La Virtualización en un Entorno Científico. Experiencia del Centro de Supercomputación de Galicia

José Ignacio López Cabido
nlopez@cesga.es

Madrid, Marzo de 2008

CESGA
XUNTA DE GALICIA
CSIC CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS
ERDF European Regional Development Fund

ESTABLISHED IN 1993 IN SANTIAGO DE COMPOSTELA

CESGA

CESGA

MISSION STATEMENT

To provide high performance computing and communication resources and services to the scientific community of Galicia and to the National Research Council (CSIC), as well as, to institutions and enterprises with R&D activity.

To promote the use of new information and communication technologies applied to research within the scientific community of Galicia.

To become a consolidated RTD Centre of Excellence serving as international scientific and technological reference in the field of computational science and numerical simulation.

LEGAL ENTITIES

- Public Company
- Public Foundation

PARTNERS

- Regional Government of Galicia 70%
- National Research Council of Spain 30%

Xunta de Galicia

CSIC

Madrid, Marzo de 2008

CURRENT CESGA'S COMMUNITY OF USERS

- **Galician Universities**
- **Galician Regional Government Research Centres**
- **Spanish National Research Council (CSIC) Centres**
- **Other public or private enterprises and institutions**
  - Hospital Laboratories
  - Private Industries R&D Departments
  - Technological & Research Centres
  - Other Universities worldwide
  - Non-profit R&D organizations

SERVICES

- HPC, HTC & GRID Computing
- User Data Storage
- Advanced Communications Network
- Video streaming broadcast & on-demand
- Remote Learning & Collaboration Room Network
- e-Learning & Collaboration Tools
- GIS (Geographical Information Systems)
- e-Business Innovation Consulting
- R&D&I Project management

TECHNOLOGY
CESGA'S TECHNOLOGYCAL EVOLUTION SERVERS INSTALLED

1993
VP 2400

2,5 GFLOPS

1998
VPP 300 **AP 3000**

14,1 GFLOPS 12 GFLOPS

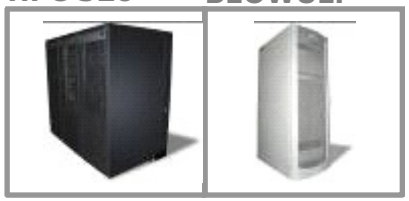
1999
HPC 4500 **STORAGETEK**

9,6 GFLOPS 51 TERABYTES

2001
SVG

9,9 GFLOPS

2002
HPC 320 **BEOWULF**



64 GFLOPS 16 GFLOPS

2003
SUPERDOME

768 GFLOPS

2004, 2005, 2006
SVG

3,142 GFLOPS

2007
FINIS TERRAE

16,000 GFLOPS

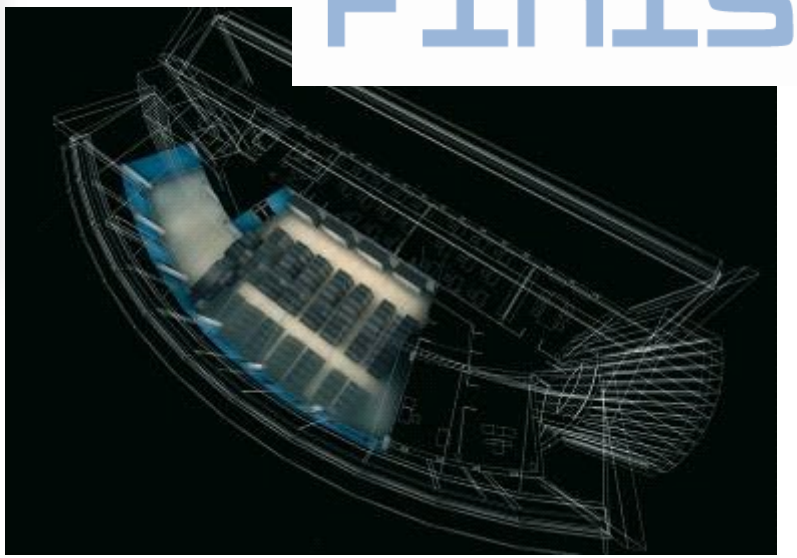
Installation Year	1993	1998	1999	2001	2002	2003	2004	2005	2006	2007
Capacity				SVG			SVG	SVG	2006	
Capability	VP2400	VPP300E AP3000	HPC4500		HPC320	SUPERDOME				FINIS TERRAE

Madrid, Marzo de 2008

CESGA XUNTA DE GALICIA CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS CSIC ERDF European Regional Development Fund

FINIS TERRAE (2007)

FINISTERRAE



New Server HPC (December 2007)

More than **16 TFLOPS** and **19TB RAM** Memory

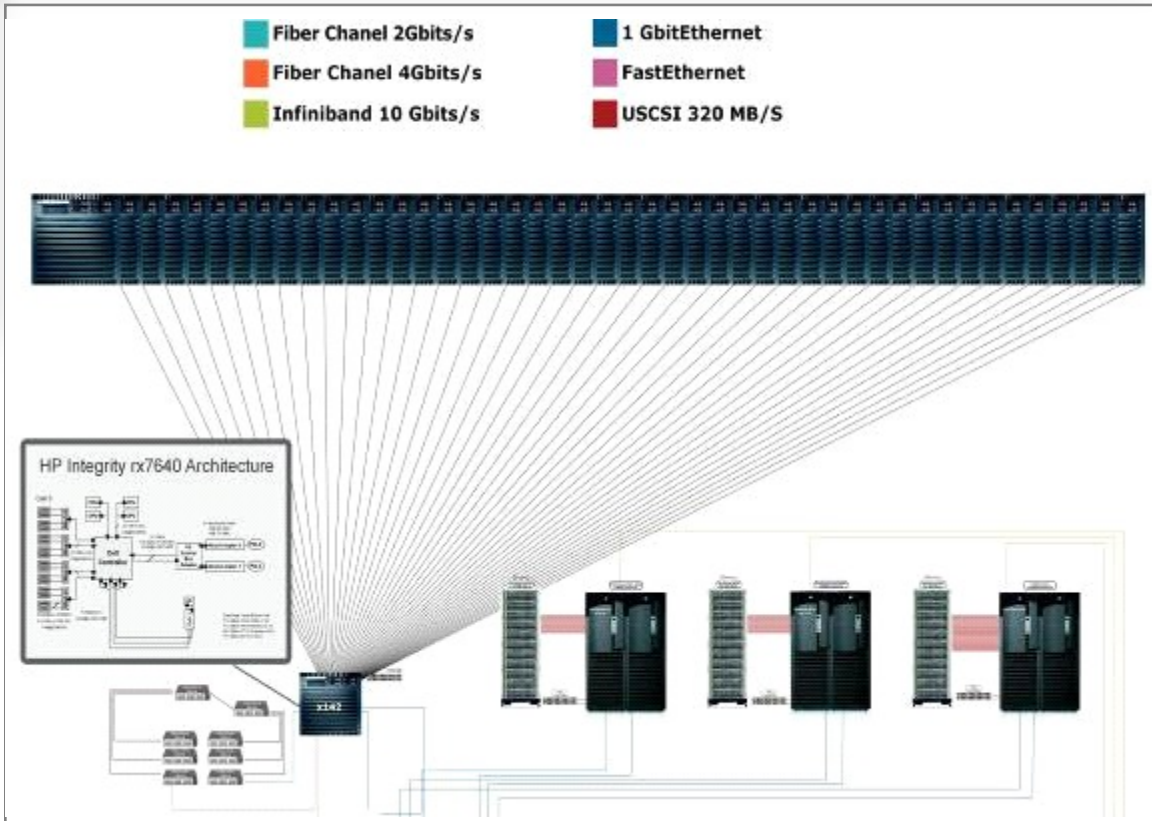
Plus Agreement among

intel hp invent CESGA

Madrid, Marzo de 2008

CESGA XUNTA DE GALICIA CSIC CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS ERDF European Regional Development Fund

FINIS TERRAE (2007) – COMPUTING NODES



SUPERCOMPUTING NODES:

144 cc-NUMA Nodes with Itanium CPUs connected through a high efficiency INFINIBAND network

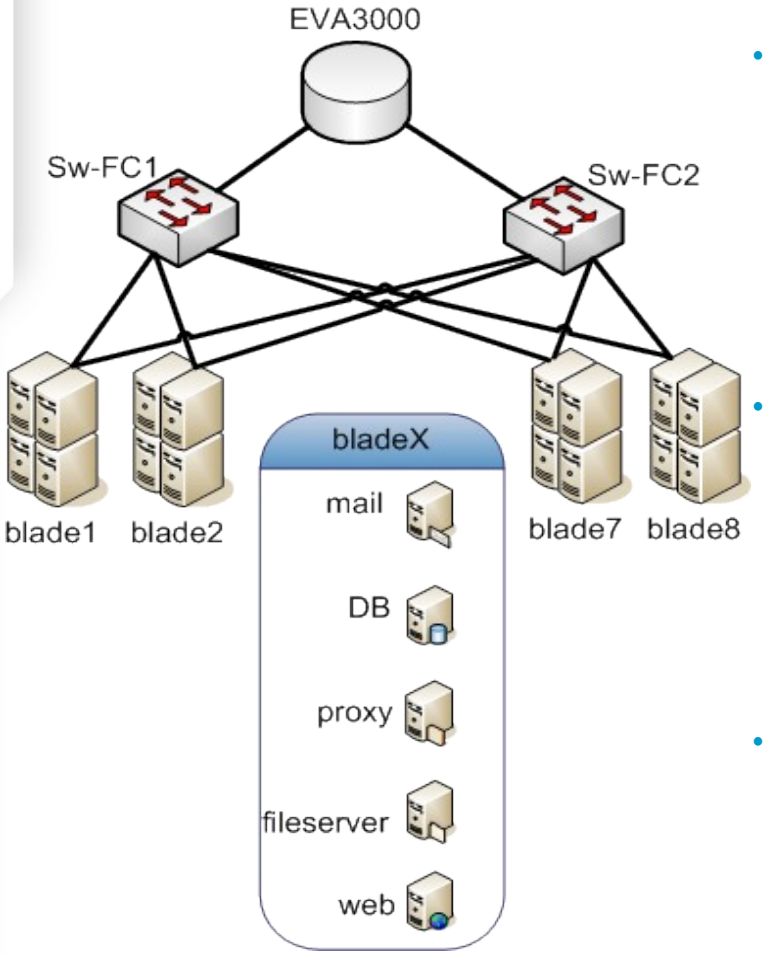
- **1 node: 128 cores, 1.024 GB memory**
- **1 node: 128 CPUs, 384 GB memory**
- **142 nodes: 16 cores, 128 GB memory**
- **2 nodes: 4 cores, 4 GB memory for testing**

Usos de la Virtualización en Cesga

- Consolidación de servidores
- Aprovisionamiento de servidores
- Uso en proyectos
- Computación cluster
- Grid
- Docencia. Virtualización de aulas

- Realidad
  - Existen más de 400 servidores en el CPD
  - Soportan servicios del CESGA y los proyectos en los que participa el CESGA (34 actualmente)
  - Prácticamente la totalidad están en S.O. linux, aunque en diferentes distribuciones (Suse, Redhat, Scientific Linux, Fedora)
  - Hay muchos servicios que requieren configuraciones distintas de los diferentes paquetes
  - Los mismos servidores se utilizan en diferentes proyectos
  - A veces cuesta mucho adaptar el software a hardware nuevo

Virtualización de Servicios: Infraestructura

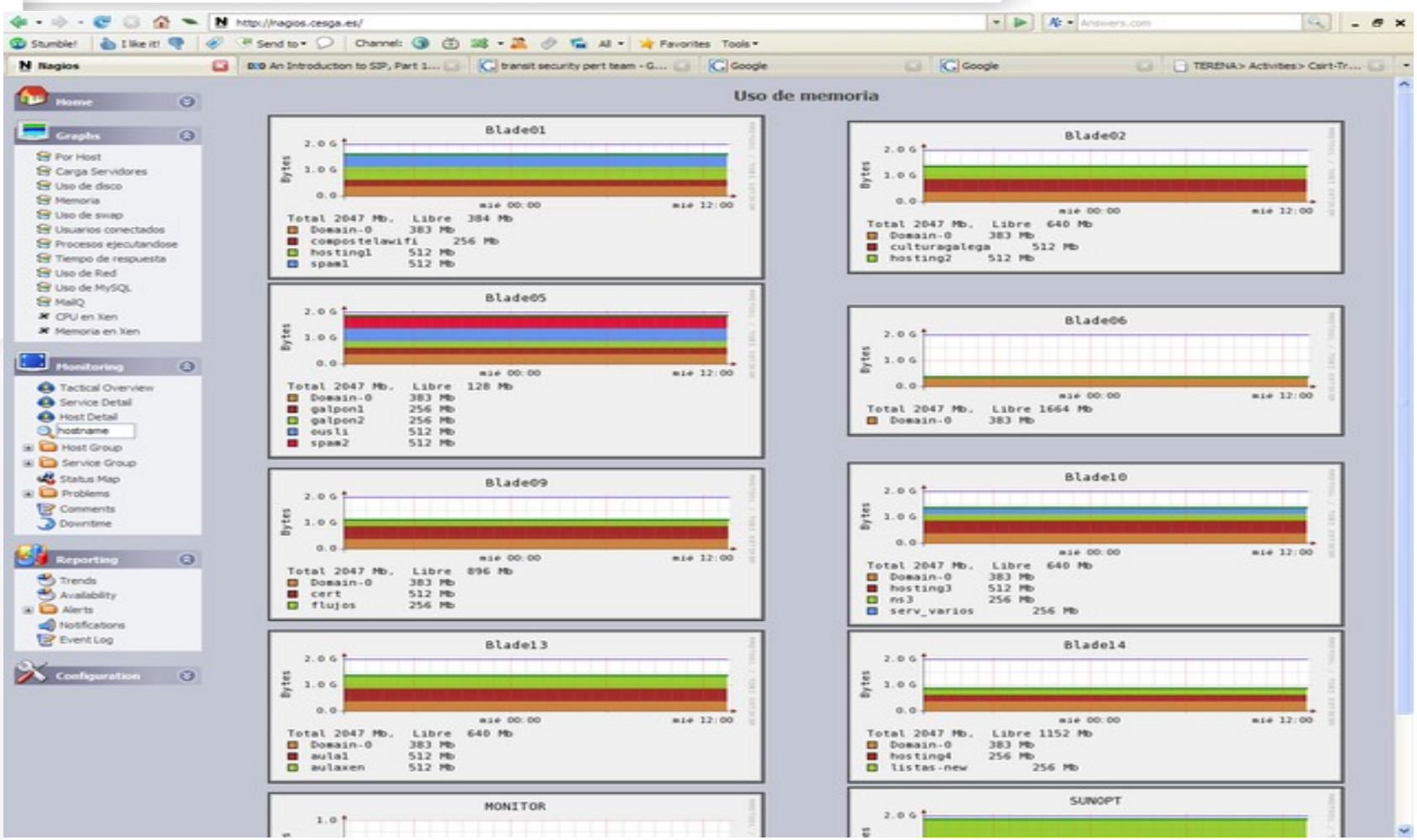


- Infraestructura:
  - Rack Blades HP Proliant (2GB RAM) – 16 en un mismo chasis.
  - Cabina de discos EVA3000
  - Switches Fiber Channel
  - Xen 3.03 (paravirtualización, mejor rendimiento)
- Almacenamiento compartido por todas las máquinas anfitrionas.
  - OCFS2 (Oracle Cluster FileSystem v2)
  - VBD (Virtual Backend Device) : fichero de imagen, mayor versatilidad sin usar LVM o GFS (cluster suite de RHEL)
  - Necesidad de sincronización entre anfitriones
- Configuración bridged (VM's en misma LAN que maquina anfitriona). Igualmente son válidas otras configuraciones más complejas

Servicios virtualizados

- > 80 webs alojados (hosting)
- email, listas de distribución
- streaming
- proxy regional (eduroam)
- news, ftp
- plataformas colaborativas/enseñanza (claroline/dokeos)
- VoIP (asterisk-trixbox)
- HostingX.cesga.es (VM base para servicios web)
- AntispamX.cesga.es (VM de procesado antivir/antispam)
- VM especificas de proyectos
- VM desarrollo y testing

Uso de memoria



Madrid, Marzo de 2008

CESGA

XUNTA DE GALICIA

CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

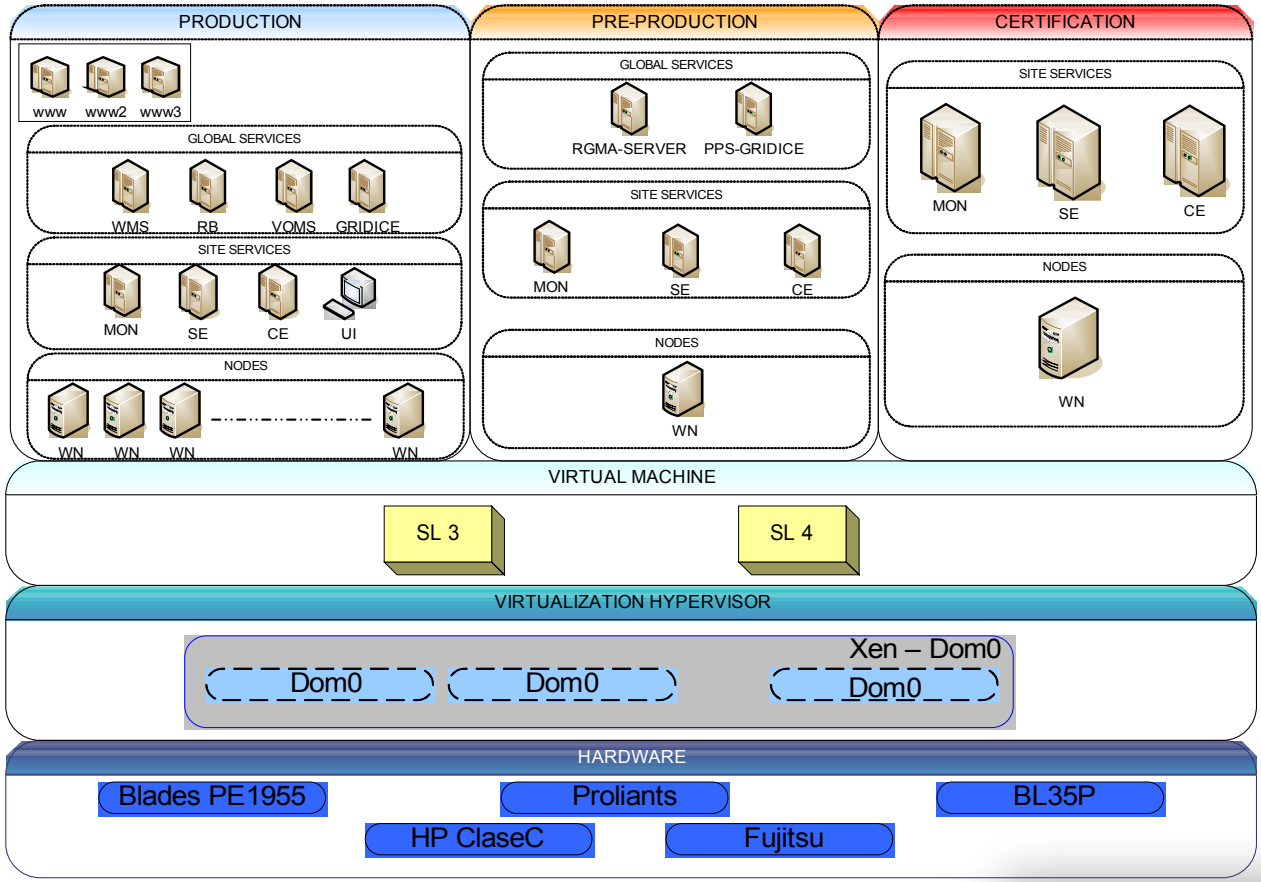
ERDF European Regional Development Fund

Ventajas e inconvenientes

- Ventajas
  - Rápido despliegue de servicios
  - Servicios complejos o especiales se sirven con una o varias VMs específicas.
  - Fácilmente escalable. Se pueden clusterizar VM's
  - Migración de máquina anfitriona en caliente (gracias al almacenamiento compartido). T < 0.5 seg (la VM mantiene las conexiones).
  - En gral. VM's pequeñas, facilmente clonables (copiar un fichero de unos cuantos GB)
  - Alto rendimiento (90-95% en comparación a sin virtualización)
  - Actualización de versiones
  - Facilita la recuperación ante desastres
- Inconvenientes
  - Más consumo de memoria.
  - El "backup" es algo más complejo, salvo que se haga directamente de las máquinas virtuales. (rsync, data protector)
  - Precaución al actualizar las maquinas anfitrionas (OCFS2 debe mantenerse con versiones homogeneas).
  - Aprovechamiento no óptimo del almacenamiento. (cada VM tiene su SO)

Virtualización en EGEE

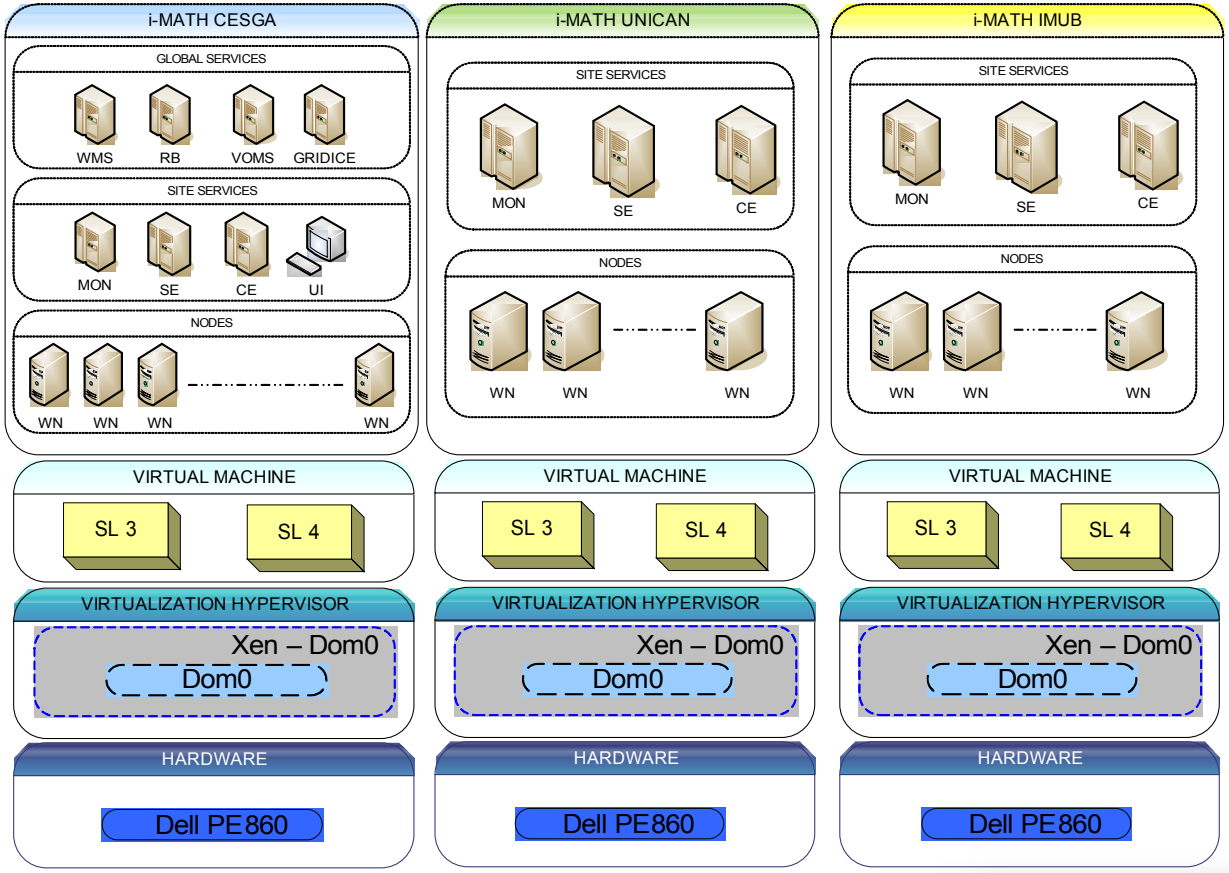
Varios clusters sobre la misma infraestructura (producción, preproducción, certificación), que a su vez requieren múltiples servicios.



Madrid, Marzo de 2008

Virtualización en Imath

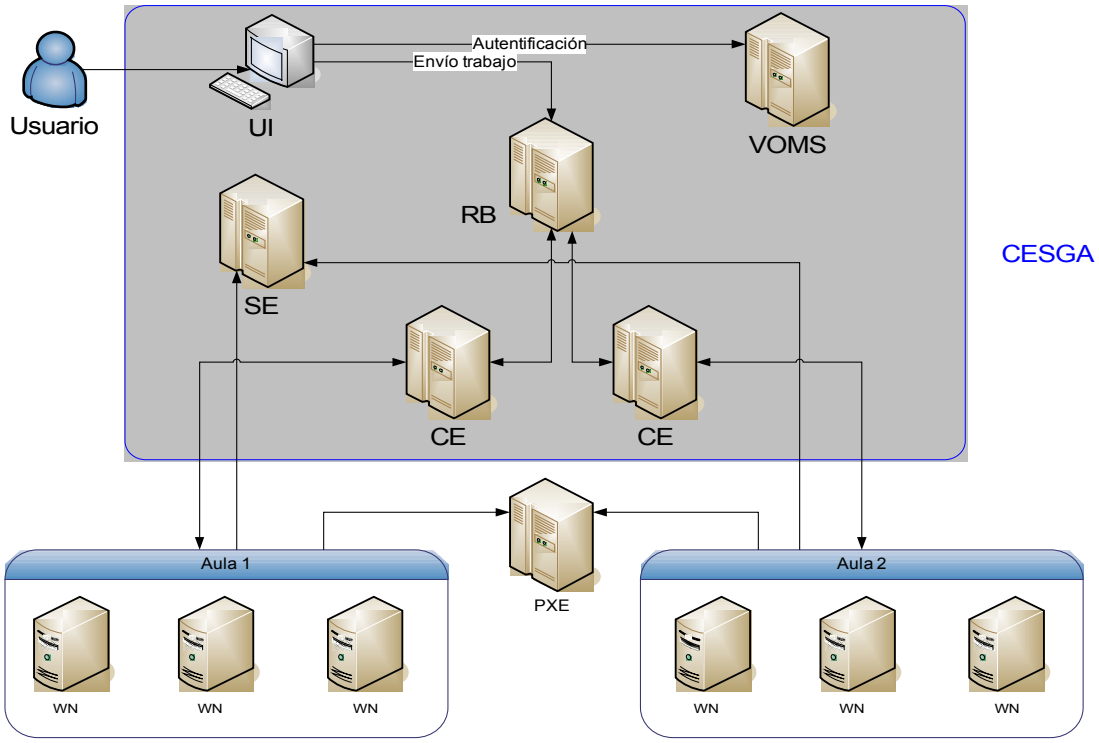
Despliegue de un grid para fines de desarrollo.
Necesidad de emular un cluster completo en un único servidor.



Madrid, Marzo de 2008

Formiga

Proyecto de reaprovechamiento de capacidad de computación de aulas informáticas en períodos de no utilización.
Necesidad de no interferir con la producción de las aulas.



Madrid, Marzo de 2008

Benchmarking the virtual grid

Description of the benchmarks :

- Linpack: numerical linear algebra. measures the time to solve a dense n by n systems of linear equations by Gaussian elimination with partial pivoting. The result is reported in millions of floating point operations per second (Mflop/s).
- Bonnie++: tests of hard drive and file system performance. Reports the bytes processed per elapsed second, per CPU second, and the % CPU usage (user and system).
- Iperf: is a tool to measure maximum TCP bandwidth. Iperf reports bandwidth, delay jitter and datagram loss.
- Gromacs: GROMACS is a package to simulate the motion of molecular systems by computing the Newtonian equations of motion of its particles. It can be run in parallel, using standard MPI communication.

Hardware & Software Configuration

HARDWARE:

- System: Dell Power Edge 1950 blade server
- CPU: two VT-enabled quad-core 1.6 GHz Intel Xeon 5310 CPUs (total eight cores) and 4 GB of RAM installed.
- IO: dualport 1Gbps Ethernet adapter and two 146GB SAS disk drives.

SOFTWARE:

- CESGA has select to use full virtualization to take advantage of Intel VT technology.
- Host: Linux OpenSUSE 10.2. Xen 3.0.3-0 release
- Virtual Machine: one virtual CPU and 512 MB RAM. Scientific Linux

Results obtained

- Linpack: The difference in the benchmarks was only of 3% less performance which can be acceptable in our environment.
- Bonnie++: the performance of the virtual server is 5-10 times slower than the host server, huge impact on the performance of the system when running grid jobs.
- Iperf: We measured 801 Mbits/second amount the physical servers that was down to 37.7 Mbits/sec when running the same benchmark across virtual servers. Enough bandwidth for typical applications but can be a trouble for parallel applications using MPI.

Results obtained (cont.)

Gromacs:

- The same application running on a single CPU is only 2.8% slower when using a virtual server than when using a physical server.
- When we try to use more than one server, we reach a good scalability when using physical worker nodes (71% of peak scalability) but the results are terrible bad when using virtual worker nodes, and in the later the application doesn't scale at all. This is a quite surprising because the network tests that we run with iperf don't show such bad results

Latency of the network:

- latencies are 2 orders of magnitude bigger in the virtual working nodes that in the physical working nodes
- In particular, there is a big latency when the information is transferred from the virtual server to its xen host
- when using paravirtualization, the results are much better.

Benchmarking: conclusions

- Network and disk performance issues that can represent a big difference to some applications, specially in parallel computing.
- Network and disk virtualization in Xen are not mature enough by now to be implemented in a production environment but are more a quick response to implement hardware virtualization with some issues not yet resolved.
- This technology has to be implemented in other high-end hardware as low latency networks like Infiniband or Myrinet to be really useful in high performance computing.

Conclusiones

- La virtualización es una tecnología que permite una gran flexibilidad en el despliegue de servicios, adaptándolos muy fácilmente a diferentes entornos hardware. Igualmente, facilita la utilización de características ampliamente demandadas en nuestro entorno, pero difíciles de implementar en la práctica, como por ejemplo checkpoint/restart.

- En entornos de cálculo de altas prestaciones, la penalización que introduce puede desaconsejar su uso, bien por la necesidad de manejo de dispositivos como disco o red, o incluso por el impacto en uso de CPU.