CESGA HPCN 2010

High Performance Computing Networking

NEW COMPUTING & STORAGE SERVICES AT CESGA

Santiago de Compostela, November 2010

Carlos Fernández Sánchez
(Systems Administrator, Galicia Supercomputing Centre)

CESGA

CSIC
CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

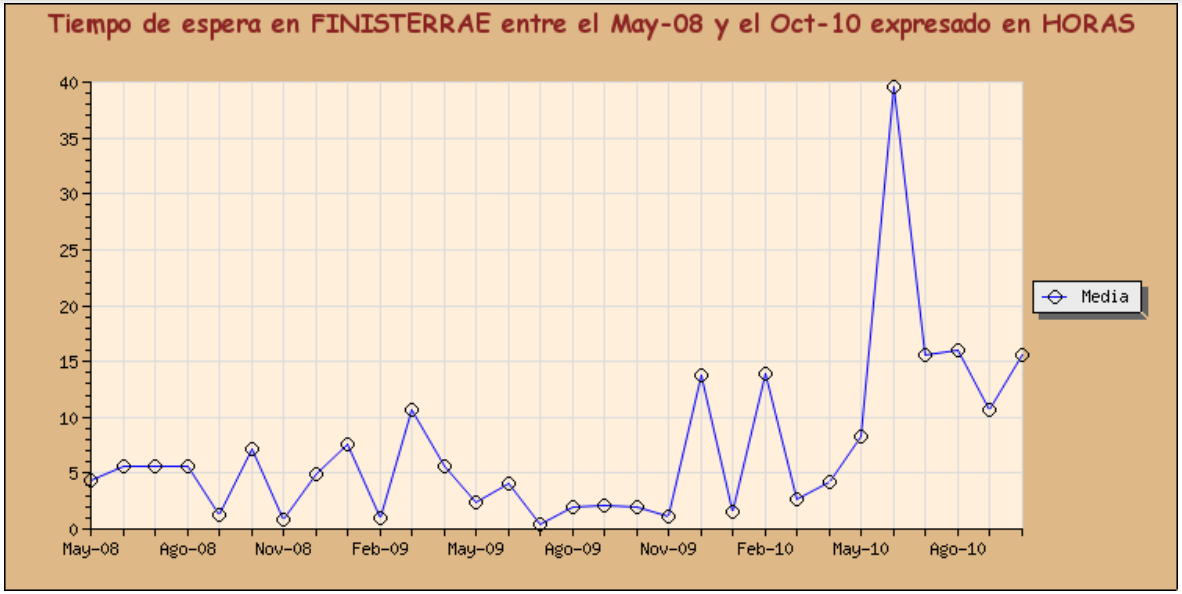
XUNTA DE GALICIA

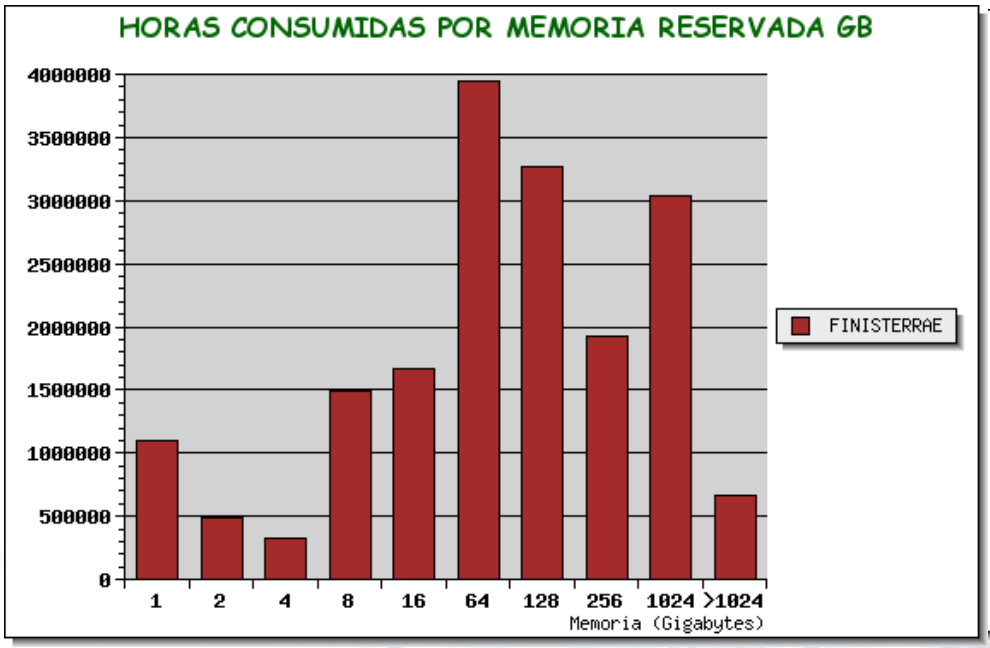
UNIÓN EUROPEA
FEDER
FONDO EUROPEO DE DESENVOLVEMENTO REXIONAL

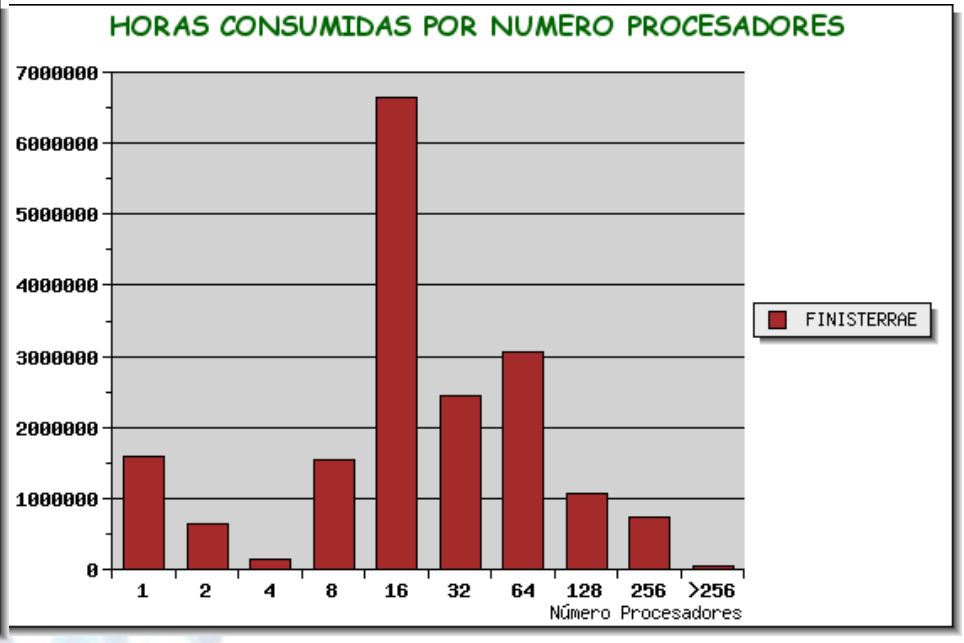
Agenda

- SVG 2010
  - Motivation
  - Hardware configuration
  - Actualization Plan
  - Benchmarks
- CESGA Supercomputers in 2011
  - Distribution of jobs
- Green Computing
- New Storage
- Questions

SVG Upgrade - Motivation







SVG Upgrade 2010 – Throughput – Pre & Post Processing - Visualization

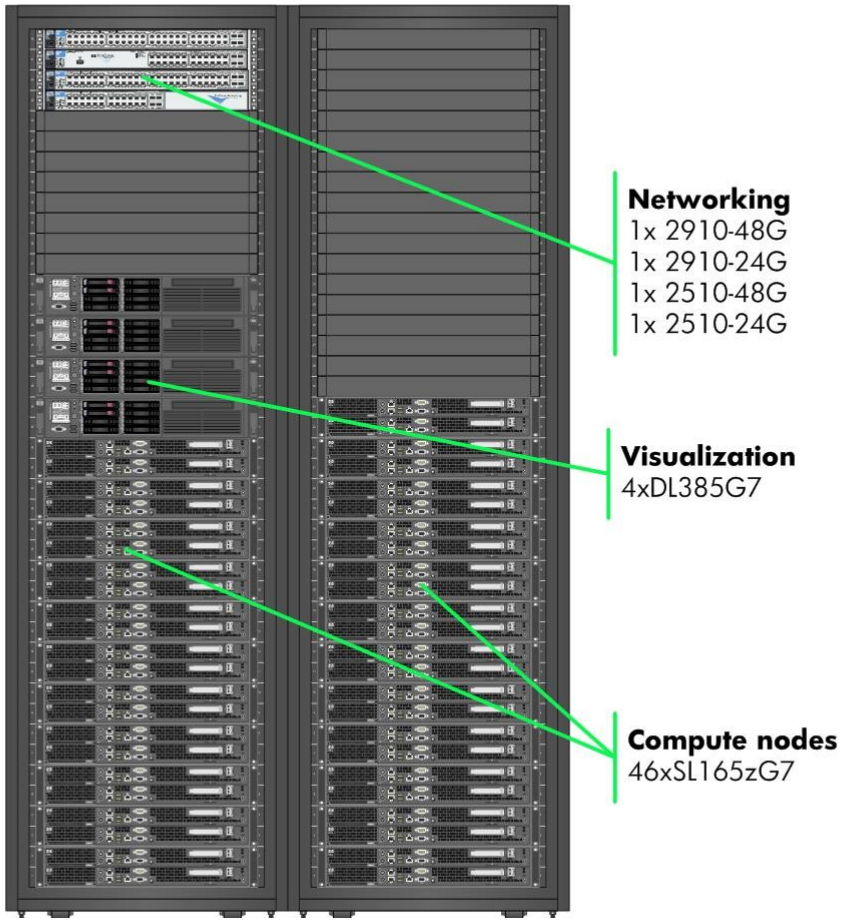
46 HP ProLiant SL165z G7 – Compute node (node configuration)	
Processor	2 x AMD Opteron™ Processor 6174 (2.2 GHz, 12MB Level 3 Cache, 80W, 12 cores per processor)
Memory	64 GB (19 nodes) & 32 GB (23 nodes)
Performance	211 GFlops
Storage	1 x 500GB 3G SATA 7.2K NHP
Networking	2 Embedded Dual Port Intel 1GbE NICs (4 total NICs)
I/O Expansion	1 x16 LP PCIe Gen 2 slot 1 x4 LP internal PCIe Gen 2 slot
Management	LO100i Management IPMI 2.0 & DCMI

4 HP DL 385 G7 - Visualization node (node configuration)	
Processor	2 x AMD Opteron™ Processor Model 6174 (2.2 GHz, 12MB Level 3 Cache, 80W)
Memory	64 GB PC3-1333R
Storage	1 x 2TB 3G SATA 7.2K NHP
Networking	2 Embedded Dual Port Intel 1GbE NICs (4 total NICs)
Vis. Card	ATI Firepro V5700

CESGA CSIC XUNTA DE GALICIA UNIÓN EUROPEA FEDER FONDO EUROPEO DE DESENVOLVEMENTO REXIONAL

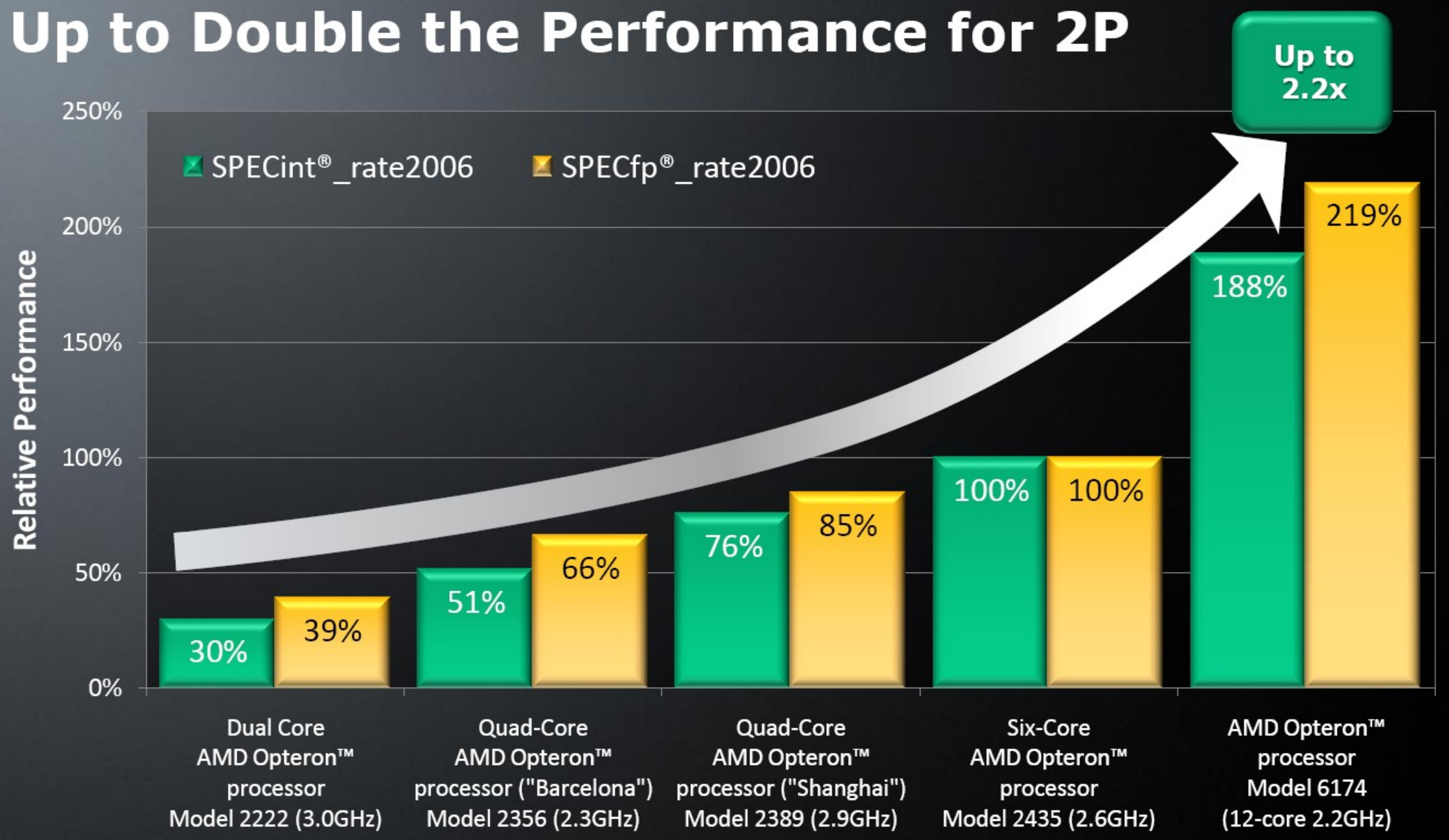
SVG Upgrade 2010

Total	
Processor	1,200 cores 2.2GHz
Memory	2,400 GB
Storage	31 TB
Networking	NFS GbE & Dedicated MPI GbE
Performance	10,240 GFlops
Power consumption	21KW
CPU hours	10 Million hours (50% increase over 2010)
Perf./KW	487GFlops/KW
Compilers	GNU & x86 Open64



AMD OPTERON 6174

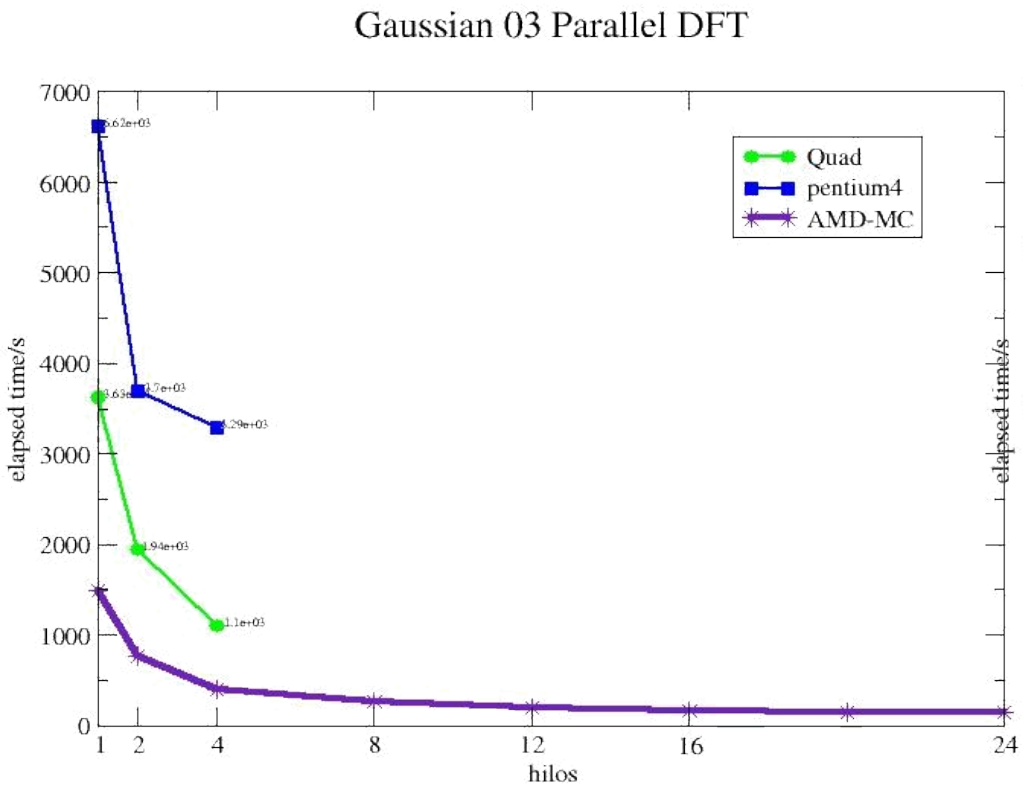
Up to Double the Performance for 2P

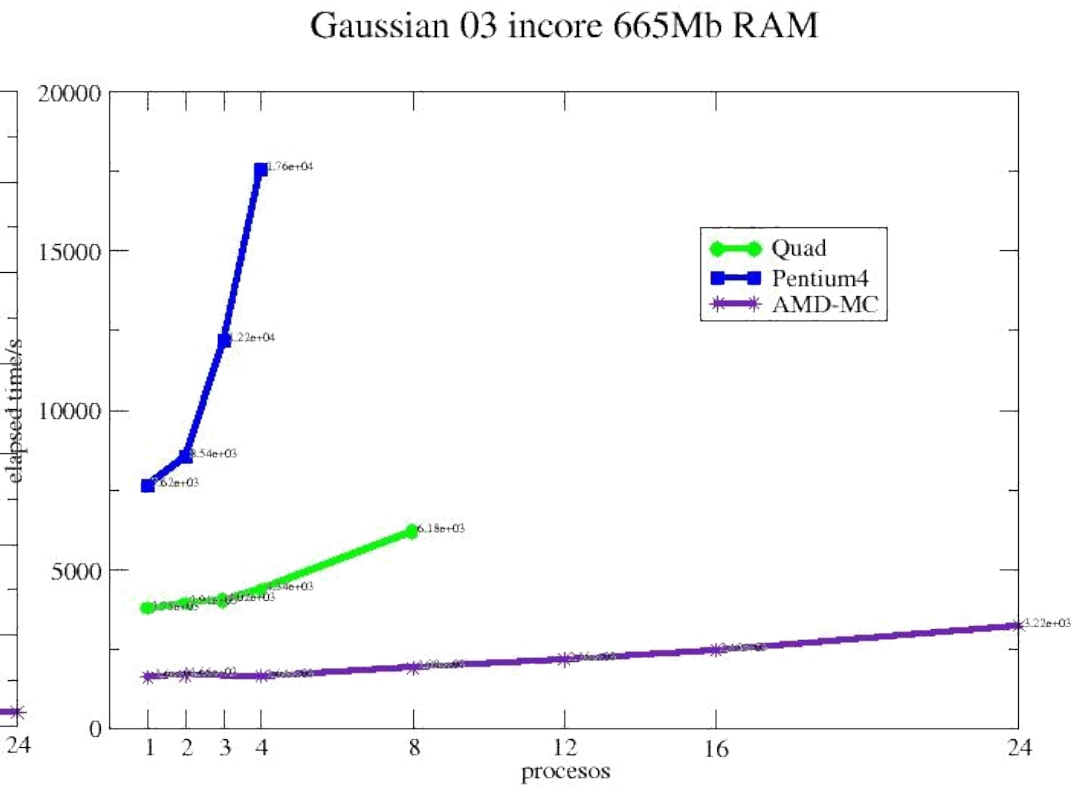


SPEC, SPECint, and SPECfp are registered trademarks of the Standard Performance Evaluation Corporation. The results for AMD Opteron™ processor Model 6174 are based upon data published on http://www.spec.org/cput2006/results as of 10/8/2010. The comparison presented above is based on the best performing two-socket servers using the specified processor model. For the latest SPECint_rate2006 and SPECfp_rate2006 results, visit http://www.spec.org/cput2006/results.

AMD The future is fusion

Benchmarks





x4 better Perf./core
x12 less Watts/core
x48 better Perf./watt

Actualization Plan

NODES: 80 x PowerEdge 750

Processor: Intel P4 3,2 GHz

RAM: 2Gb

HD: 1x SATA 160 Gb

Networking: 1x Dual 10/100/1000 Ethernet

Install date: 2004

CONTROL NODE: 1X PowerEdge 1750

Processors: 2X Intel Xeon 3,06-1 Mb Cache

RAM: 2Gb

HD: 3x SCSI ULTRA 320 146 Gb

Networking: 4x10/100/1000 Ethernet

Power Supply: redundant

Networking: 4x PowerConnect 2624

24 ports Gigabit

NODES: 40 x PowerEdge 1955

Processor: Doble Intel Xeon 5130 quad core 1,6 GHz, 8MB cache

RAM: 4Gb & 8Gb (4 nodes)

HD: 1x SAS 72 Gb

Networking: 2x Dual 10/100/1000 Ethernet

Power Supply: redundant

Install date: 2006

CONTROL NODE: 1X PowerEdge 2950

Processors: 2X Intel Xeon 5130 2 GHz

RAM: 4Gb

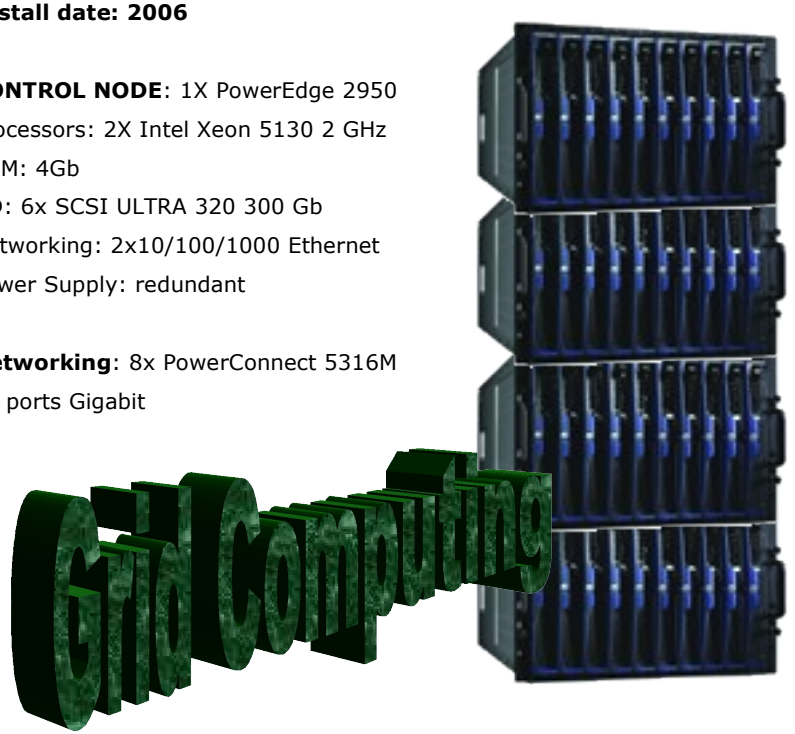
HD: 6x SCSI ULTRA 320 300 Gb

Networking: 2x10/100/1000 Ethernet

Power Supply: redundant

Networking: 8x PowerConnect 5316M

32 ports Gigabit



Actualization Plan

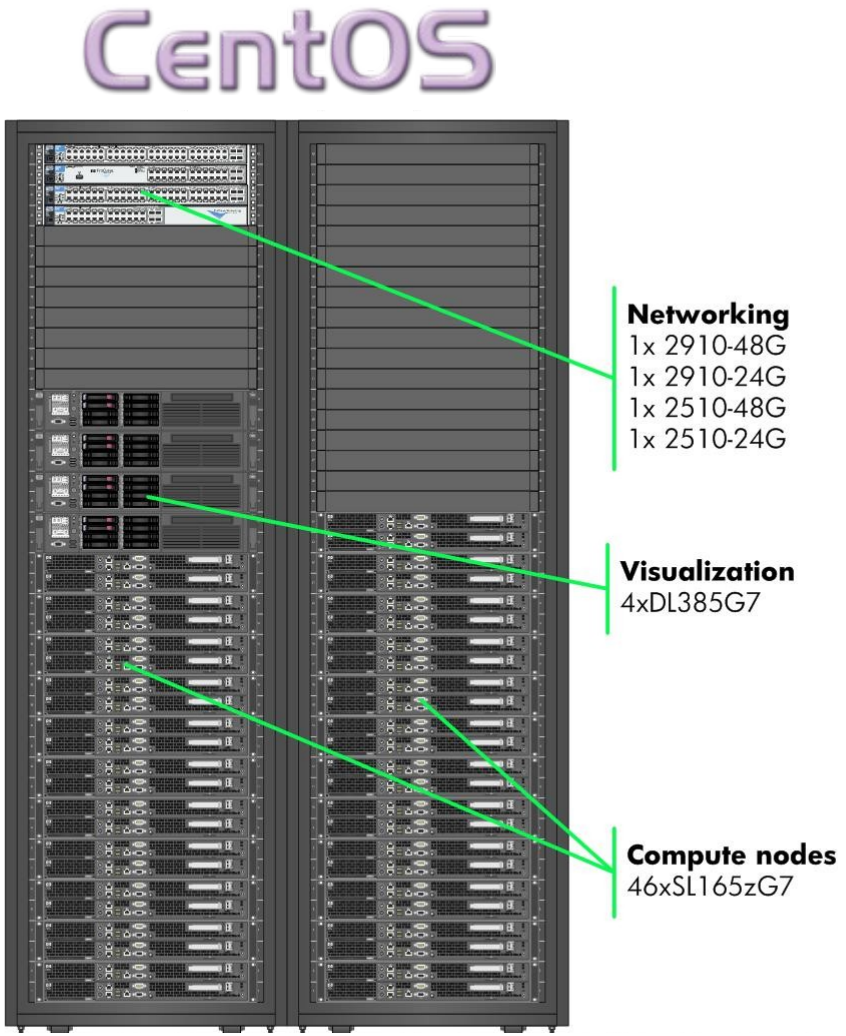
Everything Centos 6, 64-bit

Same Centos image on all nodes

Virtualization “ready” (system ready to run VM)
- Could run “custom clusters”

Users login ssh *svg.cesga.es* (new system)
- Current test/virtual machine
- Will run as a proxy to access the system

Use “*compute*” to access *interactive system*
- *Interactive system* are Visualization nodes
- Pool of compute nodes available with “compute”
  - On-demand if needed



Actualization Plan

Current “svgd” compute nodes maintained

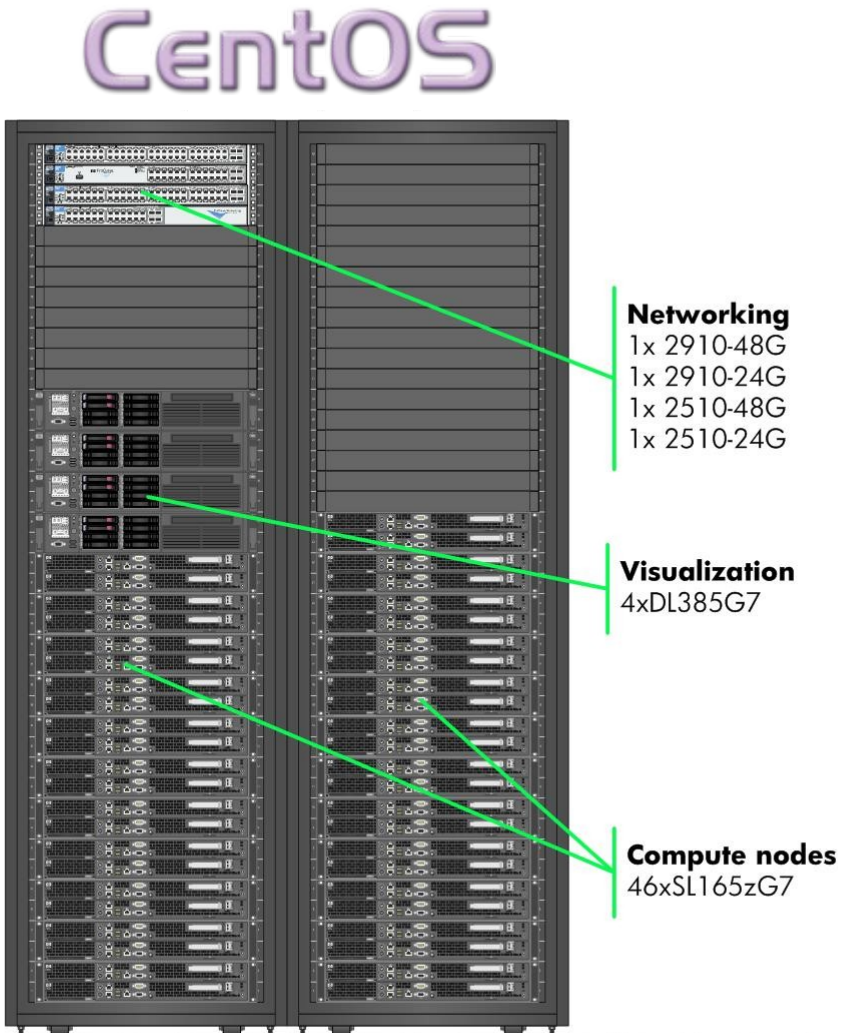
- Until all users running on the new SVG
- Current applications available
- No new applications from January 2011
- Users can login to *svgd.cesga.es* (old system)
- Shared home filesystem (same)

Batch system upgraded to GE6.2u5

- Better job to core binding functionalities
- Improve core binding over kernel default

Blade quad-core nodes available to Grid computing

- Migrated to GLITE 3.2 64 bits in Q2’2011



Network & Connectivity

CESGA Network

10GbE

4x DL385 22x SL165z svgd2 (NFS home)

Procurve 2910
48 ports

Storage Network

10GbE

Procurve 2910
24 ports

24x SL165z

MPI Network

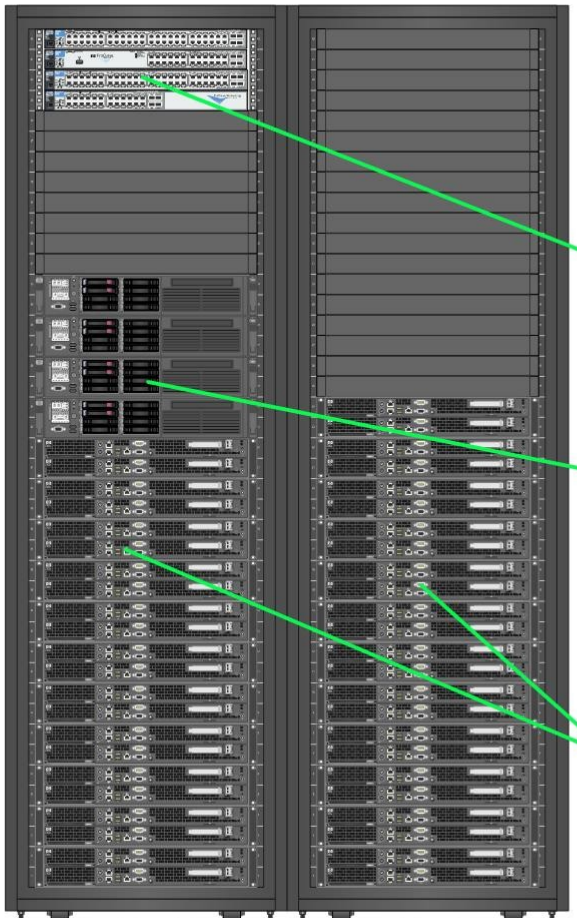
Procurve 2510
48 GbE ports

46x SL165z MPI GbE

Admin Network

PowerConnect 2624

46x SL165z + 4x DL385 ILO



Networking
1x 2910-48G
1x 2910-24G
1x 2510-48G
1x 2510-24G

Visualization
4xDL385G7

Compute nodes
46xSL165zG7

Actualization Plan - Schedule

Test node installed by this week

Configure batch system and storage

User accounts

Start installing and testing software next week

Configure batch system

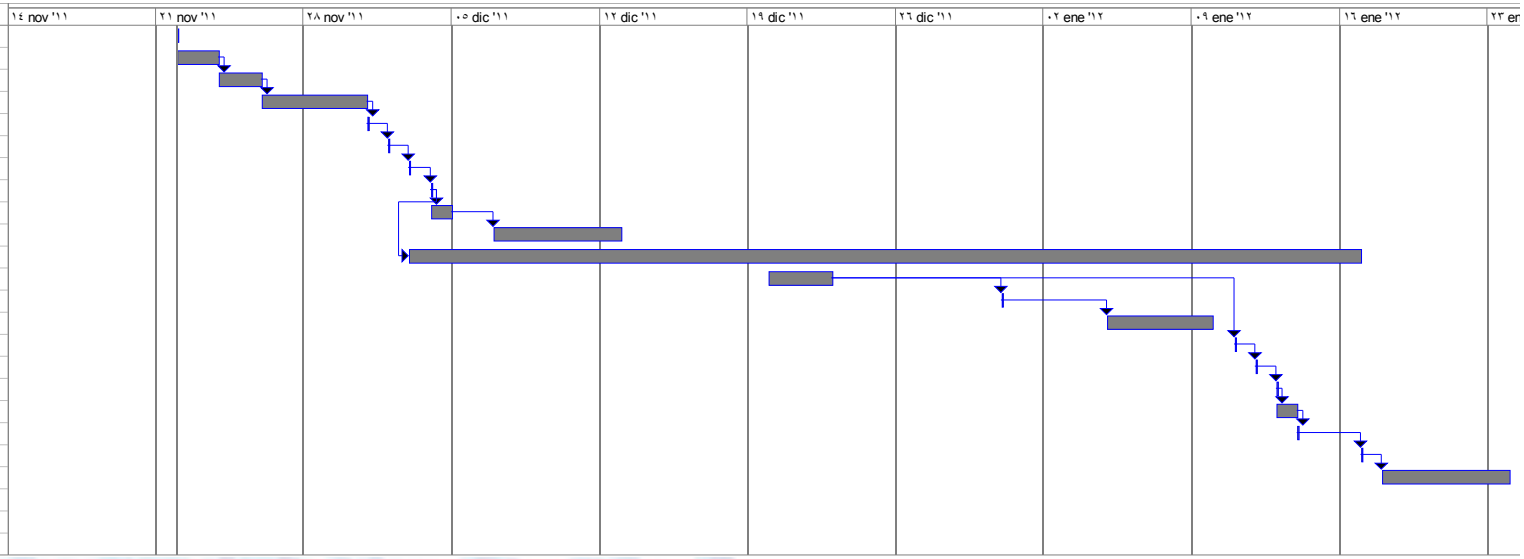
System arrives on Christmas

Installation and initial configuration by Mid-January

First test users by January, the 17

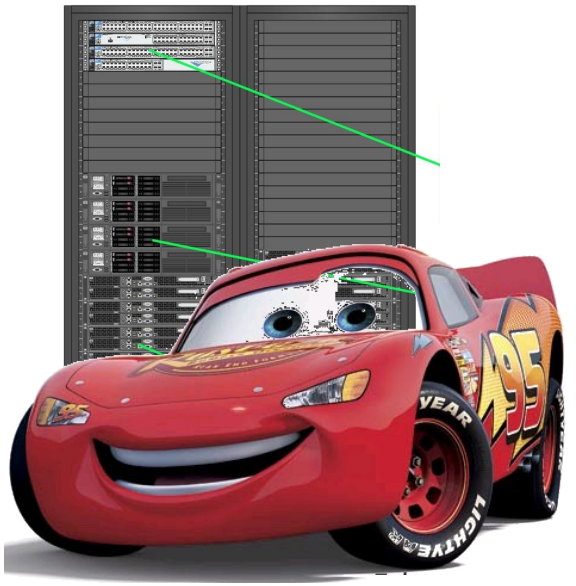
System in production in February

Id	Nombre de tarea	Comienzo	Fin
١	Preparar sistema migración	٢١/١١/١٠ ١١:٠٠	٢١/١١/١٠ ١٩:٠٠
٢	Instalar RH٦	٢١/١١/١٠ ١٩:٠٠	٢٣/١١/١٠ ٣:٠٠
٣	Sistema de colas	٢٣/١١/١٠ ١١:٠٠	٢٥/١١/١٠ ١٩:٠٠
٤	Montar home	٢٥/١١/١٠ ١٩:٠٠	٣٠/١١/١٠ ٣:٠٠
٥	Cuentas de usuario	٣٠/١١/١٠ ٣:٠٠	٣٠/١١/١٠ ١١:٠٠
٦	Software de sistema	٠١/١٢/١٠ ١١:٠٠	٠١/١٢/١٠ ١٩:٠٠
٧	Compiladores y herramientas	٠٢/١٢/١٠ ١١:٠٠	٠٢/١٢/١٠ ١٩:٠٠
٨	Comenzar instalacion aplicaciones	٠٣/١٢/١٠ ١١:٠٠	٠٣/١٢/١٠ ١٩:٠٠
٩	Generar imagen nodos cluster	٠٣/١٢/١٠ ١٩:٠٠	٠٤/١٢/١٠ ٣:٠٠
١٠	Pruebas de envio de trabajos	٠٦/١٢/١٠ ١١:٠٠	١٢/١٢/١٠ ١٩:٠٠
١١	Compilacion de aplicaciones	٠٣/١٢/١٠ ٠:٠٠	١٦/٠١/١١ ٨:٠٠
١٢	Llegada cluster	١٩/١٢/١٠ ١١:٠٠	٢٢/١٢/١٠ ١٩:٠٠
١٣	Conexion electrica	٣٠/١٢/١٠ ١١:٠٠	٣٠/١٢/١٠ ١٩:٠٠
١٤	Cableado de red	٠٤/٠١/١١ ١١:٠٠	٠٩/٠١/١١ ١٩:٠٠
١٥	Cableado de gestion	١٠/٠١/١١ ١١:٠٠	١٠/٠١/١١ ١٩:٠٠
١٦	Pruebas de hardware	١١/٠١/١١ ١١:٠٠	١١/٠١/١١ ١٩:٠٠
١٧	Instalacion de imagenes PXE	١٢/٠١/١١ ١١:٠٠	١٢/٠١/١١ ١٩:٠٠
١٨	Montaje de sistema de ficheros	١٢/٠١/١١ ١٩:٠٠	١٣/٠١/١١ ٣:٠٠
١٩	Pruebas de envio de trabajos	١٣/٠١/١١ ١١:٠٠	١٣/٠١/١١ ١٩:٠٠
٢٠	Configuracion del entorno interactivo	١٦/٠١/١١ ١١:٠٠	١٦/٠١/١١ ١٩:٠٠
٢١	Usuarios en pruebas	١٧/٠١/١١ ١١:٠٠	٢٣/٠١/١١ ١٩:٠٠
٢٢			
٢٣			
٢٤	Sistema en produccion	٠١/٠٢/١١ ١١:٠٠	٠١/٠٢/١١ ١٩:٠٠



CESGA Supercomputers

Shared storage: /home
/COMPARTIDO /sfs
Linux O. S.
Grid Engine Batch Scheduler



Finis Terrae (FT)

Capability computing

Parallel jobs (>4 ... 1024 cores)

Huge memory (>4... 1024GB)

Huge parallel scratch (>50... 10,000GB)

Superordenador Virtual Gallego (SVG)

Throughput and Capacity computing

Sequential & parallel jobs
up to 24 cores per node!

Low-medium-large memory (up to 64GB!)

Medium single node scratch (<500GB)

Customized clusters – Cloud services

Green Computing

CESGA policy on energy efficient

Improve IT Energy efficiency

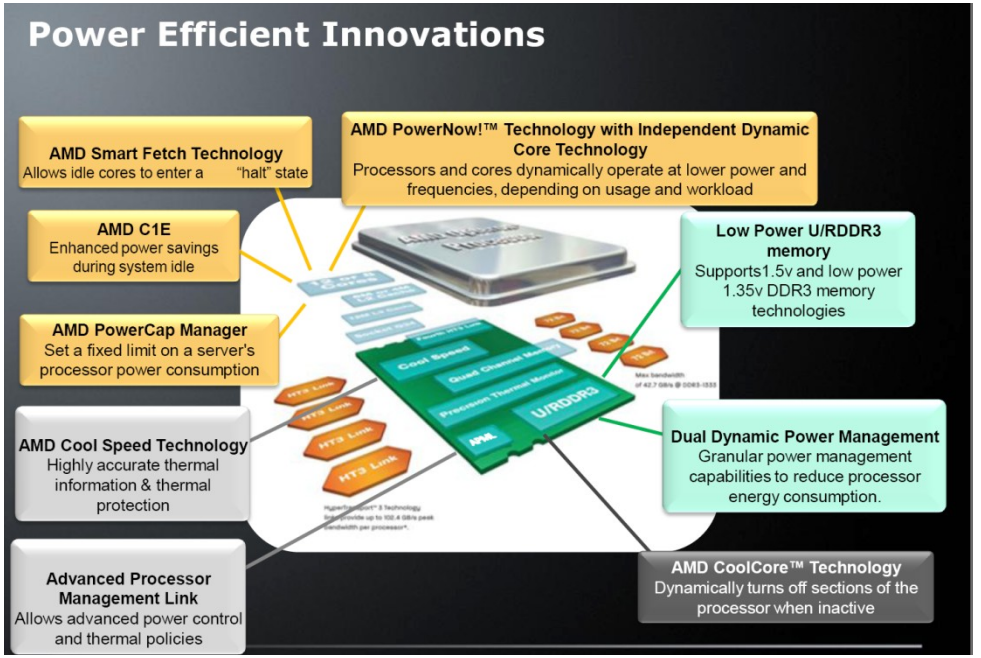
- Install power-efficient hardware
  - Valued in the public tender
- Power-off unused or low-efficiency computers
  - Beowulf (2002)
  - SVG (2004)

Current PUE is 1.8

- 1 KW of IT needs 0.8KW infrastructure
  - (UPS & cooling mainly)

Plan to Improve to 1.5 by 2015

- Every 0.1 means 38,016€/year



Green Computing

Some measures taken:

- Increase Data Center room temperature
  - From 20ºC to 25ºC (ASHRAE)
- Aisles enclosure
- Use freecooling and increase #freecooling hours
- Infrastructure monitoring and assesment
- Reduce number of working air movers
  - from 8 to 6 (2 as backup)

Other measures to be taken:

- Announce PUE in CESGA web
- VFD Fans
- Change UPS to high-efficiency
- Increase room temperature
  - 27ºC
- Analyse other cooling technologies
  - Water-cooling
  - Indirect-air cooling

Storage Upgrade 2011

Characterisitics	
Storage	> 70TB home fileystem
Performance	> 100,000 IOs
Bandwidth	> 4GB/s
Functionalities	High availability Snapshots (faster restores) Tiered storage
Other	Central & Shared storage Archiving Use energy efficient hardware

Questions & Suggestions

Thanks for your time