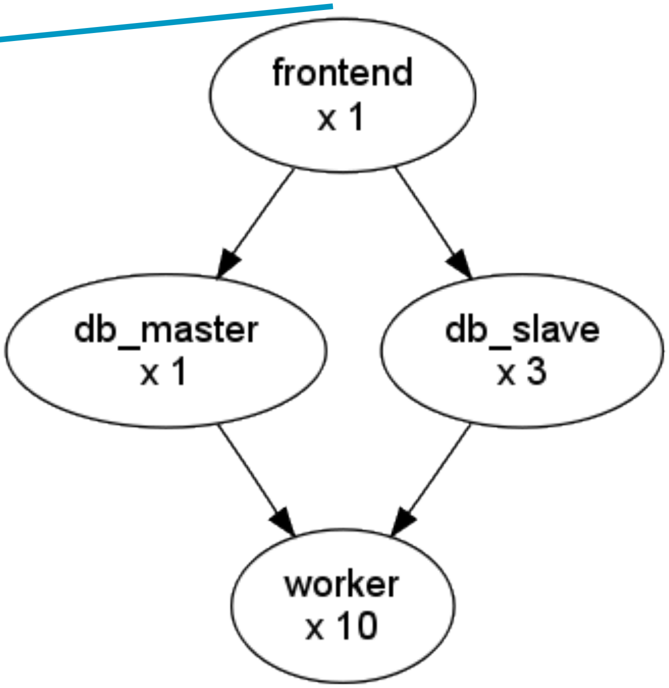
Roles

Role Name: worker

VM template: 0: tty-vm

VMs: 10

Network Interfaces	Configuration Attrs	Parent roles
□ public_network	☑ mysql_password	☑ master
☑ private_network		



Main Challenges for Private HPC Cloud

Performance Penalty as a Small Tax You Have to Pay

Overhead in Virtualization

- Single has processor performance penalty between **1% and 5%**
- NASA reported an overhead between **9% and 25%** (HPCC and NPB)[1]
- Growing number of users demanding containers (**OpenVZ** and **LXC**)

Overhead in Input/Output

- Growing number of **Big Data apps**
- Support for **multiple system datastores including automatic scheduling**

Need for Low-Latency High-Bandwidth Interconnection

- Lower performance, **10 GigE** typically, used in clouds has a significant negative (**x2-x10,** especially latency) impact on HPC applications[1]
- **PCI passthrough** available for VMs that need consumption of raw GPU devices and Infiniband access
- FermiCloud has reported MPI performance (HPL benchmark) on VMs and SR-IOV/**Infiniband** with **only a 4%** overhead[2]

(1) An Application-Based Performance Evaluation of Cloud Computing, NASA Ames, 2013
(2) FermiCloud Update, Keith Chadwick!, Fermilab, HePIX Spring Workshop 2013

Main Challenges for Private HPC Cloud

Resource Management

Optimal Placement of Virtual Machines
- Automatic placement of VM near input data
- Striping policy to maximize the resources available to VMs
- Affinity and Anti-affinity placement policies

Fair Share of Resources
- Resource quota to allocate, track and limit resource utilization

Isolated Execution of Applications
- Full Isolation of performance-sensitive applications

Management of Different Hardware Profiles
- Resource pools (physical clusters) with specific Hw and Sw profiles, or security levels for different workload profiles (HPC and HTC)

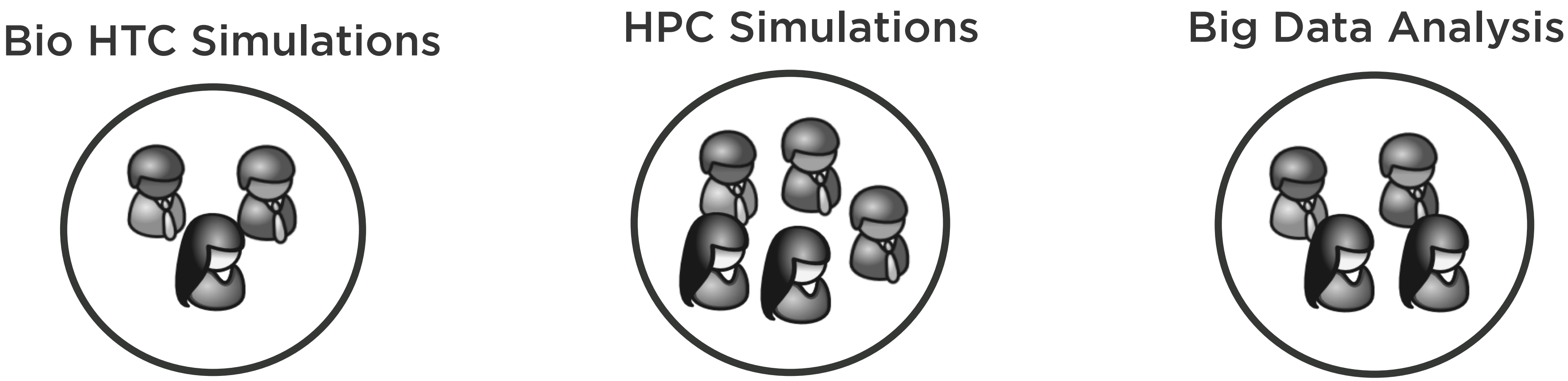
Hybrid Cloud Computing
- Cloudbursting to address peak or fluctuating demands for no critical and HTC workloads

Provide VOs with Isolated Cloud Environ
- Automatic provision of Virtual Data Centers

PCI Passthrough
- Direct connection of GPUs and network to VMs

The Resource Provisioning Framework

Challenges from the Organizational Perspective



Bio HTC Simulations

HPC Simulations

Big Data Analysis

Comprehensive Framework to Manage User Groups

- Several divisions, units, organizations…
- Different workloads profiles
- Different performance and security requirements
- Dynamic groups that require admin privileges

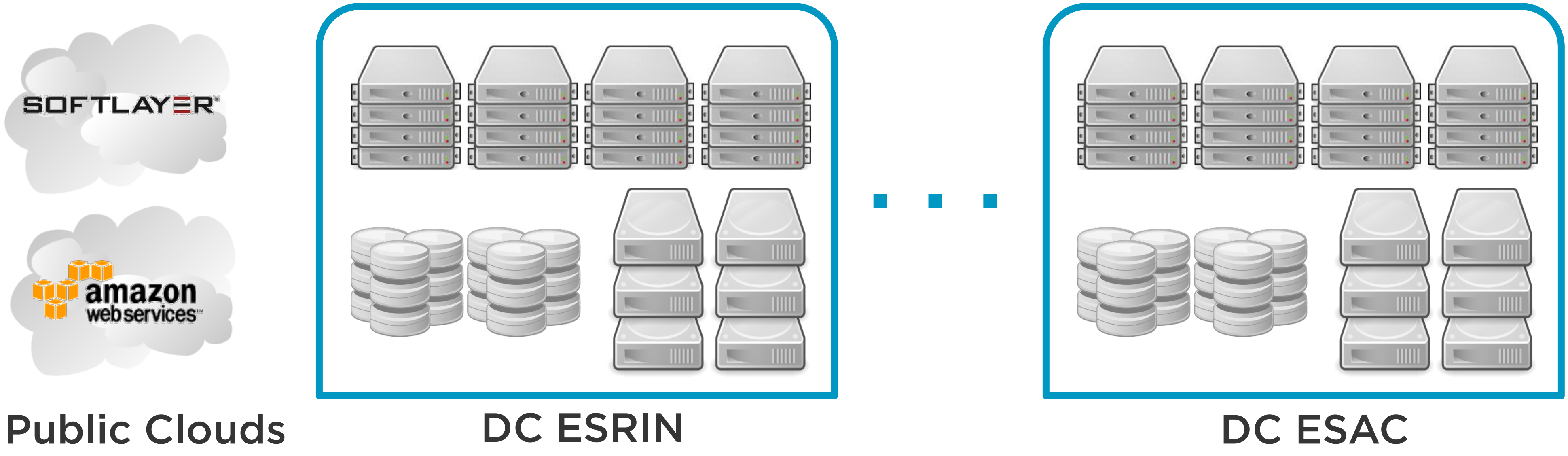
=> From many private clusters to a single consolidated environment

The Resource Provisioning Framework

Challenges from the Infrastructure Perspective

Comprehensive Framework to Manage Infrastructure Resources

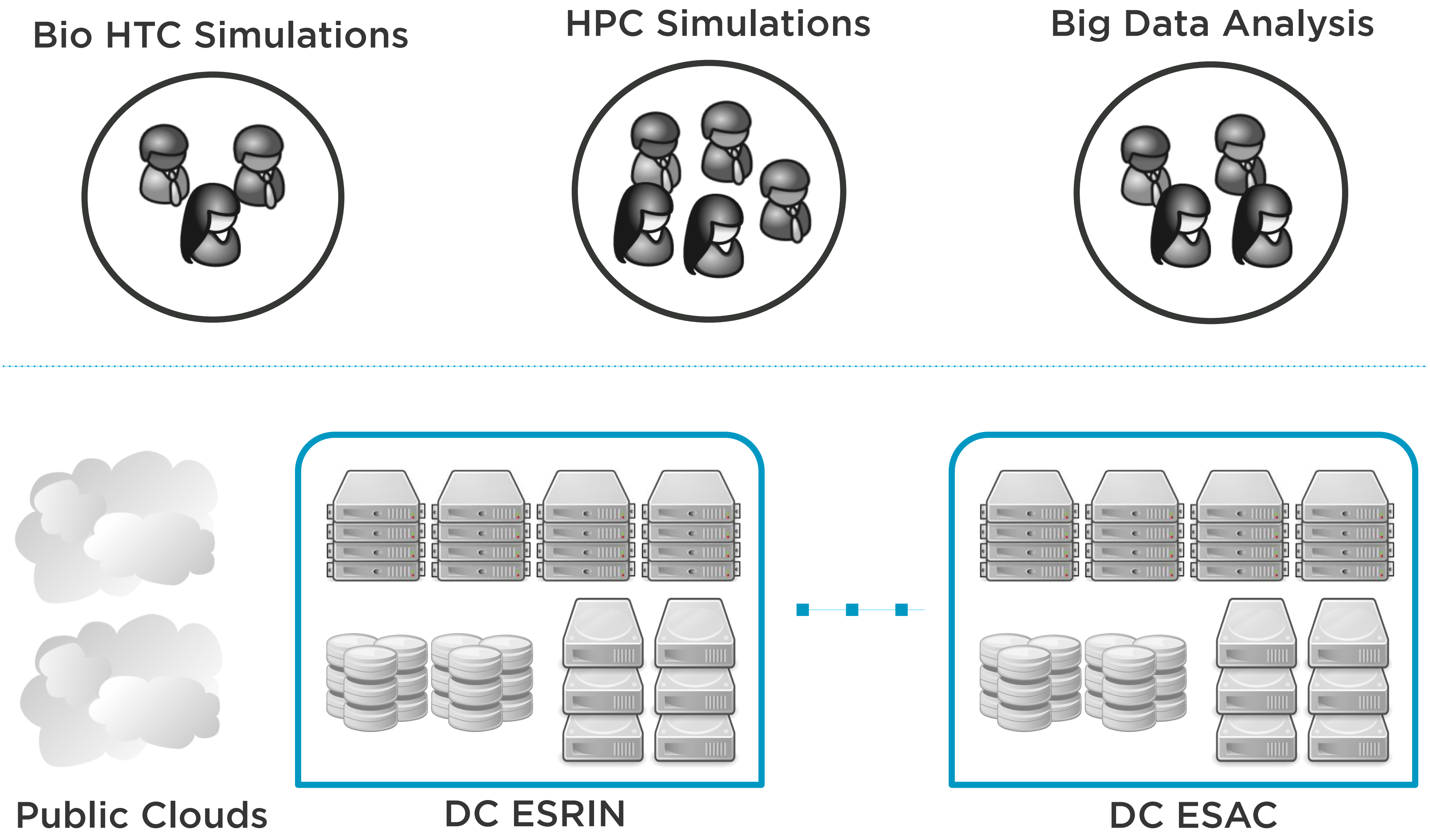
- **Scalability**: Several DCs with multiple physical clusters
- **Outsourcing**: Access to several clouds for cloudbursting
- **Heterogeneity**: Different hardware for specific workload profiles



Dr. Ignacio M. Llorente

The Resource Provisioning Framework

Dynamic Allocation of Private and Public Resources to Groups of Users



Bio HTC Simulations

HPC Simulations

Big Data Analysis

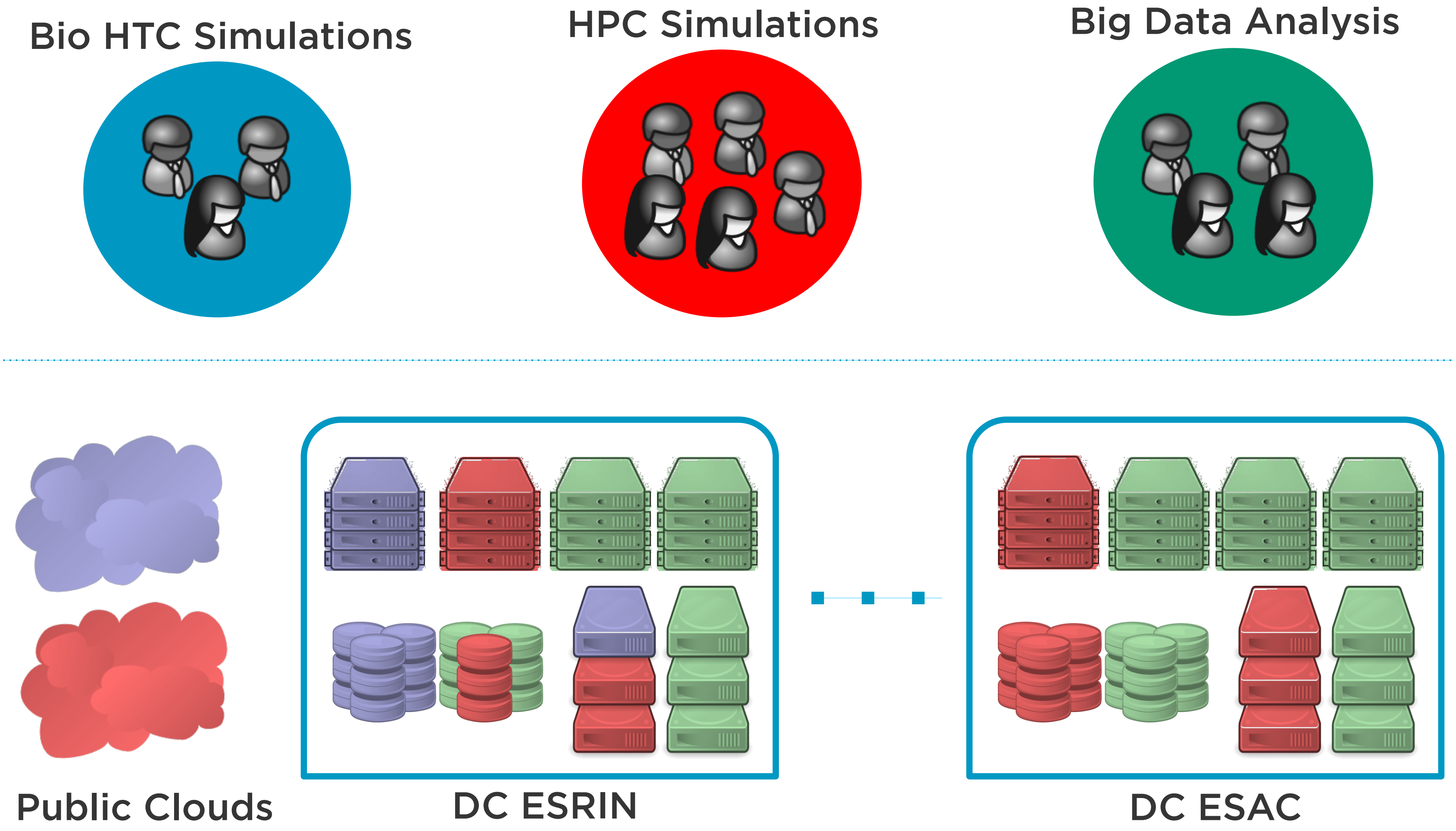
Public Clouds

DC ESRIN

DC ESAC

The Resource Provisioning Framework

Definition of VDCs



Bio HTC Simulations

HPC Simulations

Big Data Analysis

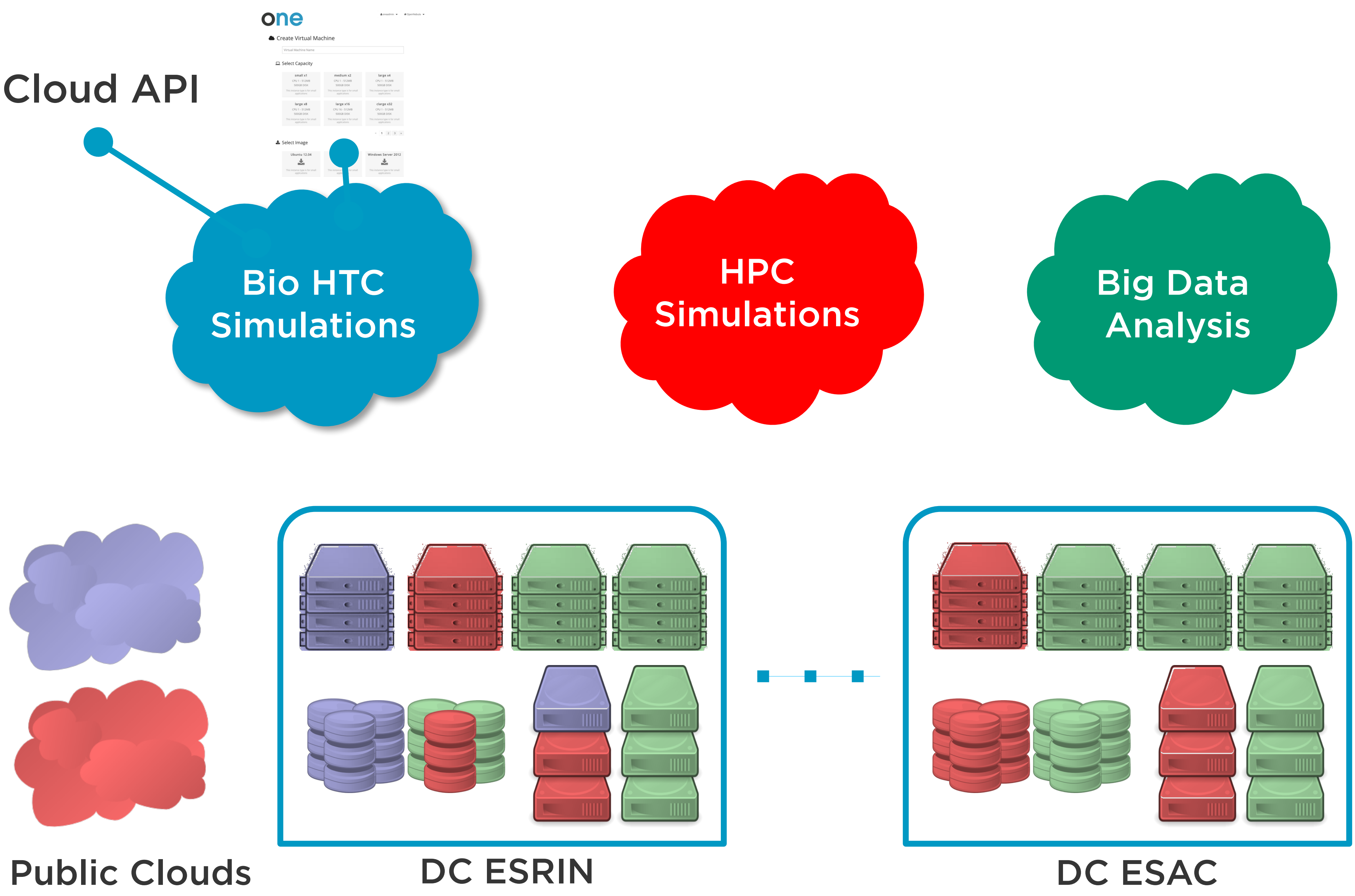
Public Clouds

DC ESRIN

DC ESAC

The Resource Provisioning Framework

Users in each Group Access to its Own Virtual Private Cloud (VDC)



one

oneadmin ▾ OpenNebula ▾

Create Virtual Machine

Virtual Machine Name

Select Capacity

small x1 CPU 1 - 512MB 500GB DISK This instance type is for small applications	medium x2 CPU 1 - 512MB 500GB DISK This instance type is for small applications	large x4 CPU 1 - 512MB 500GB DISK This instance type is for small applications
large x8 CPU 1 - 512MB 500GB DISK This instance type is for small applications	large x16 CPU 16 - 512MB 500GB DISK This instance type is for small applications	clarge x32 CPU 1 - 512MB 500GB DISK This instance type is for small applications

« 1 2 3 »

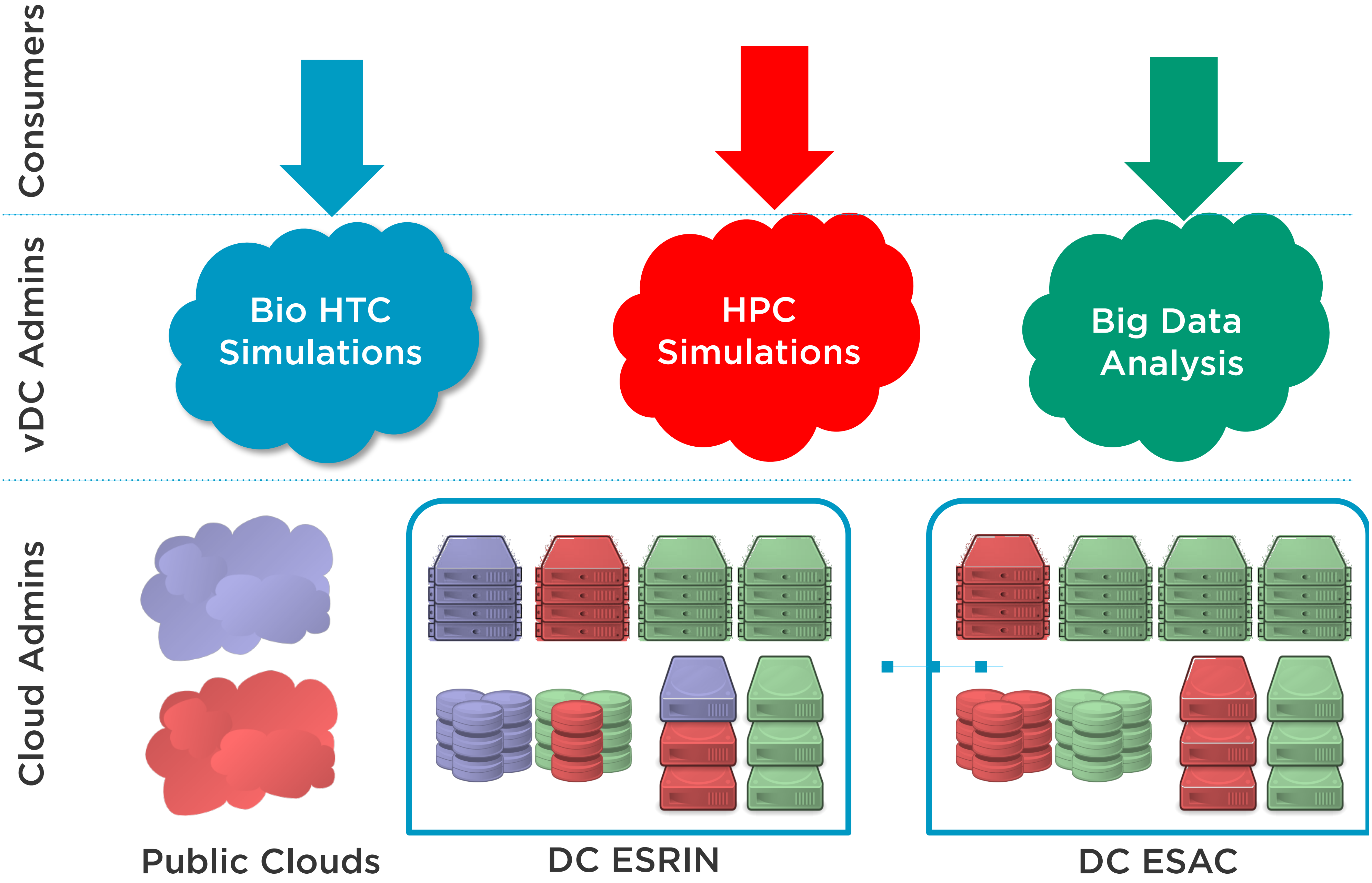
Select Image

Ubuntu 12.04 This instance type is for small applications	This instance t[illegible] for small appl[illegible]	Windows Server 2012 This instance type is for small applications

Dr. Ignacio M. Llorente

The Resource Provisioning Framework

New Level of Provisioning: IaaS as a Service



Dr. Ignacio M. Llorente

Private HPC Cloud Case Studies

Clouds for HPC and Science

Industry

AIRBUS

unity

MONTE DEI PASCHI DI SIENA BANCA DAL 1472

produban

Santander

Supercomputing, Science and Academia

PIC port d'informació científica

CESGA

SURF SARA

KTH VETENSKAP OCH KONST

CENTRE DE SUPERCOMPUTACIÓ DE CATALUNYA CESCA

lrz

Fermilab

IPB

esa

INFN Istituto Nazionale di Fisica Nucleare

MTA SZTAKI

KiSTi www.kisti.re.kr

cerit scientific cloud

Ciemat

CSIRO

KIT Karlsruher Institut für Technologie

DESY

NIKHEF

ITER Instituto Tecnológico y de Energías Renovables S.A.

cmcc Centro Euro-Mediterraneo sui Cambiamenti Climatici

Science & Technology Facilities Council Rutherford Appleton Laboratory

UNIVERSITÉ DE STRASBOURG

CSIC

UCL Université catholique de Louvain

Distributed Computing Infrastructures

EGI

PACIFIC RIM APPLICATIONS AND GRID MIDDLEWARE ASSEMBLY PRAGMA

BonFIRE

chain reds

meGha

Dr. Ignacio M. Llorente

Private HPC Cloud Case Studies

Leibniz Supercomputing Centre

lrz

https://www.lrz.de/cloud/

Nodes	**KVM** on 95 nodes (9.5 TB RAM – 852 cores)
Network	OpenvSwitch
Storage	300TB NAS with NFS
AuthN	**LDAP**
Linux	**SLES 12**
Interface	**Sunstone** Self-service and **EC2** API
App Profile	Legacy, HTC and **MPI HPC**

Private HPC Cloud Case Studies

FermiCloud

Fermilab http://www-fermicloud.fnal.gov/

Nodes	**KVM** on 29 nodes (2 TB RAM – 608 cores) Koi Computer
Network	Gigabit and **Infiniband**
Storage	CLVM+**GFS2** on shared 120TB NexSAN SataBeats
AuthN	**X509**
Linux	**Scientific Linux**
Interface	**Sunstone** Self-service and **EC2** API
App Profile	Legacy, HTC and **MPI HPC**

Typical Workloads

- Production VM-based batch system via the EC2 emulation => 1,000 VMs
- Scientific stakeholders get access to on-demand VMs
- Developers & integrators of new Grid applications

Dr. Ignacio M. Llorente

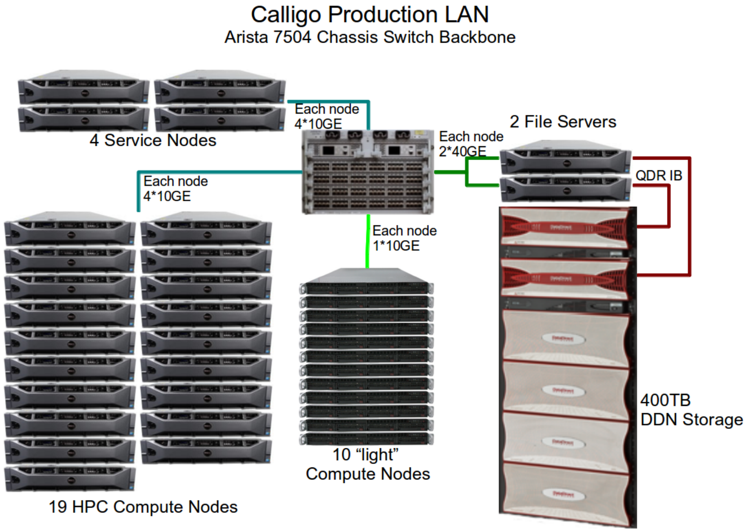
Private HPC Cloud Case Studies

SARA Cloud

SURF SARA

https://userinfo.surfsara.nl/systems/hpc-cloud

Nodes	**KVM** on 30 HPC nodes (900 cores, 8 TB RAM)
Network	2 x **Gigabit** (10G) with Arista switch
Storage	900 TB central storage on a **CEPH** cluster (50 OSD nodes)
AuthN	**Core password**
Linux	**CentOS**
Interface	**Sunstone** and **OCCI**
App Profile	MPI clusters, **windows clusters** and independent VMs



Typical Workloads

- Ad-hoc clusters with MPI and pilot jobs
- Windows clusters for Windows-bound software
- Single VMs, sometimes acting as web servers to disseminate results

Private HPC Cloud Case Studies

Research Computing at Harvard

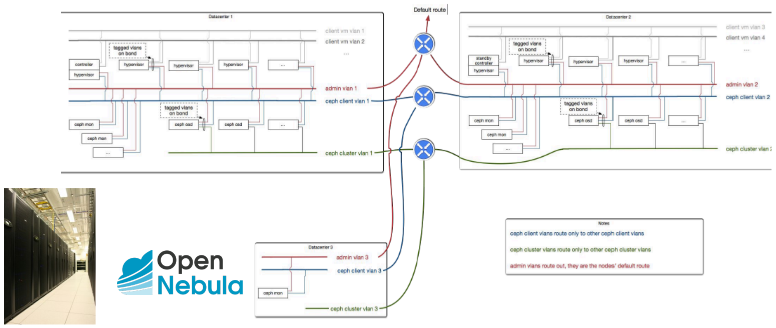
VE RI TAS FASRC

https://rc.fas.harvard.edu

Nodes	**KVM** on 8 nodes (512 cores, 2 TB RAM) in two DCs
Network	2 x **Gigabit** (10G)
Storage	500 TB central storage on a **CEPH** cluster (10 OSD nodes)
AuthN	**LDAP**
Linux	**CentOS**
Interface	**Internal** and **Sunstone**
App Profile	Internal **production** apps

Core Cloud

- Production services in HA



Notes

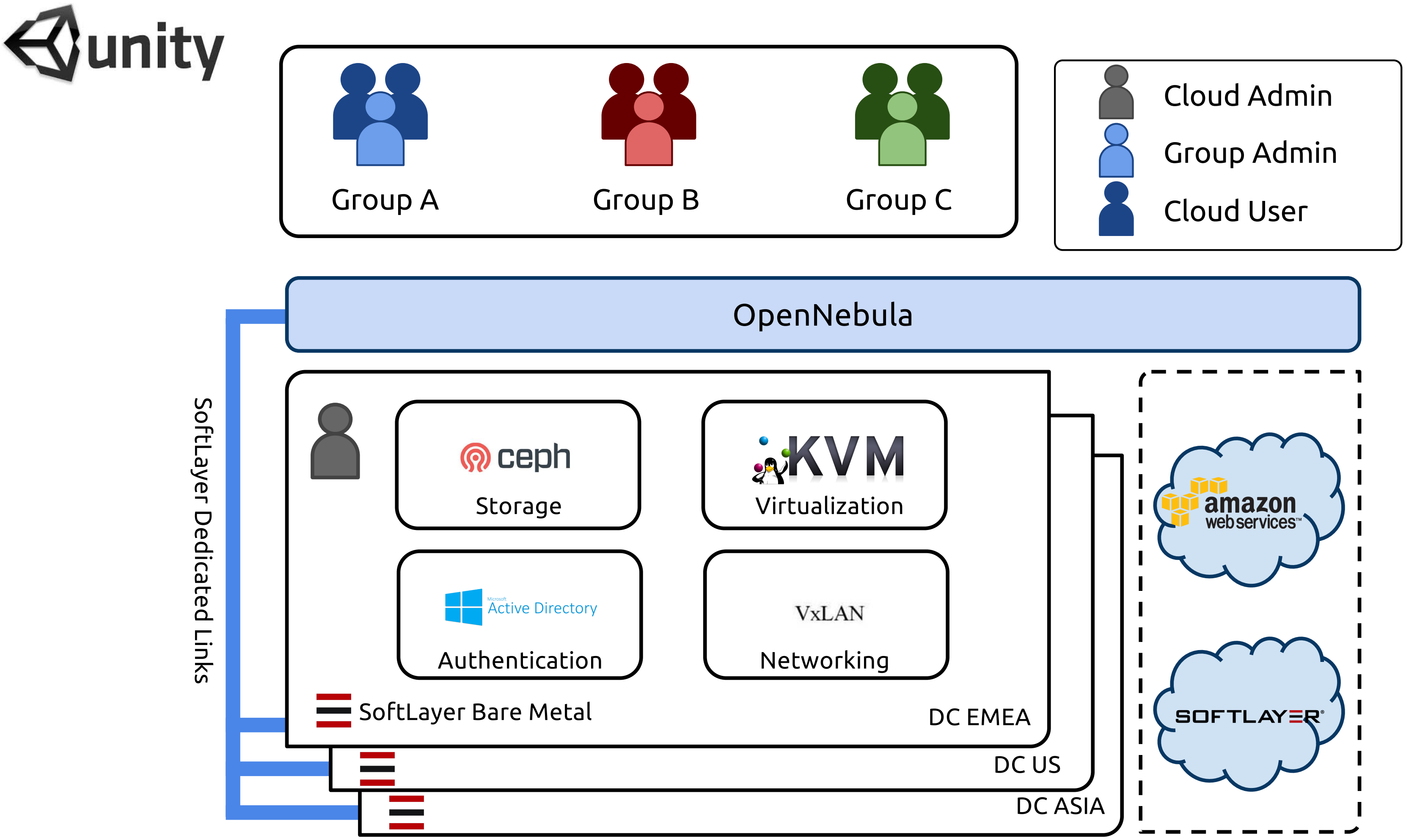
ceph client vlans route only to other ceph client vlans

ceph cluster vlans route only to other ceph cluster vlans

admin vlans route out, they are the nodes' default route

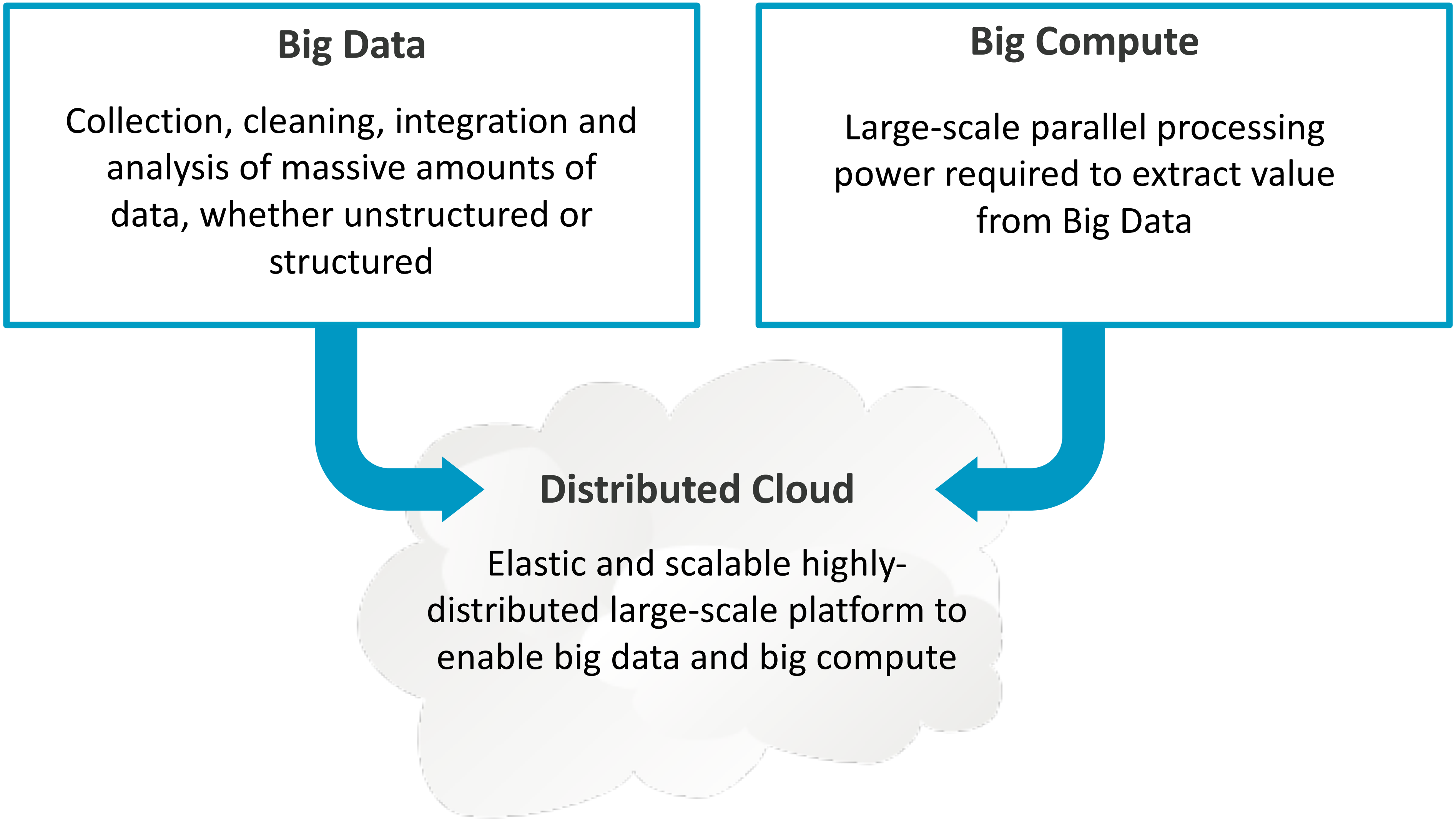
Private HPC Cloud Case Studies

Unity 3D Game Engine



The Future

Distributed Cloud as Meeting Point between Big Data and Big Compute



Big Data

Collection, cleaning, integration and analysis of massive amounts of data, whether unstructured or structured

Big Compute

Large-scale parallel processing power required to extract value from Big Data

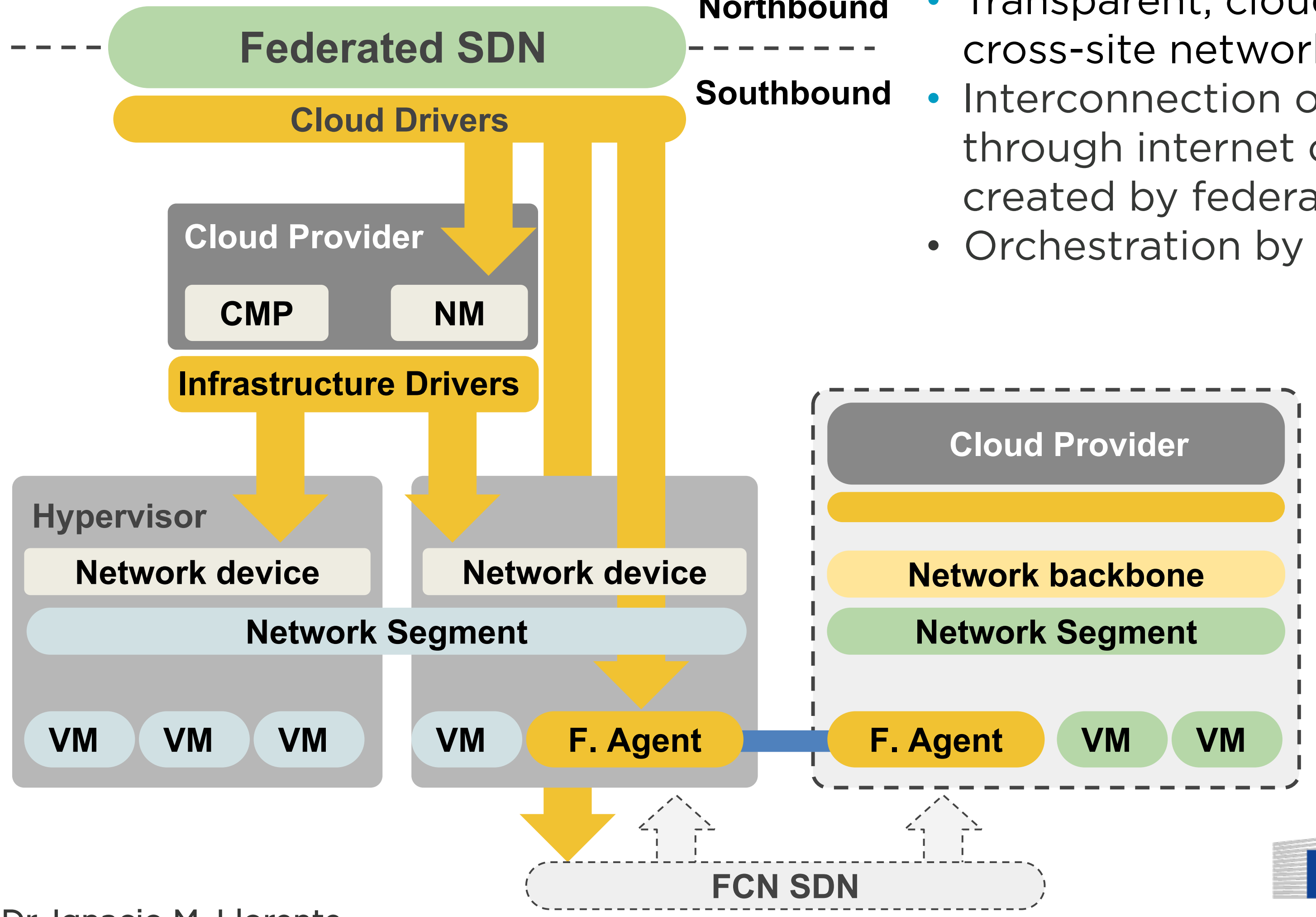
Distributed Cloud

Elastic and scalable highly-distributed large-scale platform to enable big data and big compute

The Future

Research on Federated Cloud Networking

- Federated cloud network model on heterogeneous cloud management platforms and network technologies (i.e. SDN) that can be used in all cloud federation architectures



Proposed Model

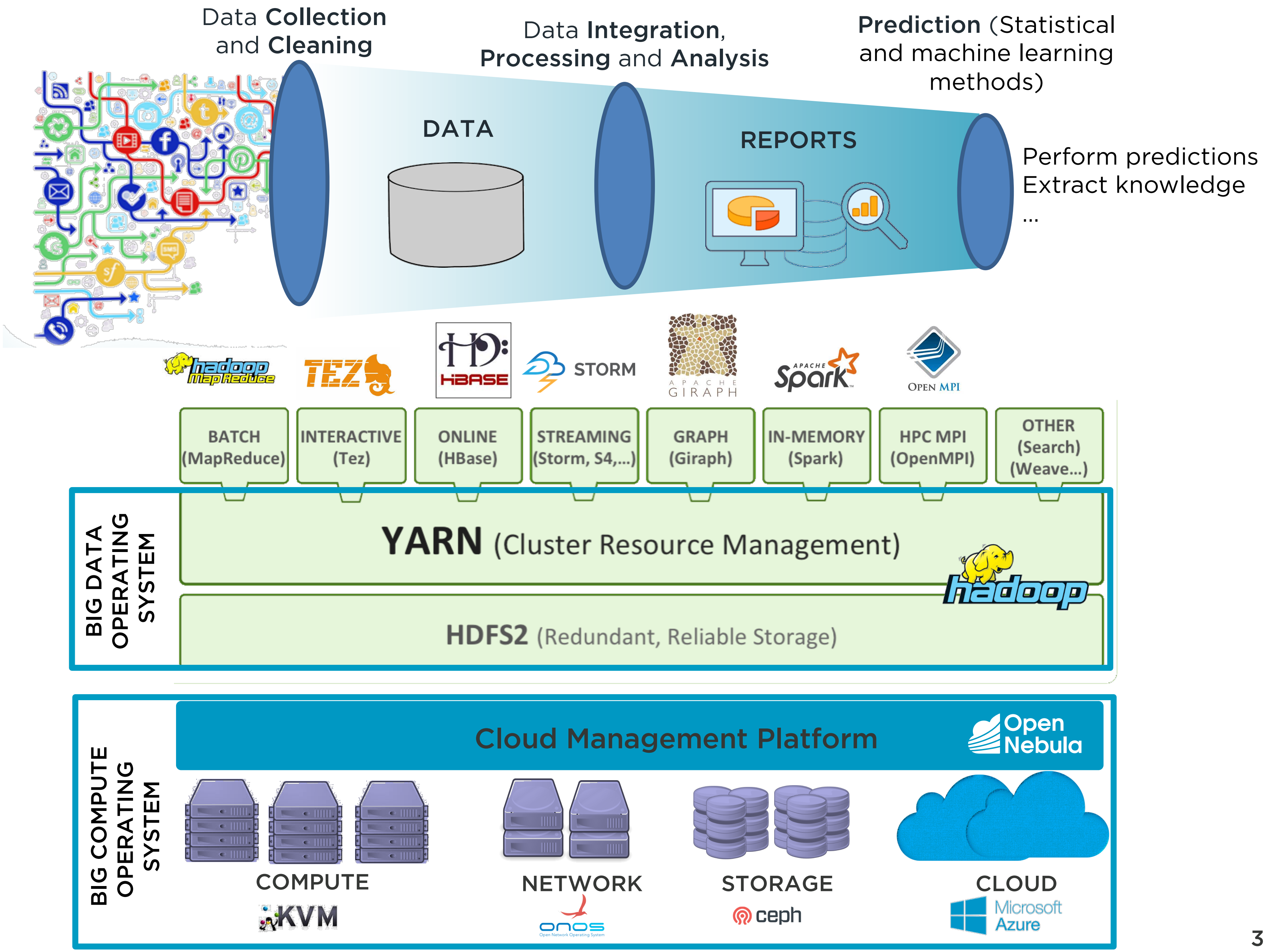
- Transparent, cloud-like provision of cross-site networks (L2/L3)
- Interconnection of network segments through internet overlays (L2/L3) created by federated agents (NFVs)
- Orchestration by a federated cloud SDN

BEACON

Horizon 2020
European Union funding
for Research & Innovation

The Future

Research on Big Data in the Cloud



Data **Collection** and **Cleaning**

Data **Integration**, **Processing** and **Analysis**

Prediction (Statistical and machine learning methods)

DATA

REPORTS

Perform predictions
Extract knowledge
…

BATCH (MapReduce)

INTERACTIVE (Tez)

ONLINE (HBase)

STREAMING (Storm, S4,…)

GRAPH (Giraph)

IN-MEMORY (Spark)

HPC MPI (OpenMPI)

OTHER (Search) (Weave…)

BIG DATA OPERATING SYSTEM

YARN (Cluster Resource Management)

HDFS2 (Redundant, Reliable Storage)

BIG COMPUTE OPERATING SYSTEM

Cloud Management Platform

COMPUTE

NETWORK

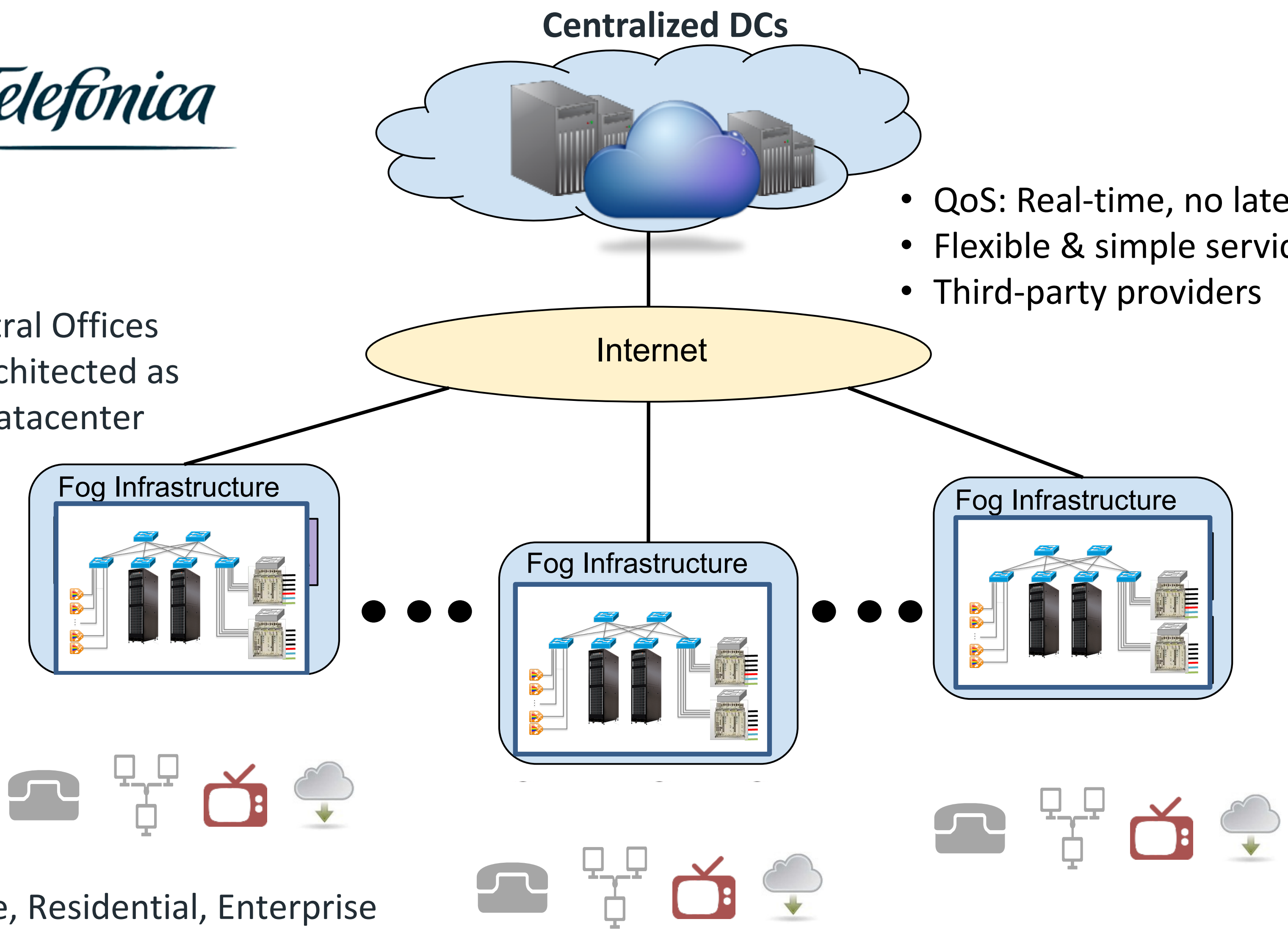
STORAGE

CLOUD

The Future

Research on Edge Computing

Telefonica



- QoS: Real-time, no latency
- Flexible & simple service deployment
- Third-party providers

Central Offices
Re-architected as
a Datacenter

Mobile, Residential, Enterprise

Research References

More about Cloud Architecture and HPC on Cloud

Innovation in Cloud Architecture

- B. Sotomayor, R. S. Montero, I. M. Llorente and I. Foster, “Virtual Infrastructure Management in Private and Hybrid Clouds”, **IEEE Internet Computing**, September/October 2009 (vol. 13 no. 5)
- Rafael Moreno-Vozmediano, Ruben S. Montero, Ignacio M. Llorente, “Multi-Cloud Deployment of Computing Clusters for Loosely-Coupled MTC Applications”, **IEEE Transactions on Parallel and Distributed Systems**, 22(6):924-930, April 2011
- Rafael Moreno-Vozmediano, Ruben S. Montero, Ignacio M. Llorente, “IaaS Cloud Architecture: From Virtualized Data Centers to Federated Cloud Infrastructures”, **IEEE Computer**, 45(12):65-72, December 2012
- Rafael Moreno-Vozmediano, Ruben S. Montero, Ignacio M. Llorente, “Key Challenges in Cloud Computing to Enable the Future Internet of Services”, **IEEE Internet Computing**, 17(4):18-25, 2012.

@imllorente