

El papel del Cloud Computing en BD & DS

Javier Cacheiro López
jlopez@cesga.es

Contenido

- Soluciones Cloud
- Cloud Federado: EGI FedCloud
- BigData en FedCloud
- Conclusiones

Cloud

Elasticidad

Pay-as-you-go

Facilidad

Rapidez

Escalabilidad

Virtualización

- Infrastructure as a Service (IaaS)
 - Amazon EC2
- Platform as a Service (PaaS)
 - Salesforce, Google App Engine
- Software as a Service (SaaS)
 - Google Docs

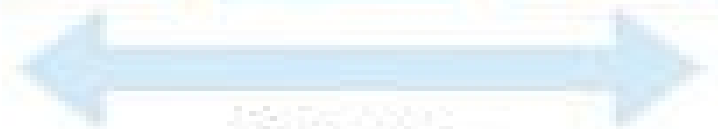
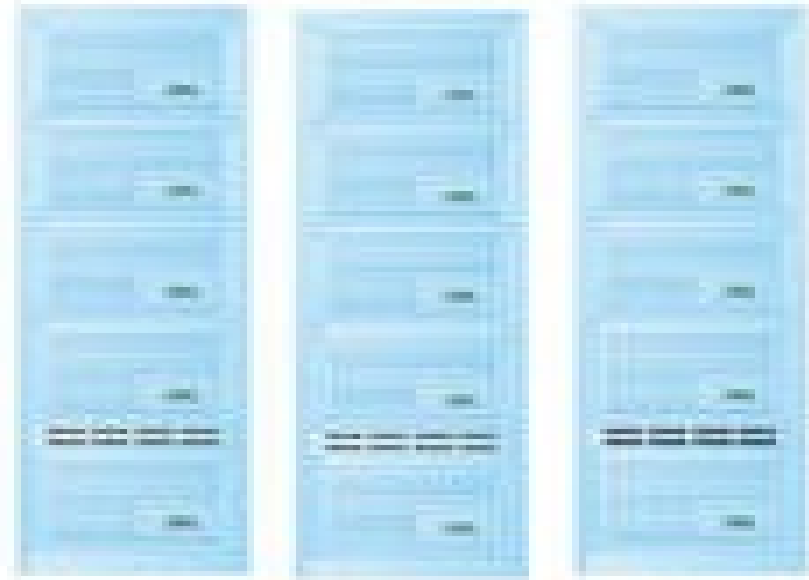
Escalabilidad

Escalabilidad vertical: Scale up

Escalabilidad horizontal: Scale out



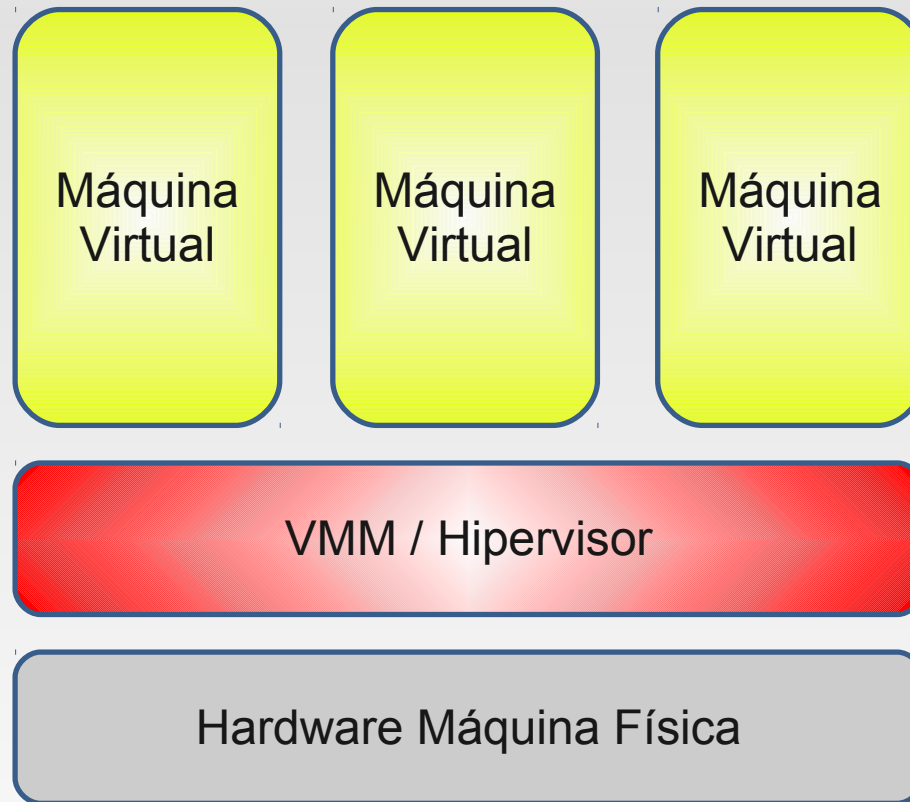
Scale up



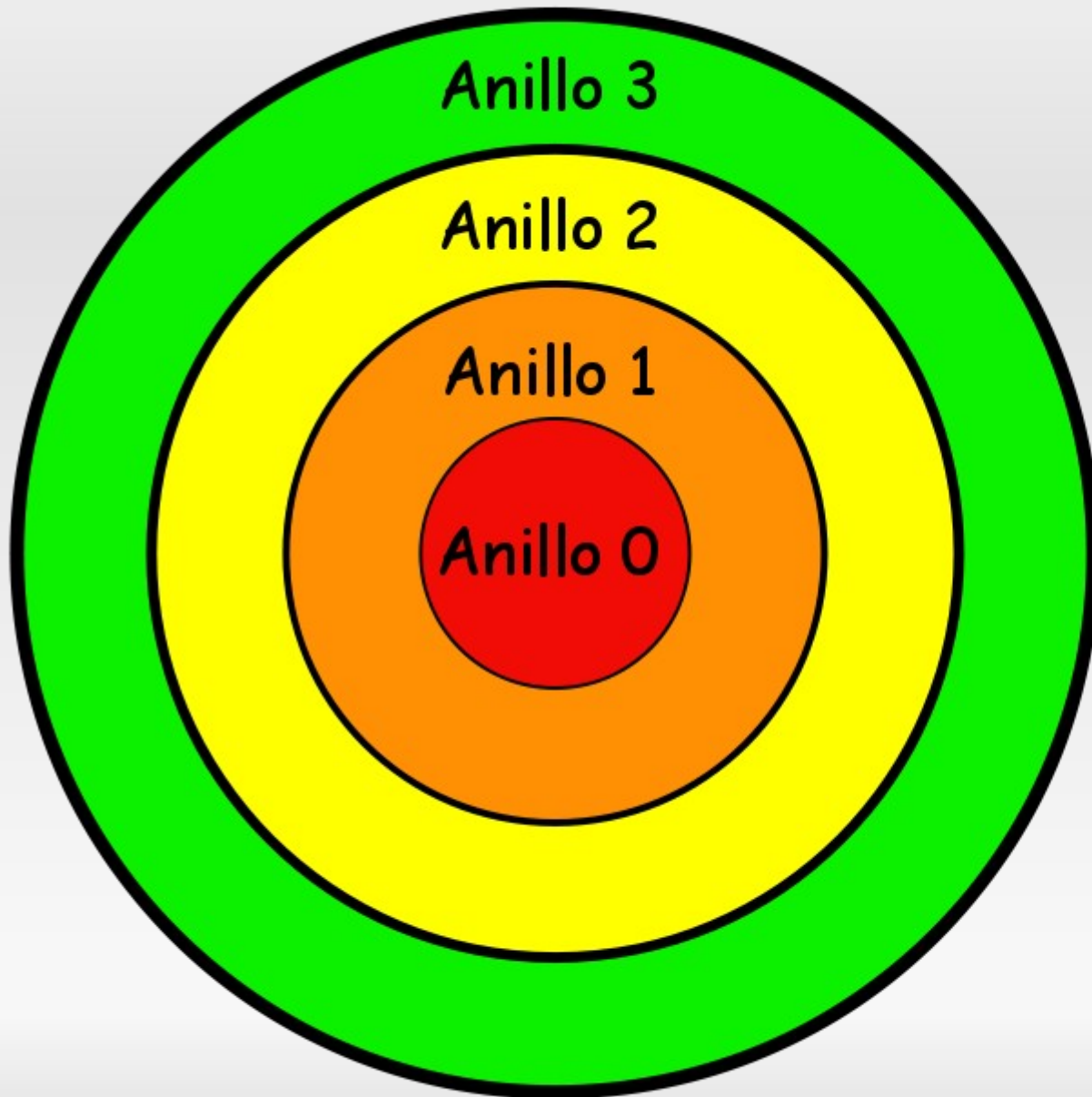
Scale out

Virtualización

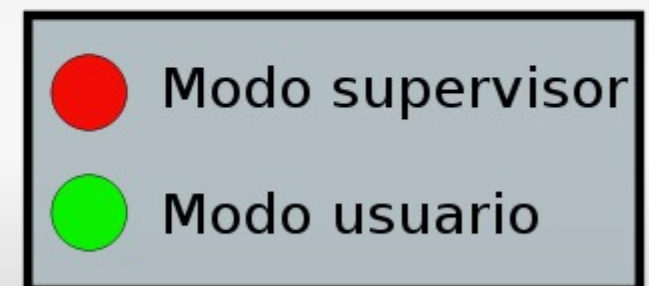
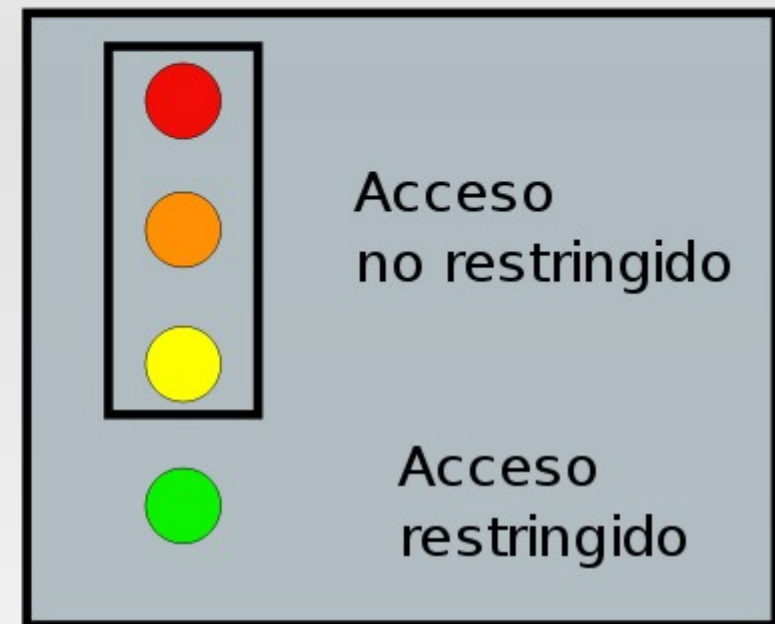
Virtual Machine Monitor (VMM)



Anillos de Protección



Niveles de acceso



Tipos de VMM

- VMM Tipo 1 (*bare metal*)
 - Anillo 0
 - Hipervisor directamente sobre el hardware
- VMM Tipo 2 (*hosted*)
 - Anillo 3
 - Hipervisor dentro de un SO

Modos de Virtualización

	Descripción	Desventajas	Ventajas	Ejemplos
Virtualización completa (<i>full virtualization</i>)	Simulación completa del hardware de la máquina física	Pérdida de rendimiento	No es necesario modificar el SO de la MV	VMWare Workstation, VirtualBox
Paravirtualización	Simulación parcial del hardware de la máquina física. La MV es consciente de que corre en un entorno virtual.	Es necesario modificar el SO de la MV para que se ejecute en el anillo 1	Rendimiento	Xen PV
Virtualización asistida por hardware (<i>Native Virtualization</i>)	Virtualización total que se aprovecha del soporte para virtualización presente en el hardware como las extensiones VT de Intel o Pacifica de AMD	Es necesario hardware específico	Rendimiento	KVM, Xen HVM

Soluciones de Virtualización en entornos cloud

- Xen
- KVM
- VMware
- Hyper-V

Xen: Conceptos Básicos

- Dom0: el SO anfitrión con el kernel de Xen
- DomU: cada una de las máquinas virtuales
- VIF: Virtual Network Interface
- VBD: Virtual Block Device
- HVM: Virtualización con soporte hardware

Linux paravirt_ops (pv-ops)

- Pv-ops es una parte del kernel de Linux que le permite ejecutarse como máquina paravirtual
- Está disponible para x86, x86_64 e ia64
- El mismo kernel puede arrancar nativamente en el hardware físico o como máquina paravirtual
- Xen y VMware tienen soporte de pv-ops

- StubDom representa un nuevo modelo de gestión de dispositivos para MV HVM disponible a partir de Xen 3.3
- El nuevo modelo utiliza un mini-SO para tareas como acceso a disco, a dispositivos de vídeo o a memoria
- Se consiguiera una mejora considerable de rendimiento

KVM: Conceptos Básicos

- Requiere de un procesador con soporte de virtualización
- Soporta paravirtualización para ciertos drivers a través de VirtIO
- RHEL6 añade KVM y elimina Xen

Virtualización Completa vs Paravirtualización

Virtualización Completa

Máquina Virtual

Máquina Virtual

Máquina Virtual

MicroKernel

API Gestión

Traducción Binaria

Driver

Driver

Hardware

Paravirtualización

Máquina Virtual

Máquina Virtual

Máquina Virtual

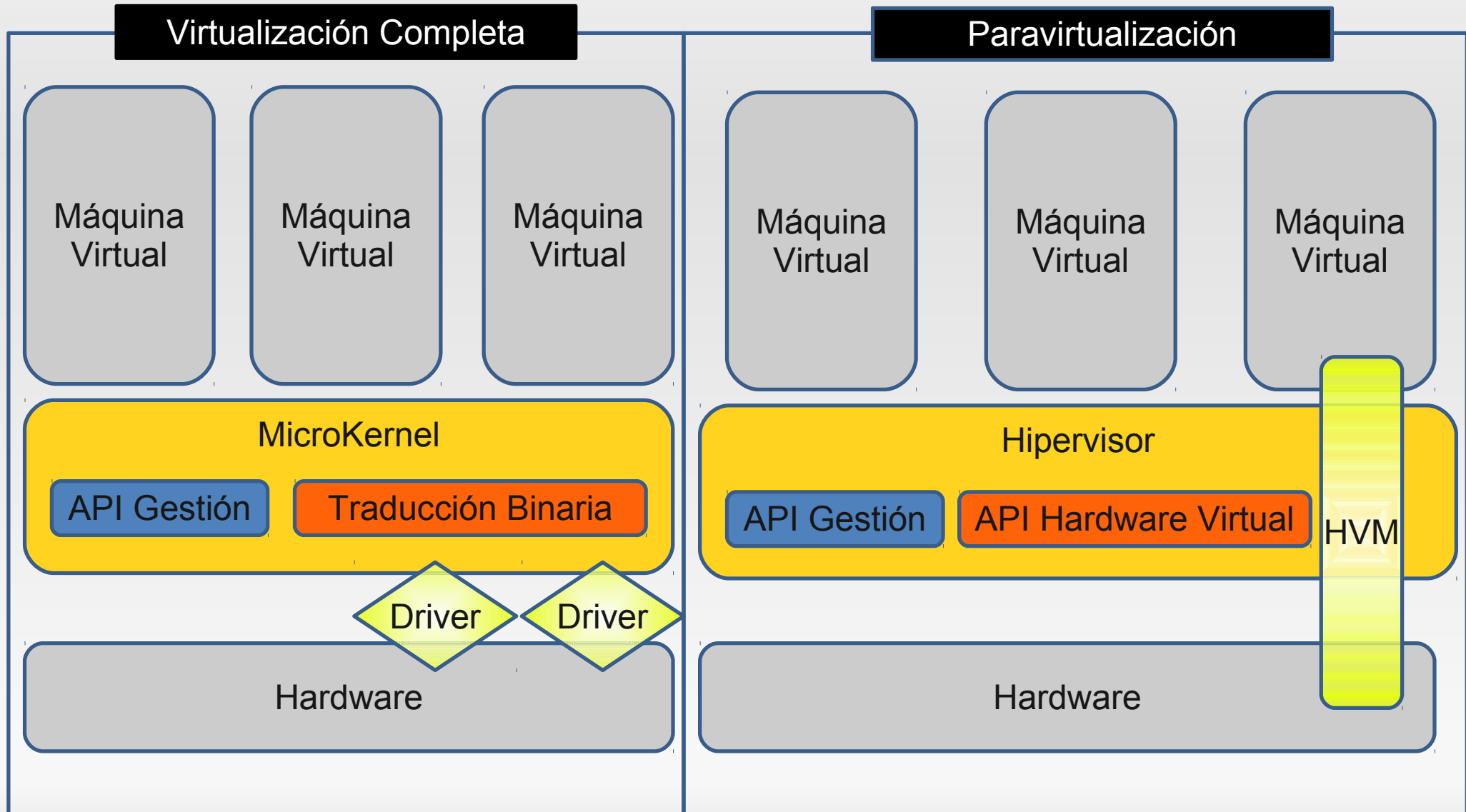
Hipervisor

API Gestión

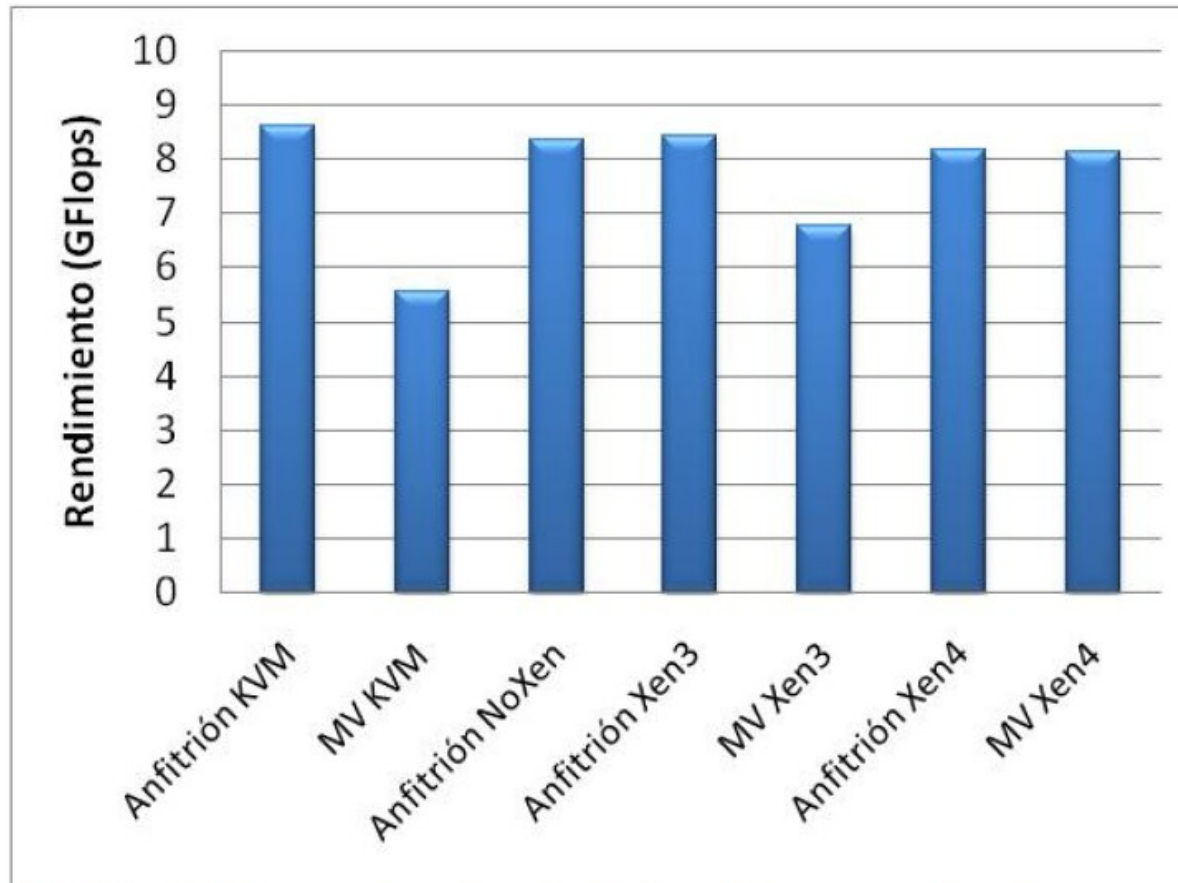
API Hardware Virtual

HVM

Hardware

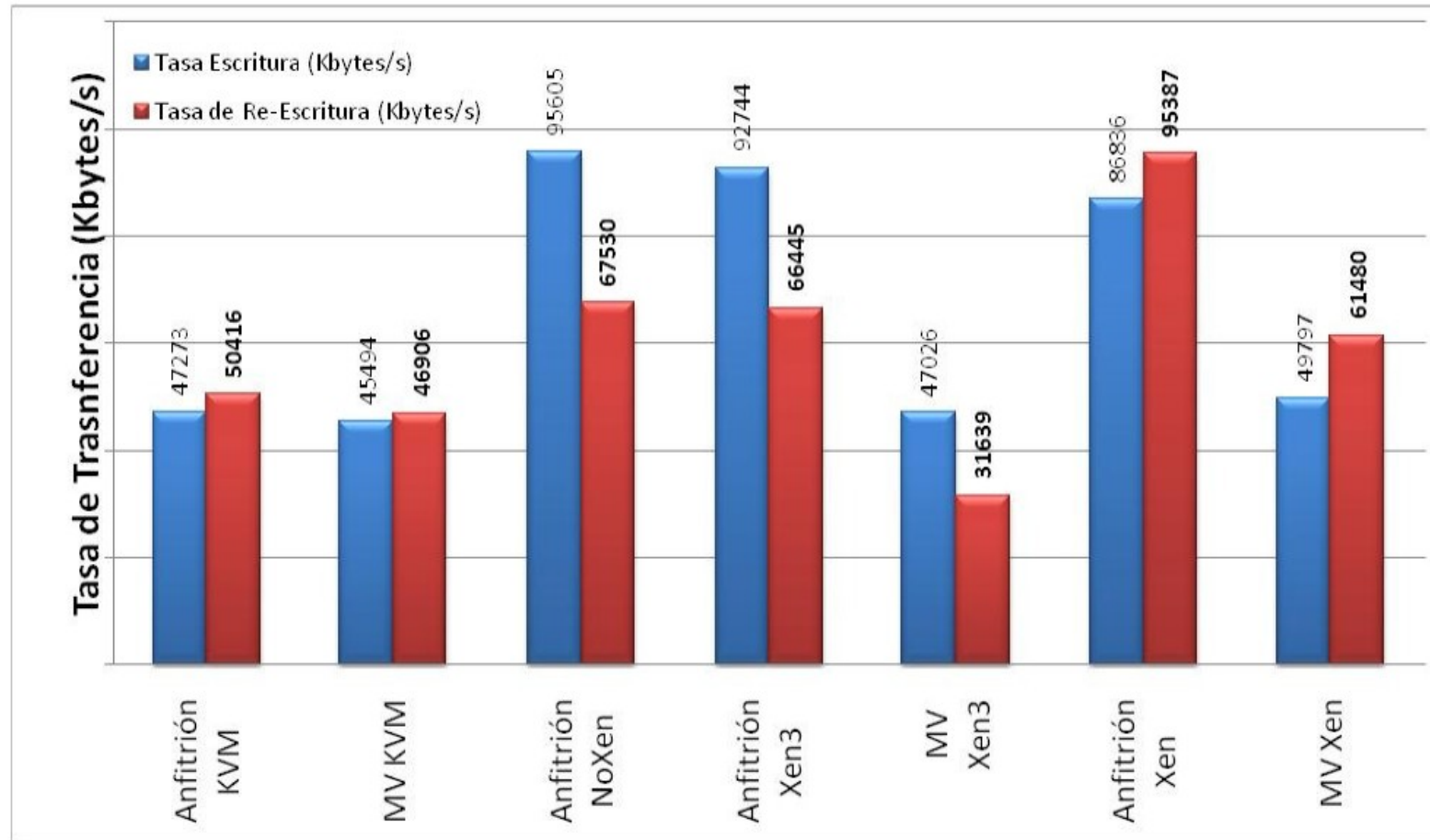


Xen vs KVM: Linpack



Gráfica 1: Comportamiento de los sistemas evaluados con Linpack

Xen vs KVM: I/O



Gráfica 2: Resultados de la prueba de IOZone.

Standards

- Open Virtualization Format (OVF): Formato standard para exportar máquinas virtuales entre VMMs
- Virtual Machine Disk Format (VMDK): Formato standard para la creación de discos virtuales (usado principalmente por VMware)
- VMI: Interfaz alternativo a pv-ops desarrollado por VMware (eliminado a partir del kernel 2.6.37)

Consideraciones prácticas

- Las máquinas virtuales HVM y con virtualización total pueden presentar problemas con el reloj del sistema
- Muchos VMM no gestionan eficientemente MV multi-procesador
- El rendimiento de ficheros y LVM para almacenar las MV es similar
- Algunas distribuciones de Linux incluyen una versión de /lib/tls que no es compatible con Xen (es recomendable deshabilitarla)

Soluciones Cloud IaaS

IaaS Pública



IaaS Privada

OpenNebula.org

cloudstack
open source cloud computing



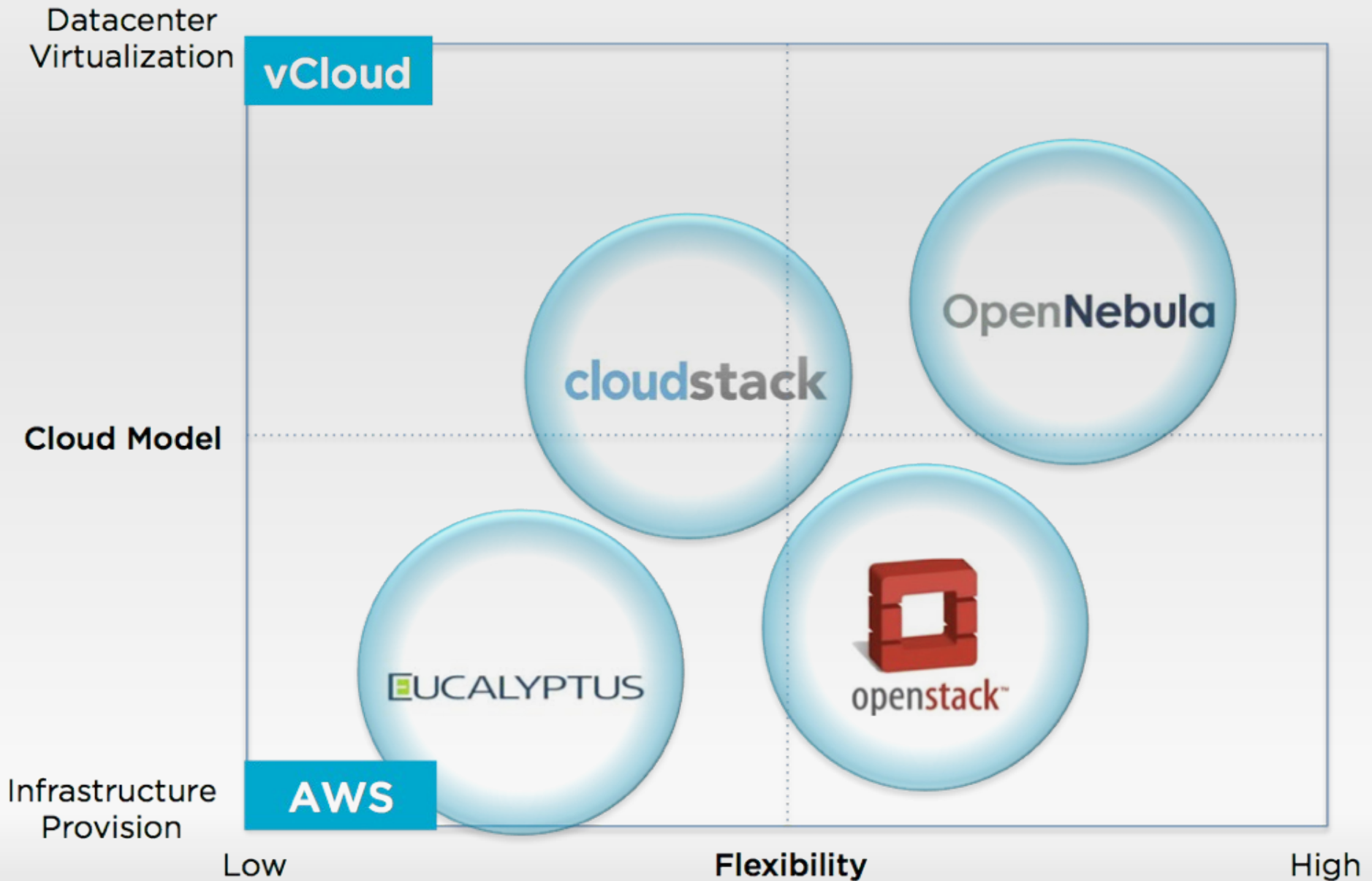
EUCALYPTUS

Public IaaS: Hypervisores

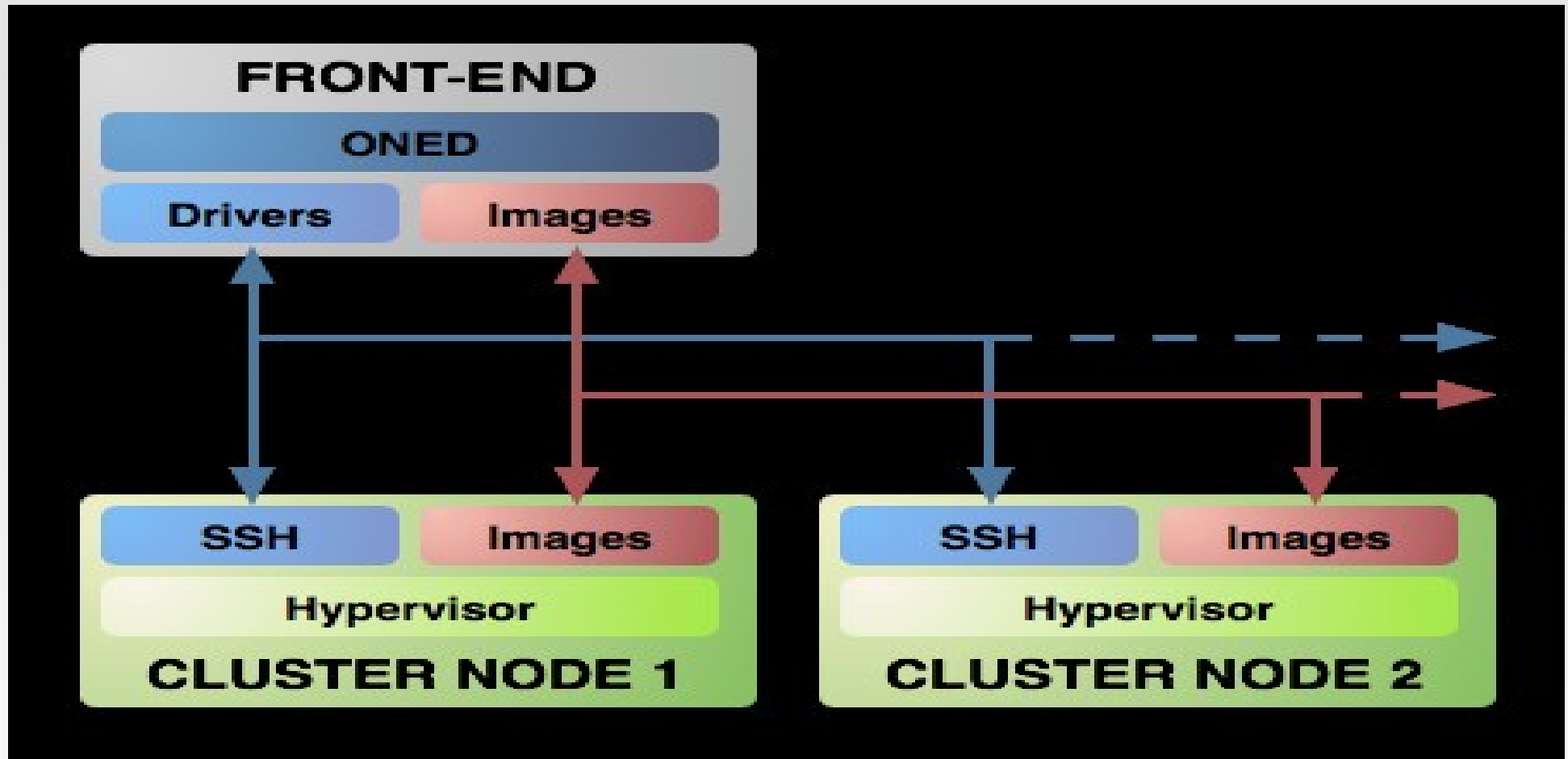
- AWS Amazon EC2:
 - Xen en general y Xen HVM para instancias tipo: cluster compute, high I/O, second generation (m3) and Windows
- Rackspace:
 - Xen para Linux y XenServer para Windows
- Google Compute Engine
 - KVM
- Windows Azure
 - Hyper-V

laaS

Comparativa



Arquitectura OpenNebula



Almacenamiento Cloud

Evolución soluciones almacenamiento

Almacenamiento local



SAN



Distributed Replicated Block Device (DRDB)



Object Storage

Tipos de Almacenamiento



Object Storage

Amazon S3

OpenStack Swift

Ceph

File Storage

NAS (NFS, CIFS)

GlusterFS

Ceph FS

Block Storage

SAS/SATA/iSCSI

Amazon EBS

Ceph RDB

Object Storage

- Escalable:
 - Posibilidad de ampliar/reducir el almacenamiento
- Barato:
 - Agrega los discos locales de muchos servidores
- Replicación automática:
 - Tolerancia a fallos
- Interfaz http: usualmente API REST

Object Storage: Ejemplos

- Amazon S3
- OpenStack Swift
- Ceph
 - Object Storage: Ceph filesystem
 - Block Storage: Ceph RADOS Block Device (RDB)
 - File Storage: Ceph FS
- GlusterFS

Object Storage: Inconvenientes

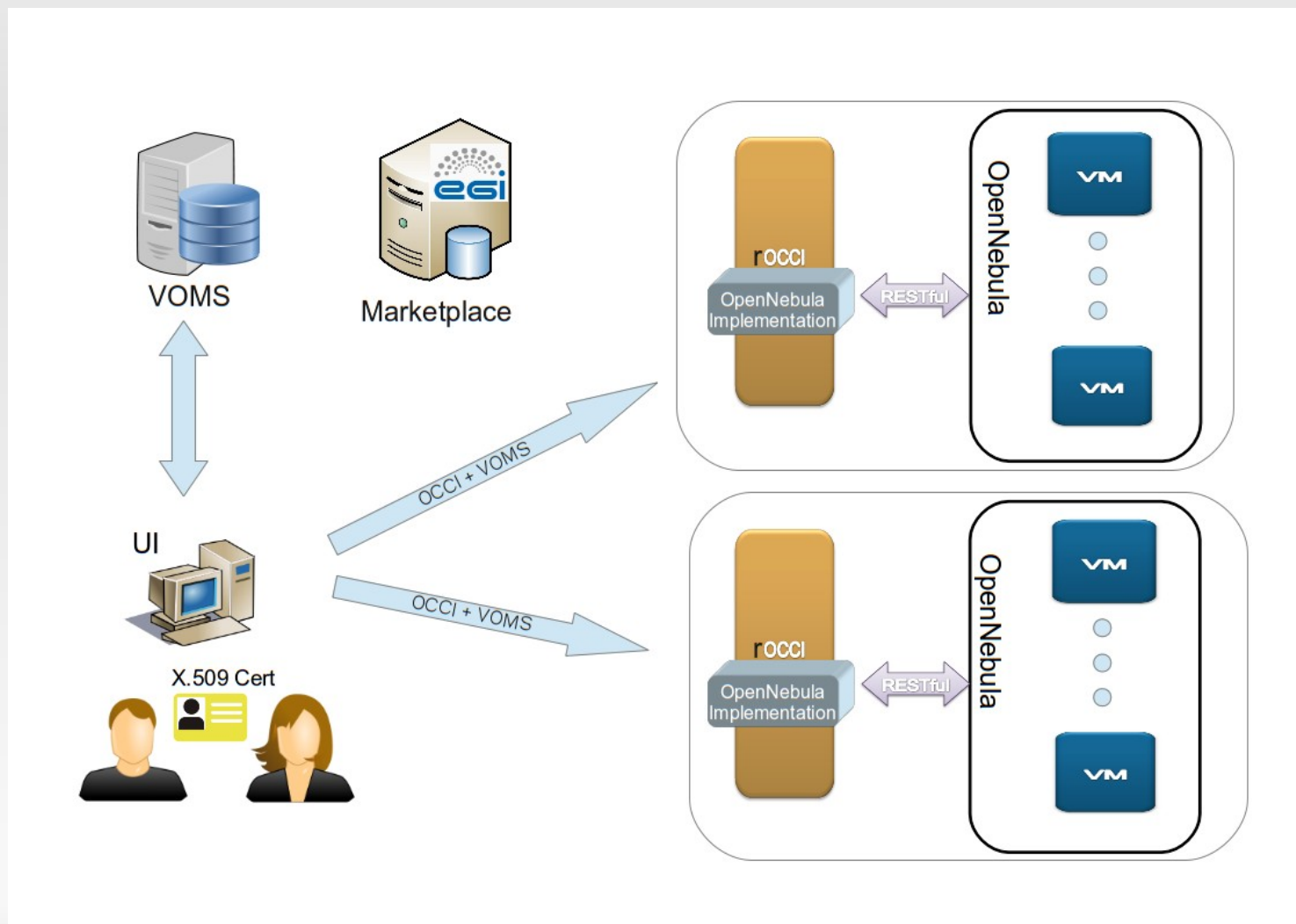
- Si se actualiza un fichero hay que esperar a que se propaguen los cambios a todas la réplicas
- Preferible para almacenar datos que no cambian mucho
- Interfaz http

Soluciones de Almacenamiento IaaS

	Object Storage	Block Storage
AWS	Simple Storage Service (S3)	Elastic Block Storage (EBS)
OpenStack	Swift	Ceph, GlusterFS, NetApp
Eucalyptus	Walrus	Storage Controller (SC)
CloudStack	Swift	Ceph
OpenNebula		Ceph

Cloud Federado

EGI FedCloud

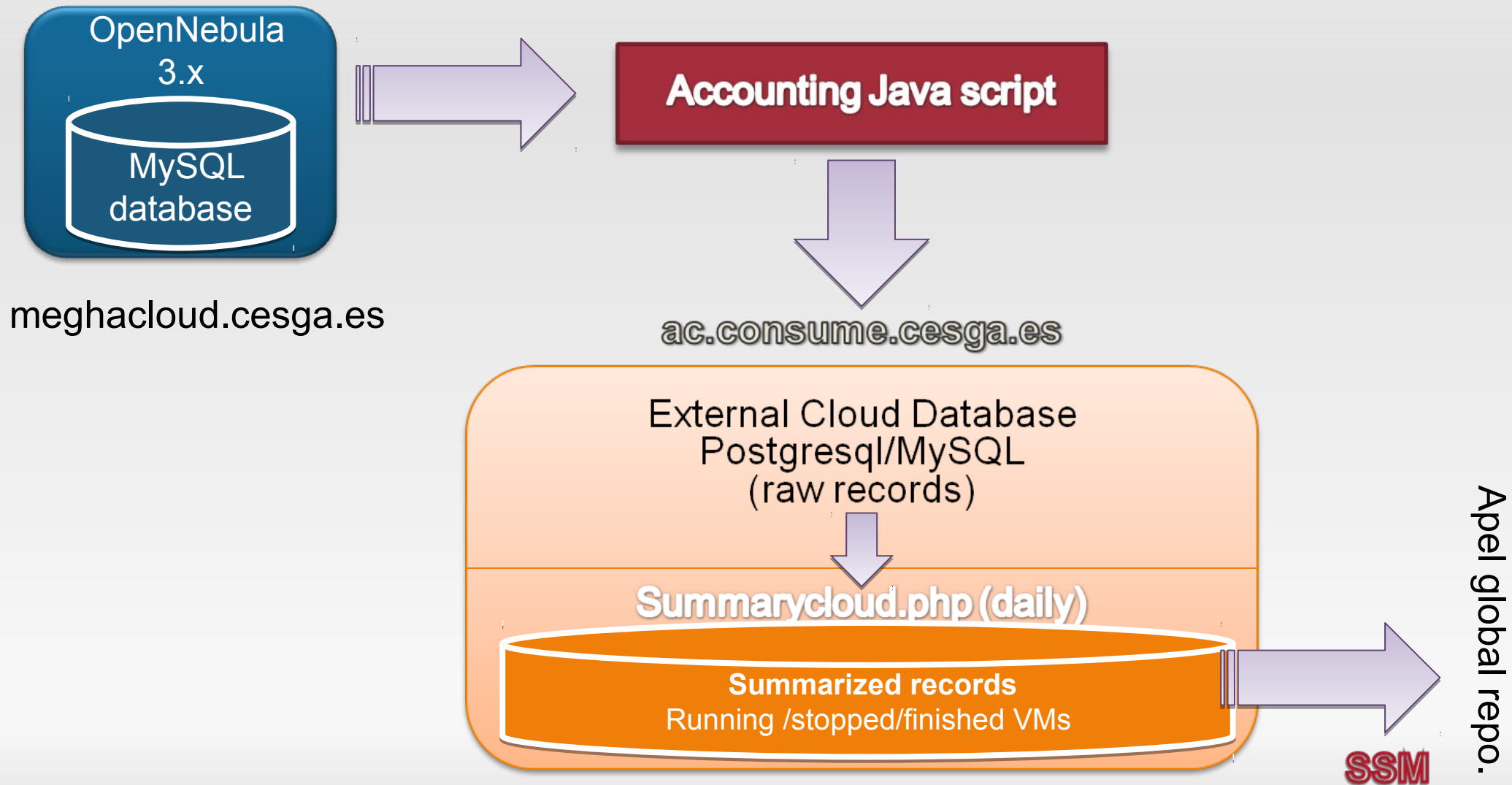


Monitorización

Service Status Details For Host Group 'FedCloud'

Host ↕	Service ↕	Status ↕	Last Check ↕	Duration ↕	Attempt ↕	Status Information
carach5.ics.muni.cz	OCCI 0.8	OK	03-28-2012 12:23:02	0d 19h 50m 20s	1/4	HTTP OK HTTP/1.1 401 Unauthorized - 250 bytes in 0.227 seconds
	OCCI 1.1	OK	03-28-2012 12:17:08	0d 22h 26m 14s	1/4	HTTP OK HTTP/1.1 401 Unauthorized - 348 bytes in 0.108 seconds
	OCCI 1.1 VM Test	OK	03-28-2012 12:18:09	2d 1h 22m 13s	1/4	[2012/03/28 12:18:09 /usr/local/bin/VMInstantiation-probe:main:::237] Started probe 'VMInstantiation-probe'...
	Web interface	OK	03-28-2012 12:17:45	9d 22h 25m 37s	1/4	HTTP OK HTTP/1.1 200 OK - 1737 bytes in 0.235 seconds
cloud-lab.grid.cyf-kr.edu.pl	OCCI 1.1	OK	03-28-2012 12:24:40	0d 19h 48m 42s	1/4	HTTP OK HTTP/1.1 401 Unauthorized - 348 bytes in 0.163 seconds
	OCCI 1.1 VM Test	OK	03-28-2012 12:04:09	4d 2h 12m 18s	1/4	[2012/03/28 12:04:09 /usr/local/bin/VMInstantiation-probe:main:::237] Started probe 'VMInstantiation-probe'...
front.pdc2.pdc.kth.se	OCA	OK	03-28-2012 12:13:23	0d 22h 14m 59s	1/4	HTTP OK HTTP/1.1 404 Not Found - 357 bytes in 3.125 seconds
meghacloud.cesga.es	OCCI 0.8	OK	03-28-2012 12:27:50	6d 2h 3m 15s	1/4	HTTP OK HTTP/1.1 401 Unauthorized - 213 bytes in 0.069 seconds
	OCCI 1.1	OK	03-28-2012 12:21:40	0d 19h 36m 42s	1/4	HTTP OK HTTP/1.1 401 Unauthorized - 348 bytes in 0.002 seconds
	OCCI 1.1 VM Test	OK	03-28-2012 12:06:09	0d 22h 16m 13s	1/4	[2012/03/28 12:06:09 /usr/local/bin/VMInstantiation-probe:main:::237] Started probe 'VMInstantiation-probe'...
	Web interface	OK	03-28-2012 12:16:43	0d 19h 56m 39s	1/4	HTTP OK HTTP/1.1 200 OK - 1724 bytes in 0.003 seconds
occi.cloud.gwdg.de	OCCI 0.8	OK	03-28-2012 12:19:40	0d 19h 53m 42s	1/4	HTTP OK HTTP/1.1 401 Unauthorized - 213 bytes in 0.202 seconds
	OCCI 1.1	OK	03-28-2012 12:27:20	0d 22h 16m 2s	1/4	HTTP OK HTTP/1.1 401 Unauthorized - 348 bytes in 0.111 seconds
	OCCI 1.1 VM Test	OK	03-28-2012 12:08:09	4d 17h 42m 58s	1/4	[2012/03/28 12:08:09 /usr/local/bin/VMInstantiation-probe:main:::237] Started probe 'VMInstantiation-probe'...
	Web interface	OK	03-28-2012 12:25:00	0d 19h 48m 22s	1/4	HTTP OK HTTP/1.0 200 OK - 1761 bytes in 0.233 seconds

Accounting



Ejecutando Hadoop en FedCloud

N+1 Hadoop cluster

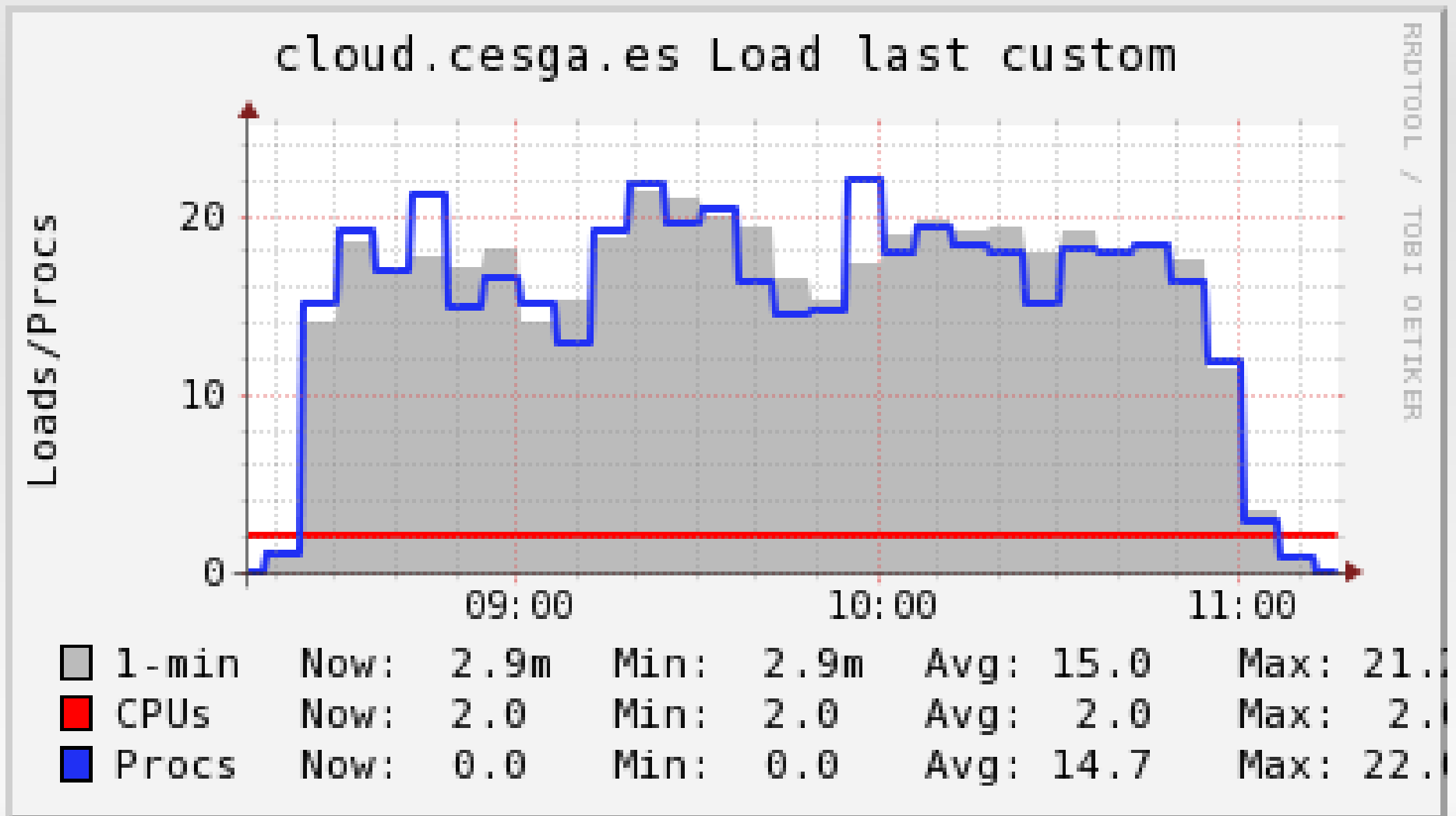
- 1 Master
 - NameNode / Secondary NameNode
 - JobTracker
- N Slaves
 - DataNode
 - TaskTracker

Despliegue del cluster

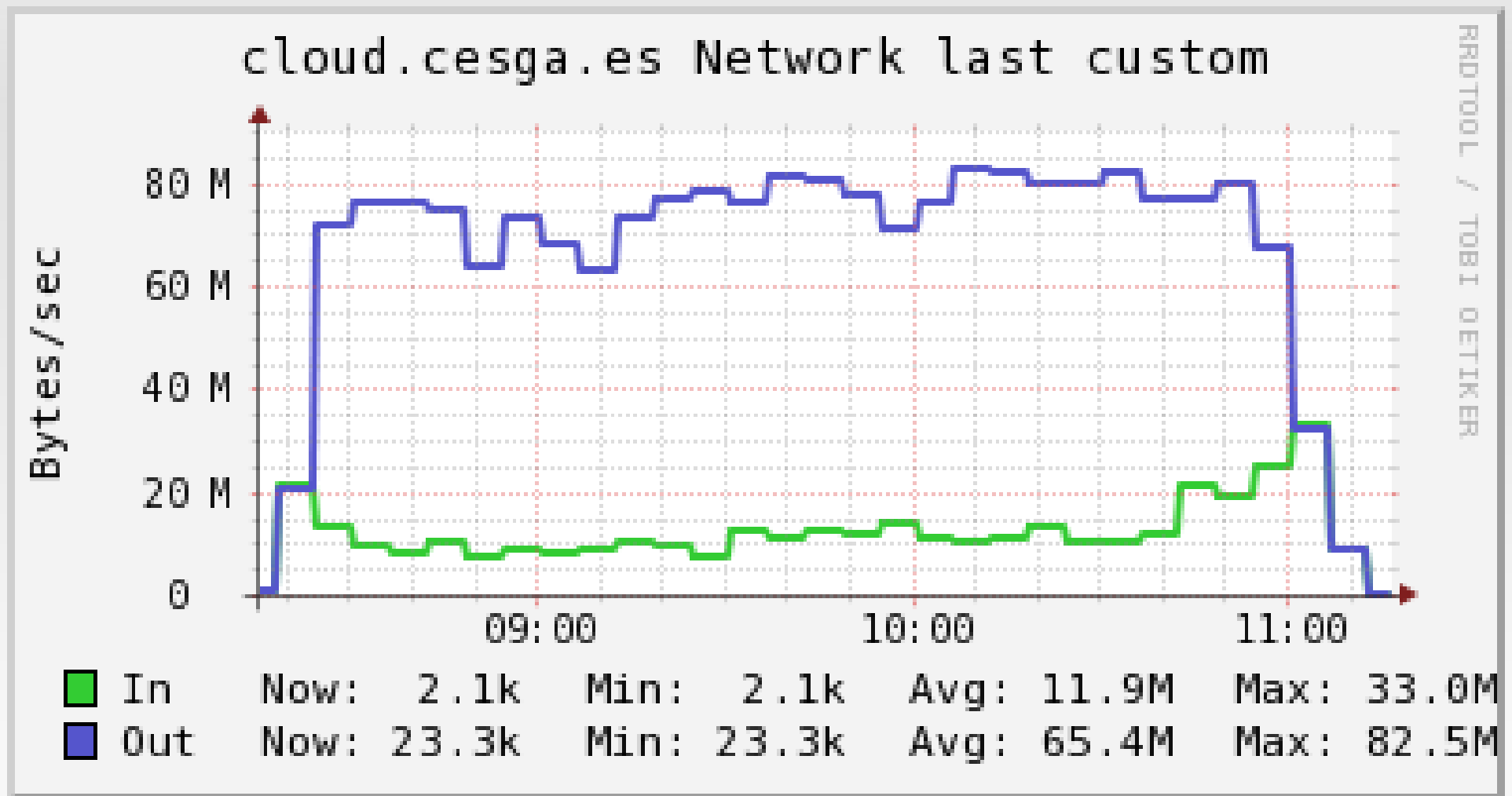
Tiempo de despliegue de un cluster Hadoop de 101 nodos

- Obtener identificador: 71-86 min
- Tiempo total: 2,5-3 horas

Despliegue: Carga en el frontend



Despliegue: Uso de red



Cluster size	Federated		One-site-only	
	put (s)	wordcount (s)	put (s)	wordcount (s)
10 (CESGA)			47 ± 2	169 ± 1
10 (CESNET)			9.5 ± 0.2	160 ± 1
21	235 ± 12	108 ± 2	37 ± 1	97 ± 1
31	189 ± 4	90 ± 2	36 ± 2	78 ± 2
41	190 ± 11	81 ± 4	33 ± 2	71 ± 3
51	133 ± 7	73 ± 3	33 ± 1	66 ± 2
101	94 ± 7	67 ± 22	33 ± 4	52 ± 5

Cluster size	Federated		One-site-only	
	put (s)	wordcount (s)	put (s)	wordcount (s)
51	19190	1001 ± 39	6705	1347 ± 117
101	13208	705 ± 14	5665	725 ± 18

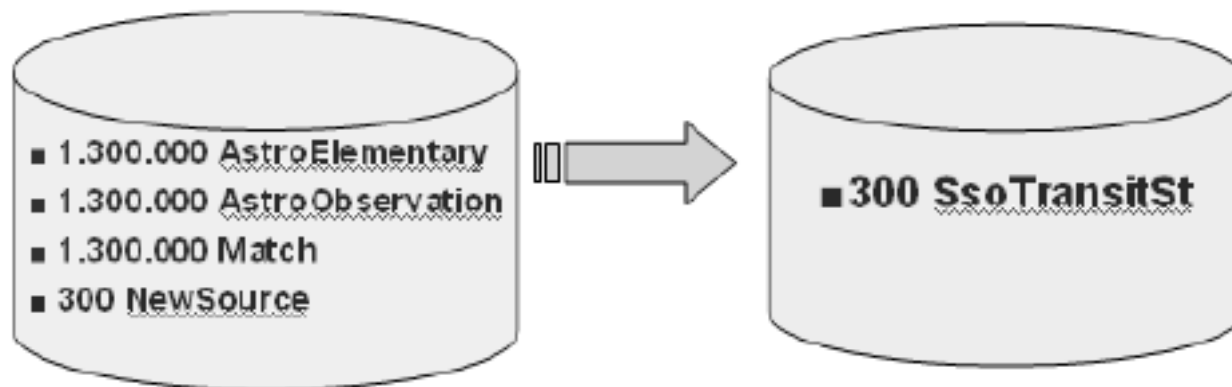
GAIA

- Misión de la ESA para elaborar un mapa tridimensional de nuestra Galaxia, la vía láctea: <http://sci.esa.int/gaia/>
- El satélite será lanzado en agosto
- Datos a analizar: ~1PB

Se analizaron distintas soluciones:

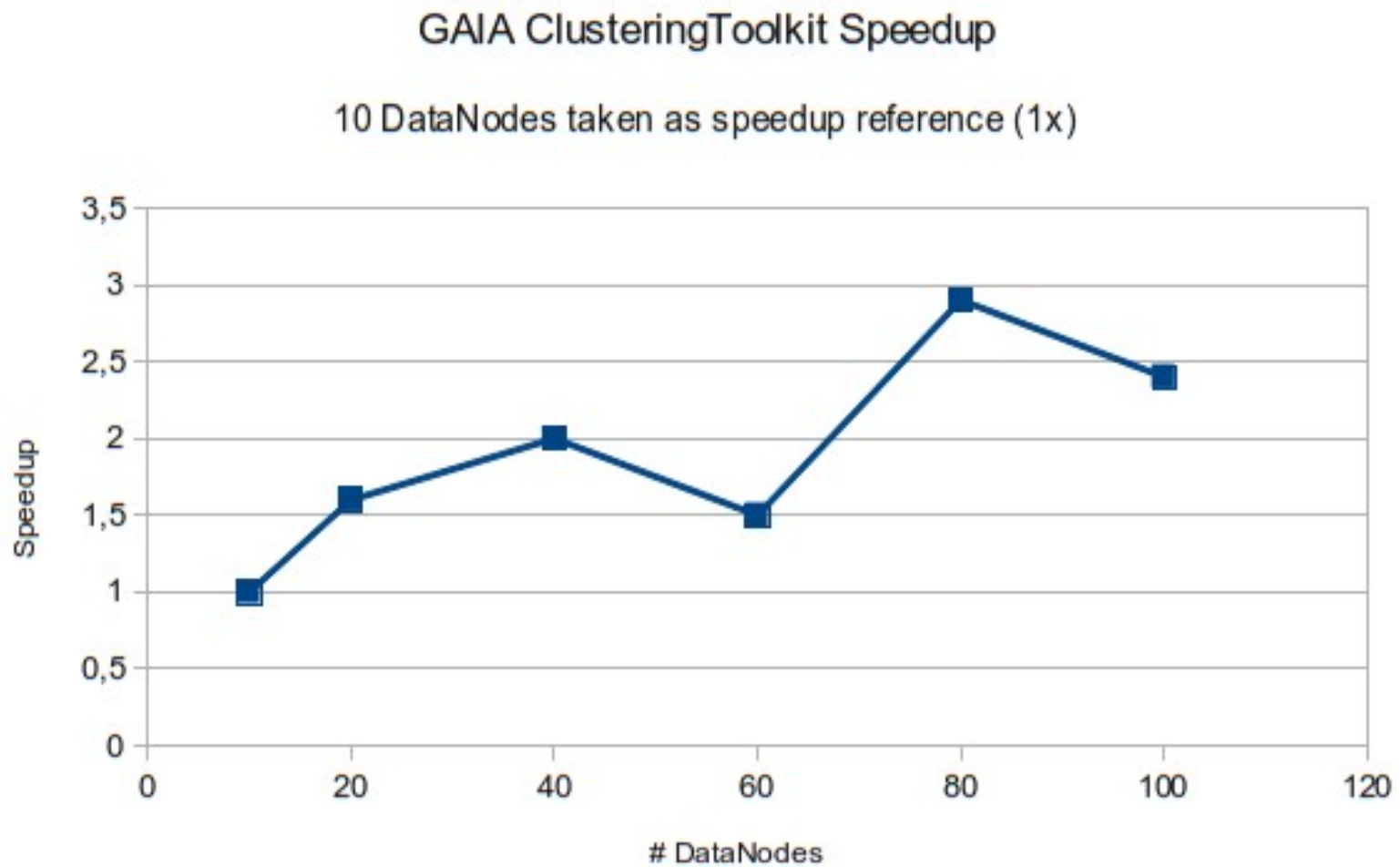
- Cassandra
- Hadoop
- InterSystems Cache
- IBM DB2
- PostgreSQL 9 + PI/Proxy2 + Cassandra

- Solución elegida: Hadoop



Existing solution (PostgreSQL 8)	Cassandra	Hadoop	Cache	DB2	PI/proxy+ PostgreSQL9
14min50s	3min37s	3min37s	2min25s	13min55s	6min50s

GAIA: Escalabilidad en FedCloud



Conclusiones

Ejecutar BigData sobre una plataforma IaaS ofrece ventajas e inconvenientes



Pros

- Elasticidad: tanta como tenga la aplicación
- Rapidez para obtener recursos
- Facilidad en la configuración
- Sencillez en el despliegue
- Ideal para pruebas

Contras

- Tiempo de despliegue:
 - Es necesario optimizar el gestor cloud para mejorar el tiempo de despliegue
- Rendimiento
 - Es necesario optimizar las MV para mejorar su rendimiento de E/S
- Entorno heterogéneo
- Rendimiento de cada MV variable dependiendo de la carga del anfitrión