HP-CAST 10, Singapur 2008

Finis Terrae: Large Itanium Cluster Experience

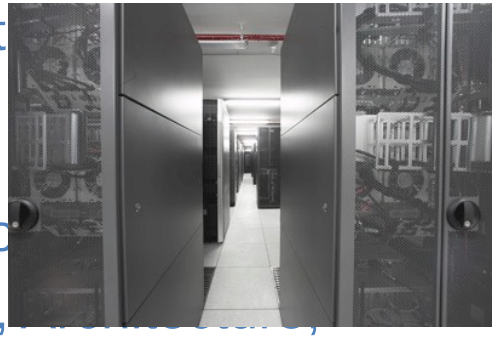
Sysadmin point of view

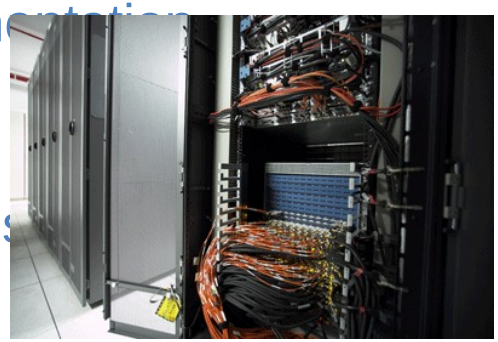
Carlos Fernández Sánchez
(HPC & Grid Systems Manager, Galicia Supercomputing Centre)

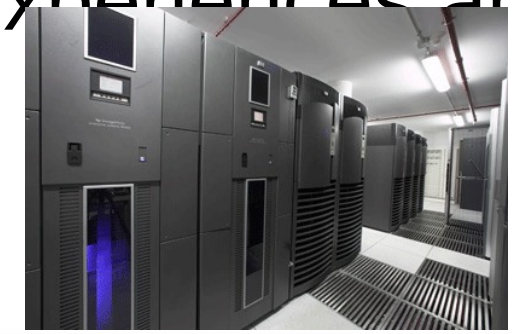
CESGA
XUNTA DE GALICIA
CSIC CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS
ERDF European Regional Development Fund

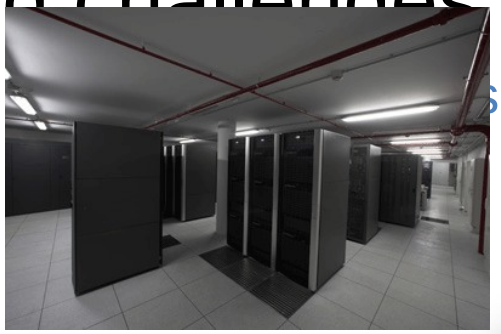
Agenda

- Introducing CESGA
  - Where, what, past, present
- Finis Terrae Constellation
  - #100 To
  - Desing, Implementation
- Experiences and challenges
  - sons









ESTABLISHED IN 1993 IN SANTIAGO DE COMPOSTELA (SPAIN)

CESGA

UNESCO World Heritage 1985
End of St. James Way
100,000 pilgrims in 2007

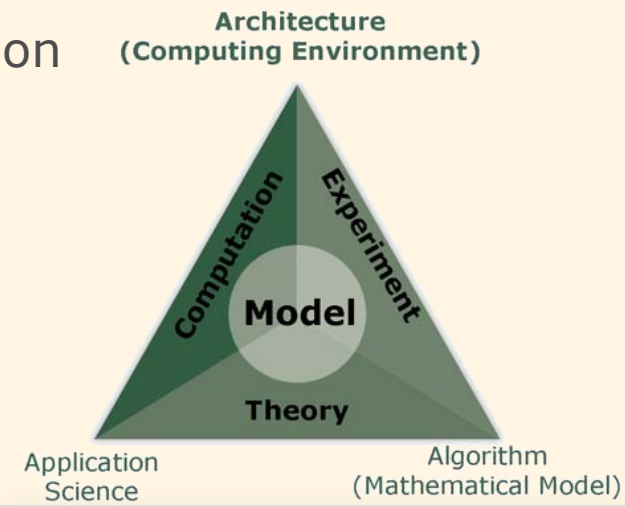
PAST, PRESENT, FUTURE (in 2 minutes)

Costumers: Three main Galician Universities and Spanish Research Council, Regional weather forecast service

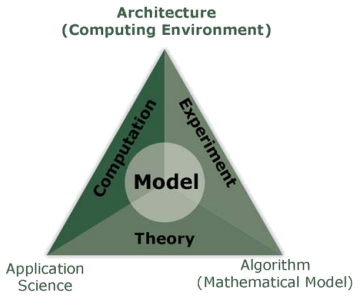
Services: High performance computing, storage and communication resources (RedIris PoP)

Promote new information and communication technologies (HPC & Grid projects)

Future: Centre of Excellence in Computational Science – C²SRC
141 research staff
75 MM€ (31% building, 23% HPC)



NEW CENTER STRUCTURE



R&D Excellence Centers

- Centro de Investigación en Ciencias del Mar.
- Centro de Investigación en Ciencias y Tecnologías de la Vida.
- Centro de Electrónica para Vehículos Inteligentes.
- Centro Hispano-Portugués de Investigación en Nanotecnología.

- Life Sciences
- Nanotechnology
- Ocean Sciences
- New Energies

HPC + Simulation

Application Areas

CESGA - C²SRC

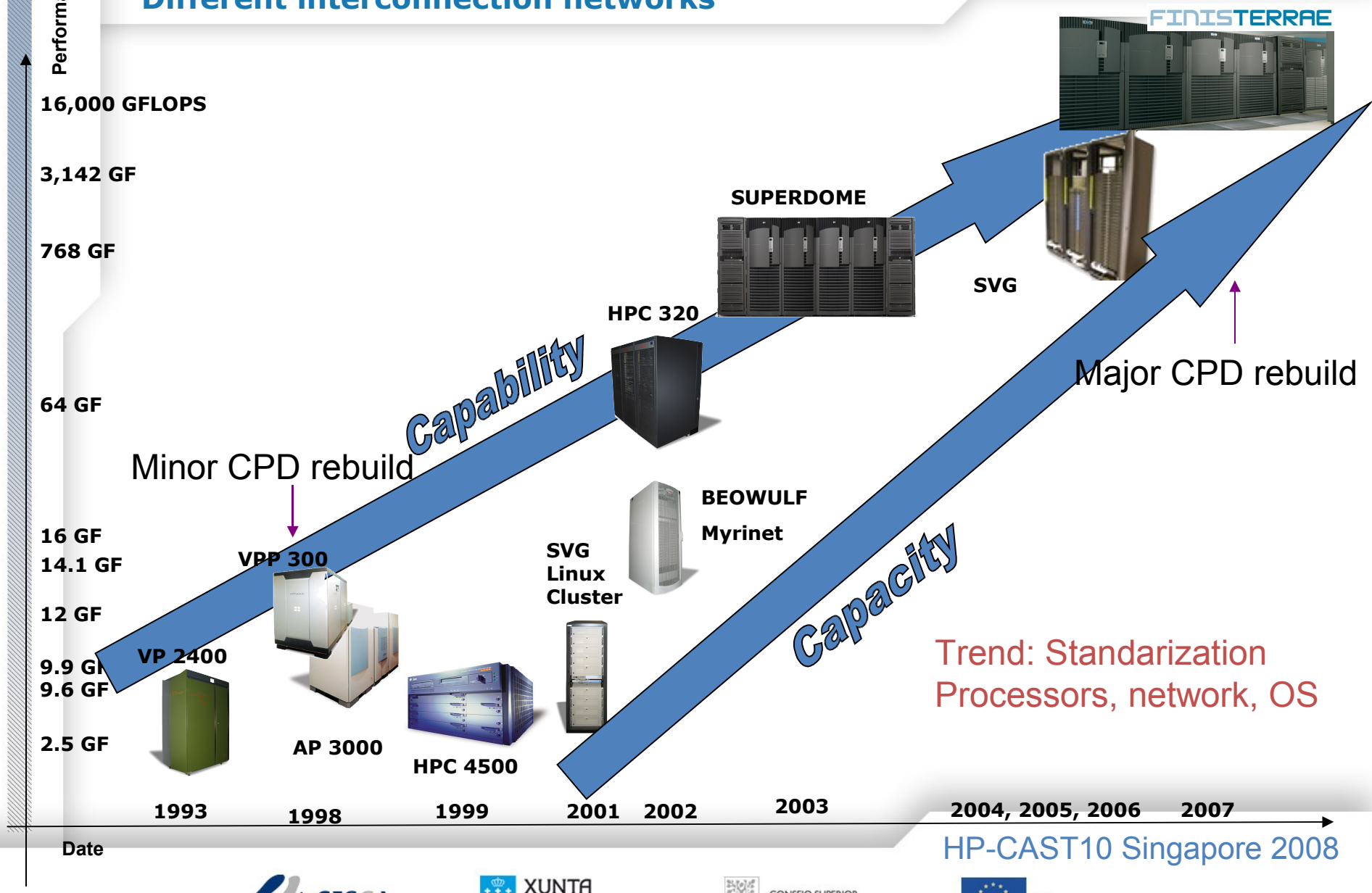
R&D Galician Plan 2006-2010

- Biotechnology.
- Nanotechnology.
- ICT.
- New Energies.

Strategic Actions R&D Spanish National Plan 2008-2011

- Health.
- Biotechnology.
- New Energies and Climate Change.
- ICT.
- Nanoscience and Nanotechnology.

CESGA'S TECHNOLOGICAL EVOLUTION (PAST & PRESENT)
Different architectures different applications
Different interconnection networks



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

FUTURE: NEW BUILDING 2010

FINISTERRAE DESIGN (2006)

THE BEST ARCHITECTURE FOR:

ALMOST ALL APPLICATIONS

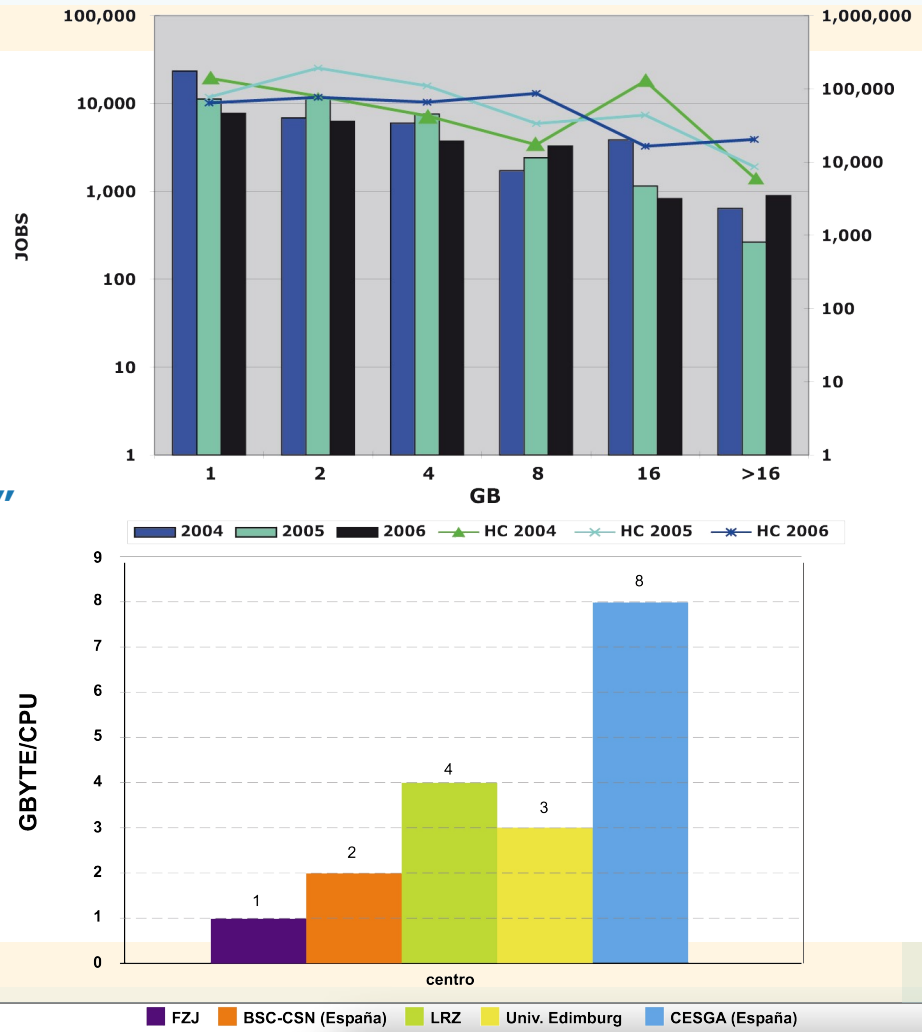
COMPUTING CHALLENGES
& "EVERYDAY" RESEARCH

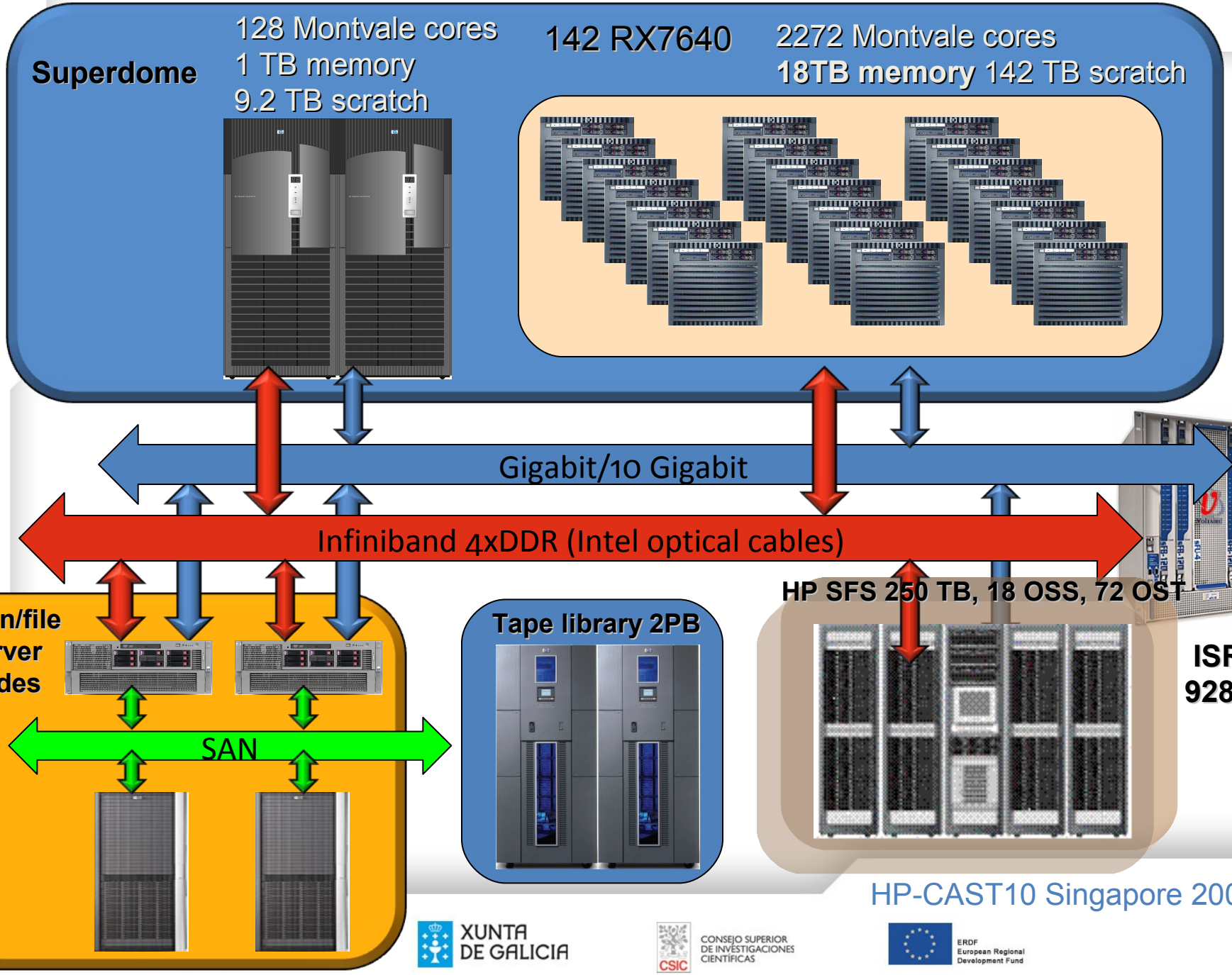
"CAPABILITY" AND "CAPACITY"

MINIMIZES "TIME TO SOLUTION"
(SIMPLER PROGRAMMING)

TAKES ADVANTAGE OF OLD, THOUGH EFFICIENT CODES

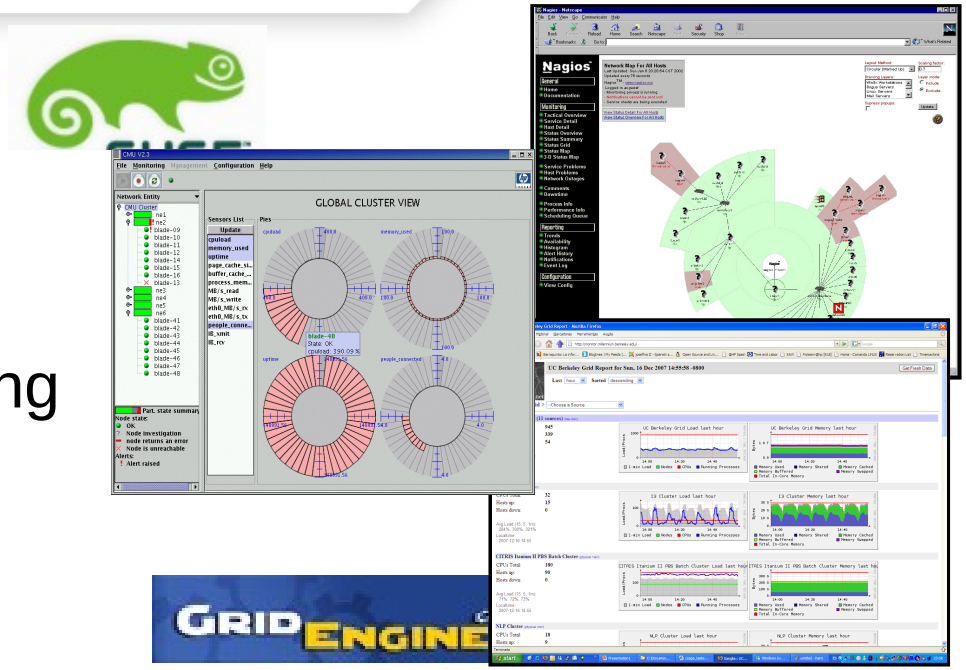
Highest GB/core in Europe
No Top500/Green500 focus





Software Stack

- Suse Linux 10sp1
- HP-CMU
- HP-SIM + RSP
  - Automatic hw ticketing
- Nagios & Ganglia
- N1GE batch system
- HP-Service Guard
  - (HA SSH, NFS, LDAP)
- Lustre 2.3beta (SLES10sp1)
- HP-MPI & Intel compilers
- RBB (on & off EVERYTHING)

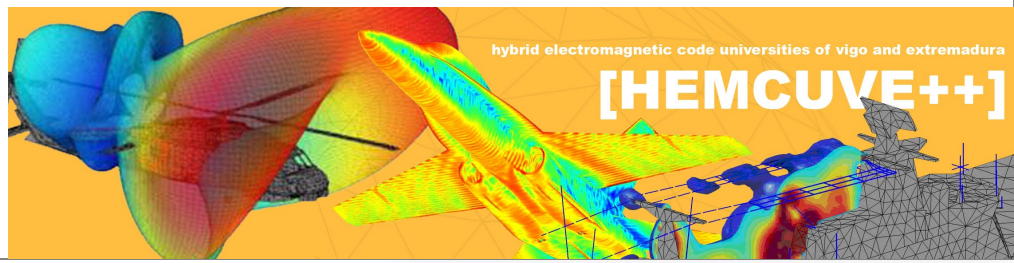


Finis Terrae installation

- Building adaptation CPD & Infra March-Oct 07& .
- System pre-integration in Germany Sep-Oct 07
- Systems arrive and HW install Nov-Dec 07
- SW configuration / acceptance tests Jan-Feb 08
- Challenges and local configuration Feb-March 08
- Started production 1st abril 2008
- Legacy Superdomes integration July 2008

APPLICATION AREAS AT CESGA
SOME CURRENT PROJECTS

- **Project: HEMCUVE++ Hybrid electromagnetic Code Universities of Vigo and Extremadura:**
  - Fernando Obelleiro Basteiro (UVIGO)
  - Luis Landesa Porras (UNEX)
- **Applications:**
  - Electromagnetic compatibilities studies (EMC), interferences (EMI), and risky radiation levels for radiating systems on board real platforms (cars/planes/ships).
  - Surface Equivalent Radar (SER) prediction for real targets. Analysis and design of practical antenna problems involving wire antennas, arrays, broadband antennas, etc
- **Computing requirements:**
  - 0.5-1 TB of memory
  - 500-1000 processors
  - 1-10 CPU days.



Challenges & lessons learned

- Parallel jobs allocation
  - Resource reservation & backfilling
  - On-the-fly re-adjust policies
  - Jobs checkpointing and migration (virtualization?)
- Filesystem
  - NFS vs. SFS
- Compatibility
  - (HP-SFS/HP-Serviceguard/HP-CMU/HP-SIM)
- Parallel jobs performance and bottleneck detection and analysis / Monitoring
- Infiniband on fat nodes (latency & BW/core)
- Memory is expensive
- Lots of hardware -> 53 hw failures, decreasing

carlosf@cesga.es

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

Reservas

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

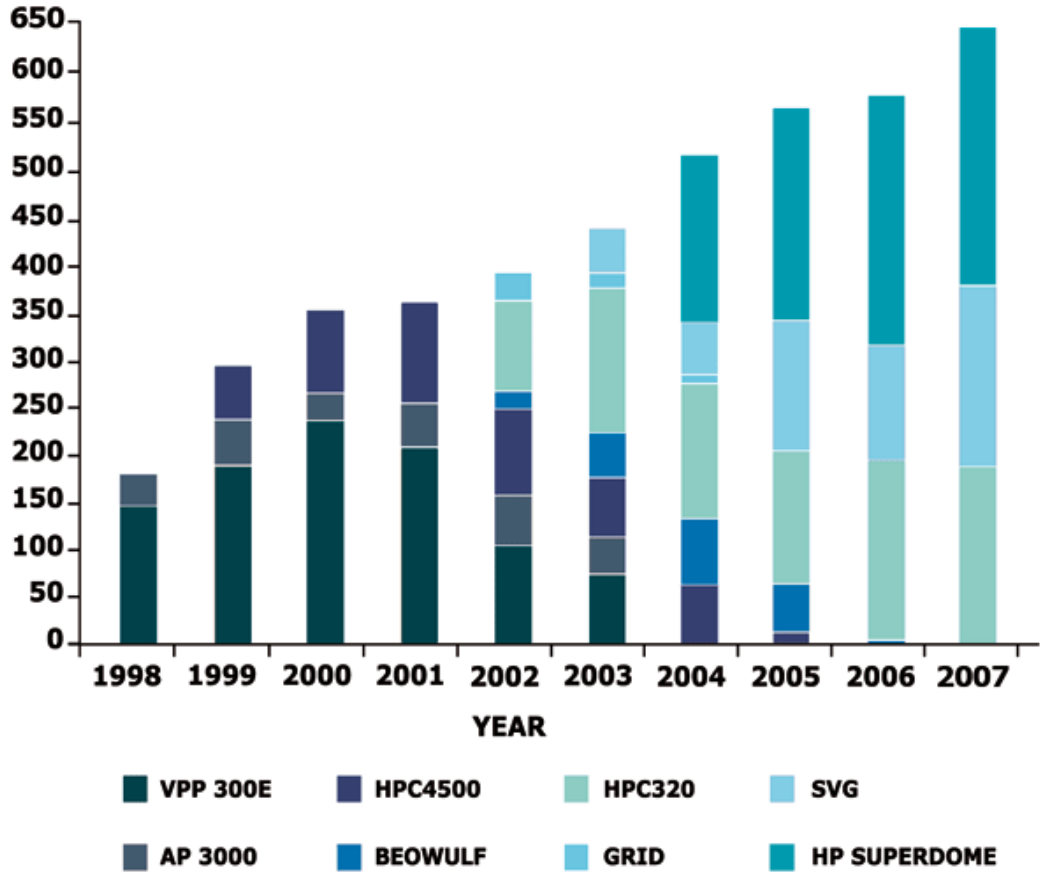
Job submission
Resources based

Slots	Number of MPI tasks
Num_proc	Minimum number of threads/node Total processors requested = num_proc*slots
S_rt	Maximum wallclock job time
S_vmem	Maximum memory per MPI task
H_fsize	Maximum scratch disk per MPI task SFS if slots > 1

qsub -l num_proc=nproc, s_rt=hh:mm:ss, s_vmem=memory, h_fsize=disksize pe mpislots job.sh

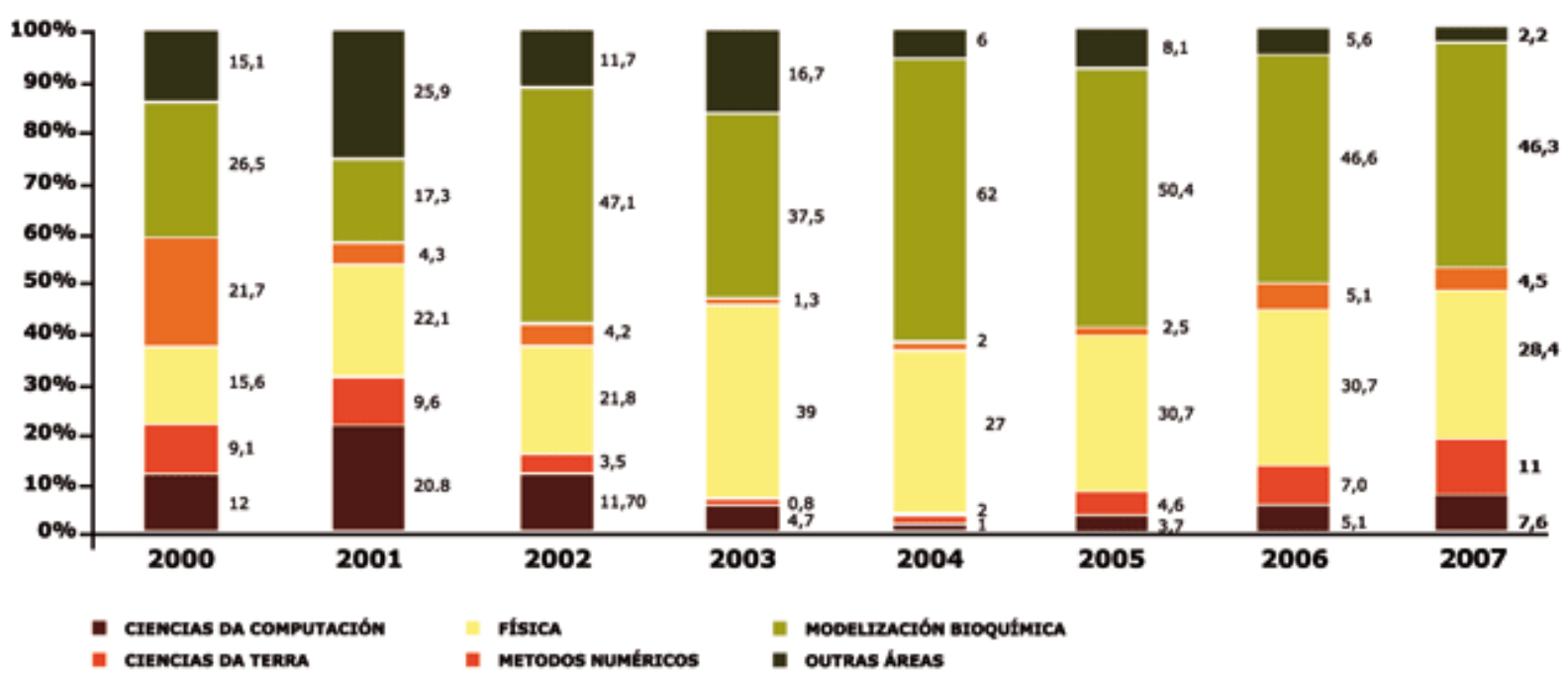
DEMAND OF COMPUTING RESOURCES AT CESGA

NUMBER OF USER ACCOUNTS PER SYSTEM PER YEAR



APPLICATION AREAS AT CESGA

CPU USE DISTRIBUTION PER AREA



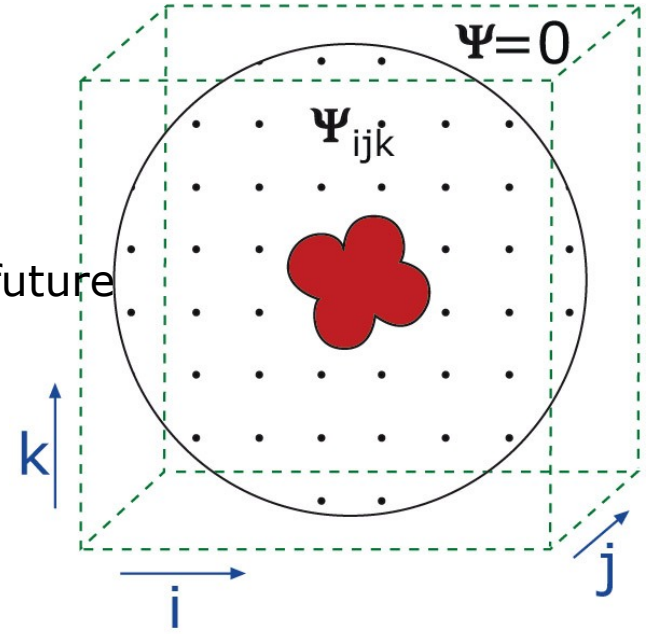
APPLICATION AREAS AT CESGA
SOME CURRENT PROJECTS

- **Project: Research in nanostructured Materials**

  - Dr. Manuel María González Alemany (USC)
  - Study of structural, electronic and optic properties of physical systems of technological interest by means of simulation techniques from first principles

- **Applications:**

  Predicting the structural properties and electronic of nanostructured material like nanowire, which have big technological implications. The nanowires could be the future materials of semiconductor industry



- **Computational requirements:**

  64 GB shared memory
  32 processors

APPLICATION AREAS AT CESGA
SOME CURRENT PROJECTS

- **Project: Study of the phase separation in magnetic oxides combining theory and experiment**

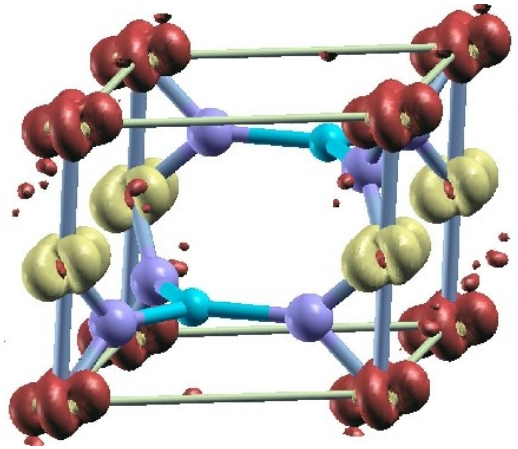
  - Dr. Daniel Baldomir Fernández (USC)
  - Simulations of magnetic materials (using DFT).

- **Applications:**

  Predicting electromagnetic properties at nanometric scale. Very useful in materials design to be used in: data storage, memories, drug administration systems, health monitoring, computer science, etc.

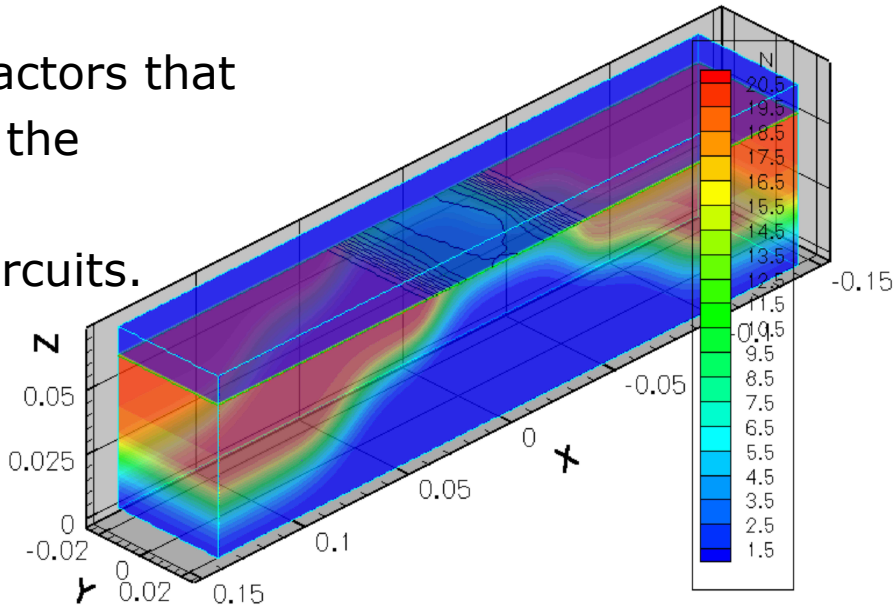
- **Computing requirements:**

  284 GB of memory
  128 processors

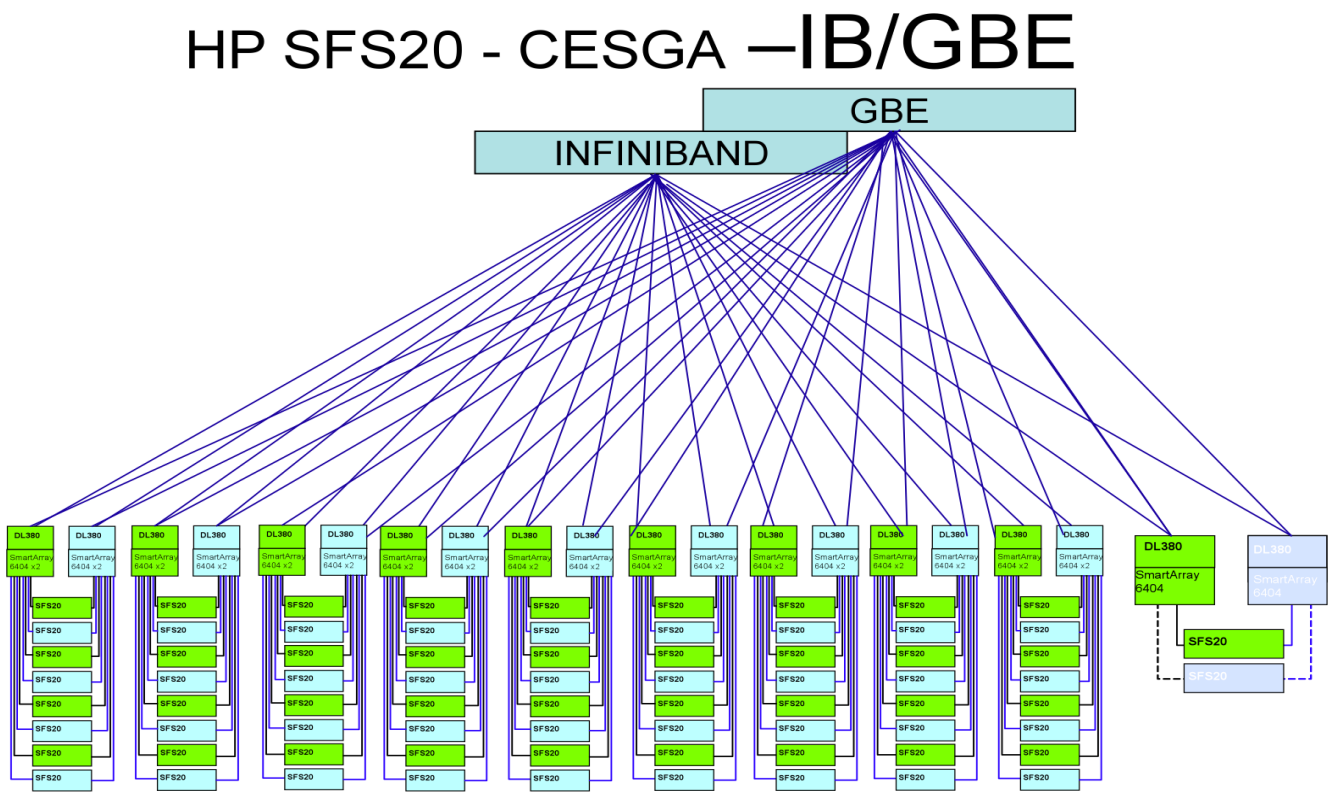


APPLICATION AREAS AT CESGA
SOME CURRENT PROJECTS

- **Project: Fluctuations in nanometric MOSFET devices**
  - Dr. Antonio Garcia Loureiro (U. Santiago) & A. Asenov (U. Glasgow)
  - Simulations of semiconductors including quantum effects (FEM and Monte Carlo)
- **Applications:**
  - Prediction of the dominant factors that degrade the performance of the transistors below 100nm.
    Development of electronic circuits.
- **Computing requirements:**
  80 GB of memory
  64 processors x 200 runs



HP-SFS



LEGAL ENTITIES

- Public Company
- Public Foundation

PARTNERS

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

Xunta de Galicia

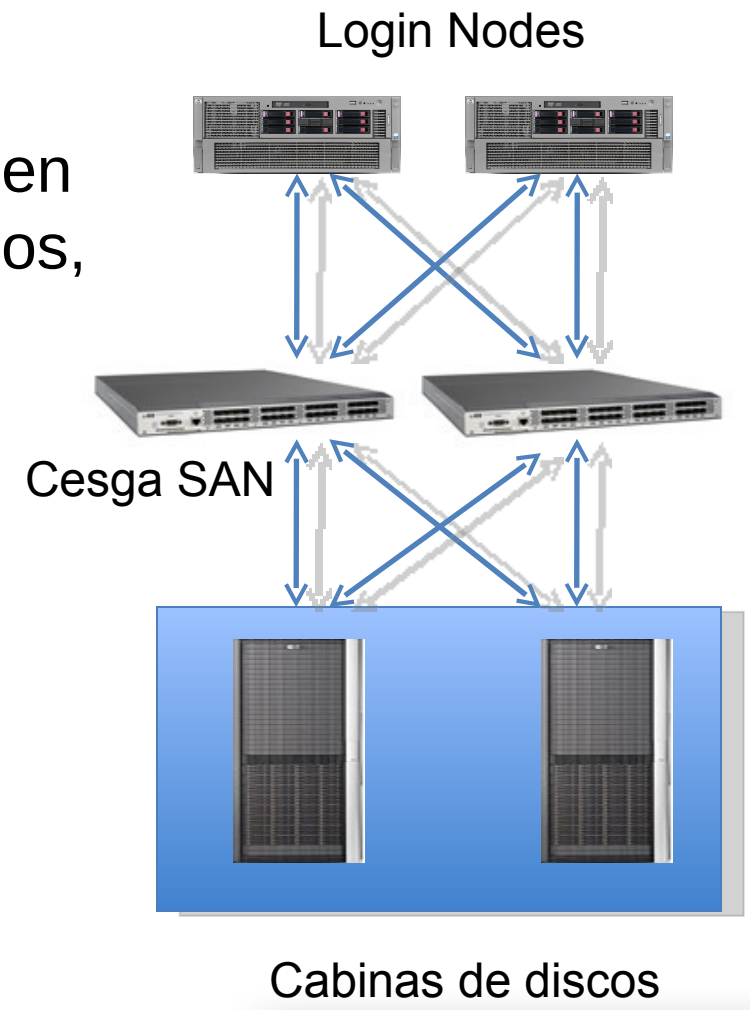
CESGA XUNTA DE GALICIA CSIC CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS ERDF European Regional Development Fund

CURRENT CESGA'S COMMUNITY OF USERS

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

Login/fileserver nodes

- Configuración redundada mediante servidores basados en Xeon (Proliant DL580, 4 núcleos, 8 GB), para:
  - login al cluster
    - Chequeo estado de los trabajos
    - Envío nuevos trabajos a cola
    - Borrar trabajos de la cola
    - Crear y editar ficheros
    - Realizar transferencias de ficheros a/desde Finis Terrae
    - Realizar algunas acciones de preprocesado y postprocesado
    - Acceder a los nodos interactivos (comando *compute*)
  - Sistemas de ficheros /home (nfs)



Job submission (GE) – Resources based

qsub -l num_proc=nproc, s_rt=hh:mm:ss,
s_vmem=memory, h_fsize=disksize pe mpislots job.sh

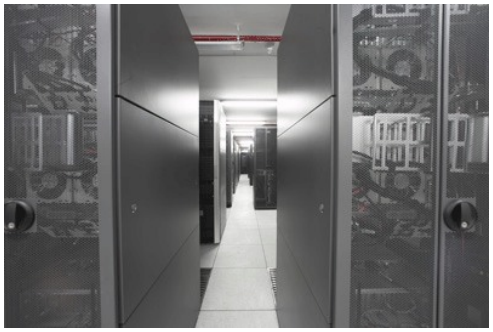
- *slots* # MPI tasks
- *num_proc* minimum # threads/node
  -
- *s_rt* maximum wallclock job time
- *s_vmem* maximum memory (per MPI task)
- *h_fsize* maximum scratch disk (per MPI task)

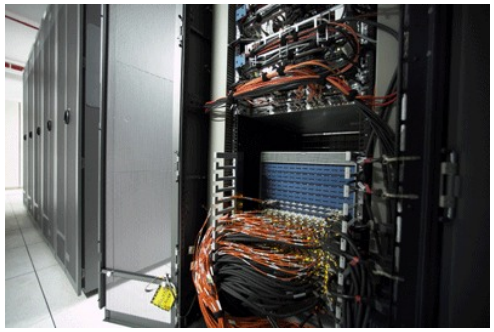
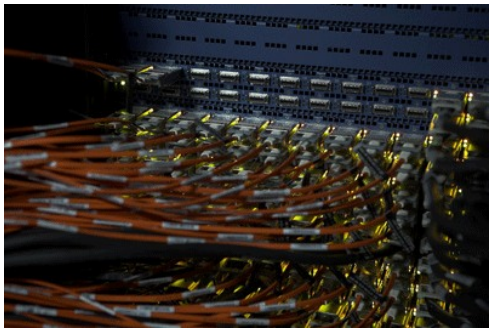
Total processors requested = num_proc*slots

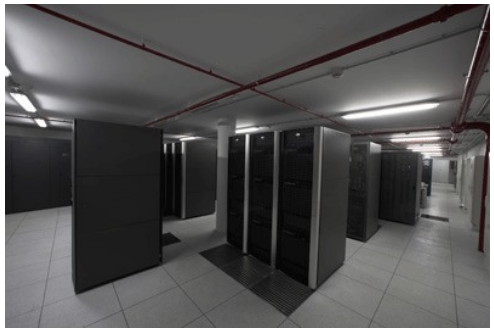
SERVICES

- HPC, HTC & GRID Computing
- User Data Storage
- Advanced Communications Network
- e-Learning & Collaboration Infrastructures
- e-Business Innovation Support
- GIS (Geographical Information Systems)

TECHNICAL INFRASTRUCTURE





Motivación

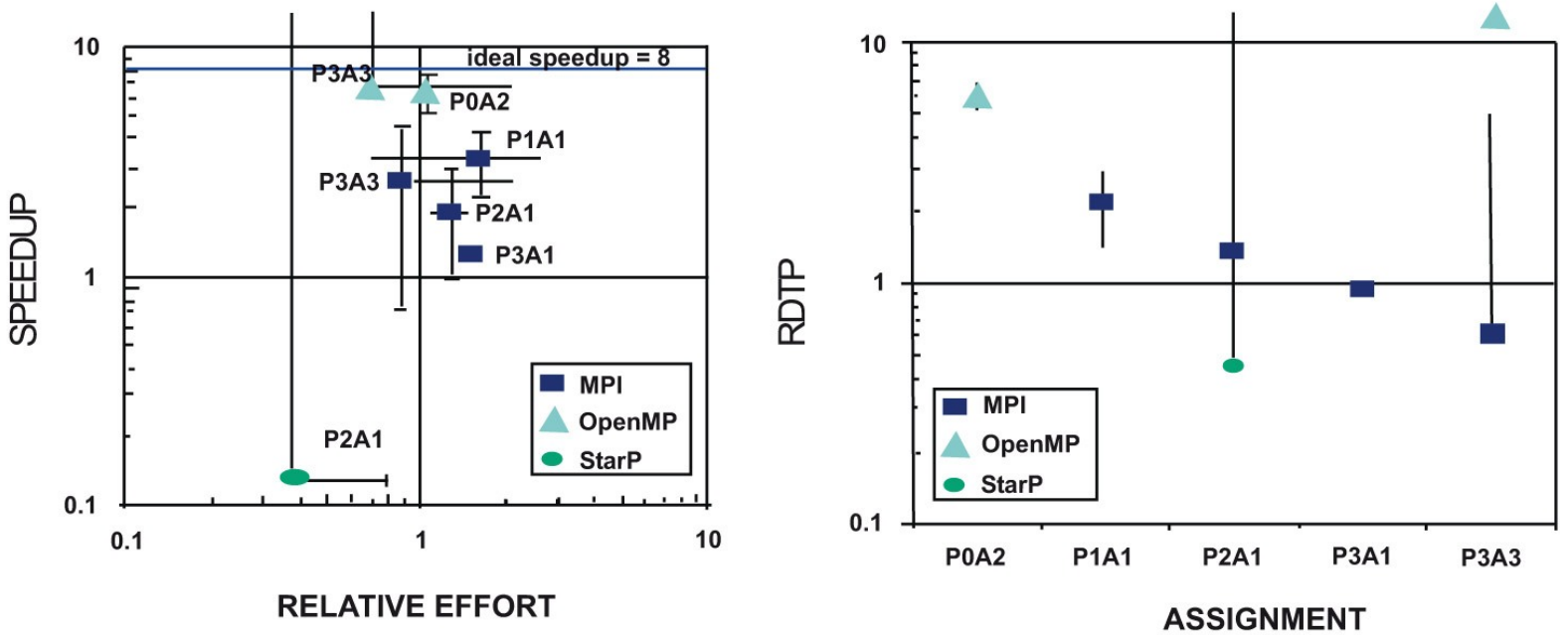
- Oportunidad!
- Estudio de la demanda de los usuarios y de su evolución en cuanto a :
  - Memoria requerida
  - Grado de paralelización
  - Tiempos de espera en cola
  - Tiempo total de resolución del problema

TECHNOLOGY
ANALYSIS OF PARALLEL SOFTWARE DEVELOPMENT

RDTP = speedup/relative effort

RDTP = Relative Development Time Productivity

Andrew Funk
MIT Lincoln Laboratory
Victor Basili
University of Maryland, College Park
Lorin Hochstein
University of Nebraska, Lincoln
Jeremy Kepner
MIT Lincoln Laboratory



Speedup vs Relative Effort and RDTP for the classroom experiments.

TECHNOLOGY
CESGA'S TECHNOLOGICAL EVOLUTION: INSTALLED SERVERS

1993
VP 2400

2.5 GFLOPS

1998
VPP 300

14.1 GFLOPS

AP 3000

12 GFLOPS

1999
HPC 4500

9.6 GFLOPS

2001
SVG

9.9 GFLOPS

2002
HPC 320

64 GFLOPS

BEOWULF

16 GFLOPS

2003
SUPERDOME

768 GFLOPS

2004, 2005, 2006
SVG

3,142 GFLOPS

2007
FINISTERRAE

16,000 GFLOPS

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

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

CESGA - C²SRC

CESGA COMPUTATIONAL SCIENCE RESEARCH CENTRE

CESGA

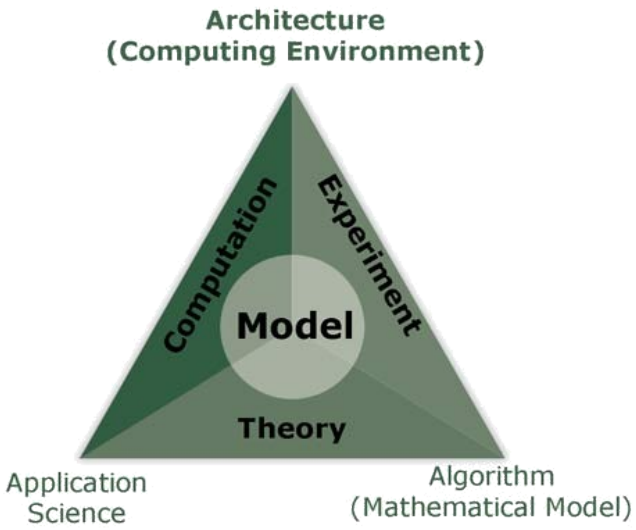
CSIC CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

XUNTA DE GALICIA

MINISTERIO DE EDUCACIÓN Y CIENCIA

ERDF European Regional Development Fund

COMPUTATIONAL SCIENCE



Computational Science is the field of study concerned with constructing mathematical models and numerical solution techniques and using computers to analyze and solve scientific, social scientific and engineering problems.
(Wikipedia)

C²SRC MISIÓN Y VISIÓN

- **MISIÓN**: realizar investigación de alta calidad en Ciencia Computacional, en estrecha colaboración con la comunidad investigadora gallega y estatal, contribuyendo de este modo a la evolución del conocimiento, la transferencia de tecnología y, como consecuencia, al bienestar social.

- **VISIÓN**: alcanzar y mantener el reconocimiento internacional como Centro de Excelencia en Investigación, aportando avances significativos en Ciencia Computacional.

DEPARTAMENTOS

- **Aplicaciones**
  - Ciencia Computacional y Ciencias de la Vida y la Salud
  - Ciencia Computacional y Ciencias del Mar
  - Ciencia Computacional y Nanotecnología
  - Ciencia Computacional y Energías Renovables
- **Computación de Altas Prestaciones**

PARTICIPACIÓN DE LAS UNIVERSIDADES GALLEGAS Y EL CSIC

- Incorporación de 145 Investigadores hasta el 2012
- Participación de investigadores de las Universidades a través de:
  - Programa de intensificación de la investigación de la Xunta, tipo I3.
  - Plazas de intensificación financiadas por el propio Centro y reguladas con cada universidad según convenio.
  - Profesores adscritos a la Universidad y al centro concurrentemente (tipo IMDEA).
  - Proyectos coordinados Centro – Universidad
- Participación de investigadores del CSIC, mediante:
  - Incorporación de investigadores de plantilla
  - Dotar con plazas de nueva creación.
  - Becas de formación predoctoral y contratos a doctores de su convocatoria JAE.
  - Titulados Superiores y Titulados de Grado Medio así como técnicos de la convocatoria JAE para el Área de Servicios, Innovación y Gestión.

CESGA

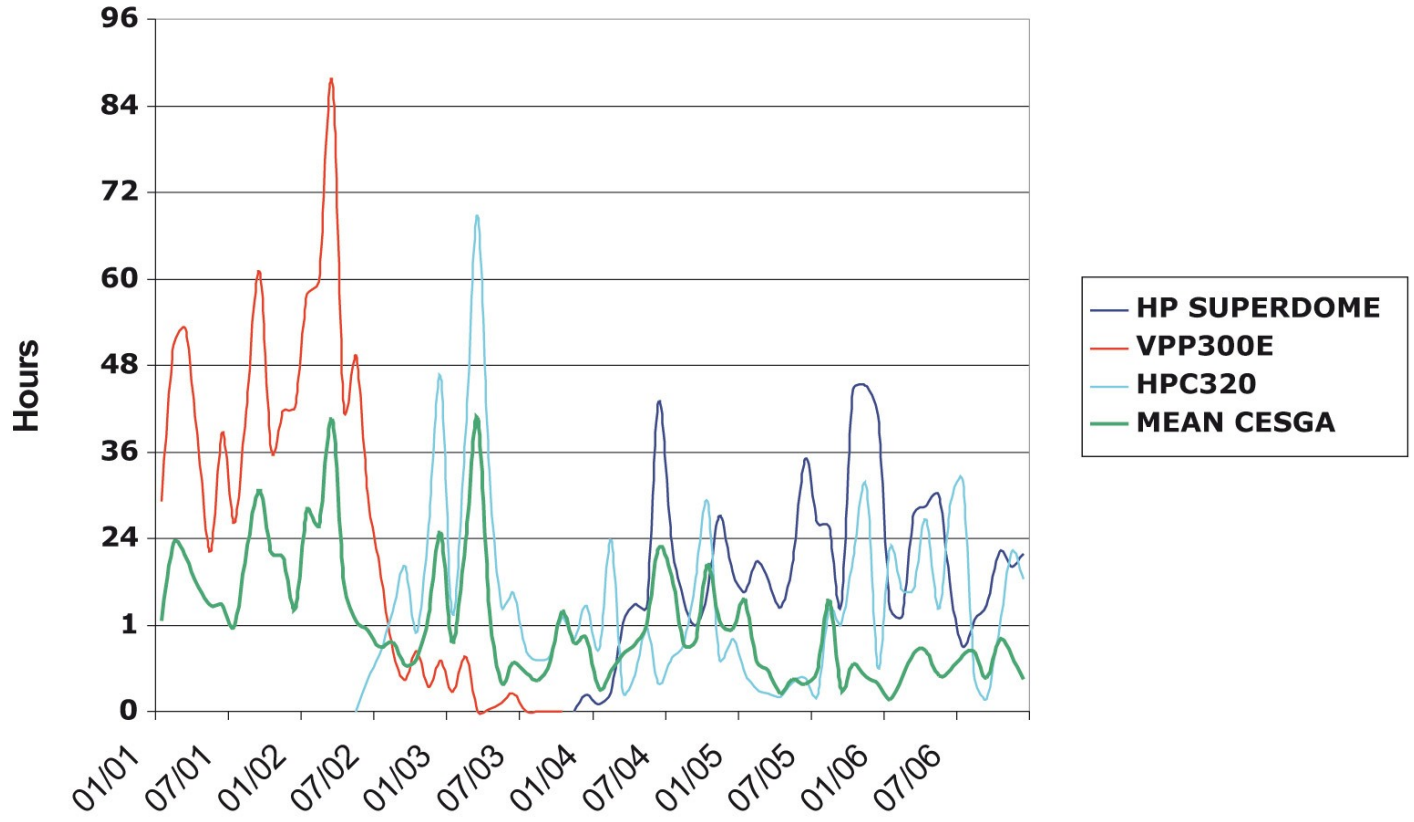
XUNTA DE GALICIA

CSIC CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

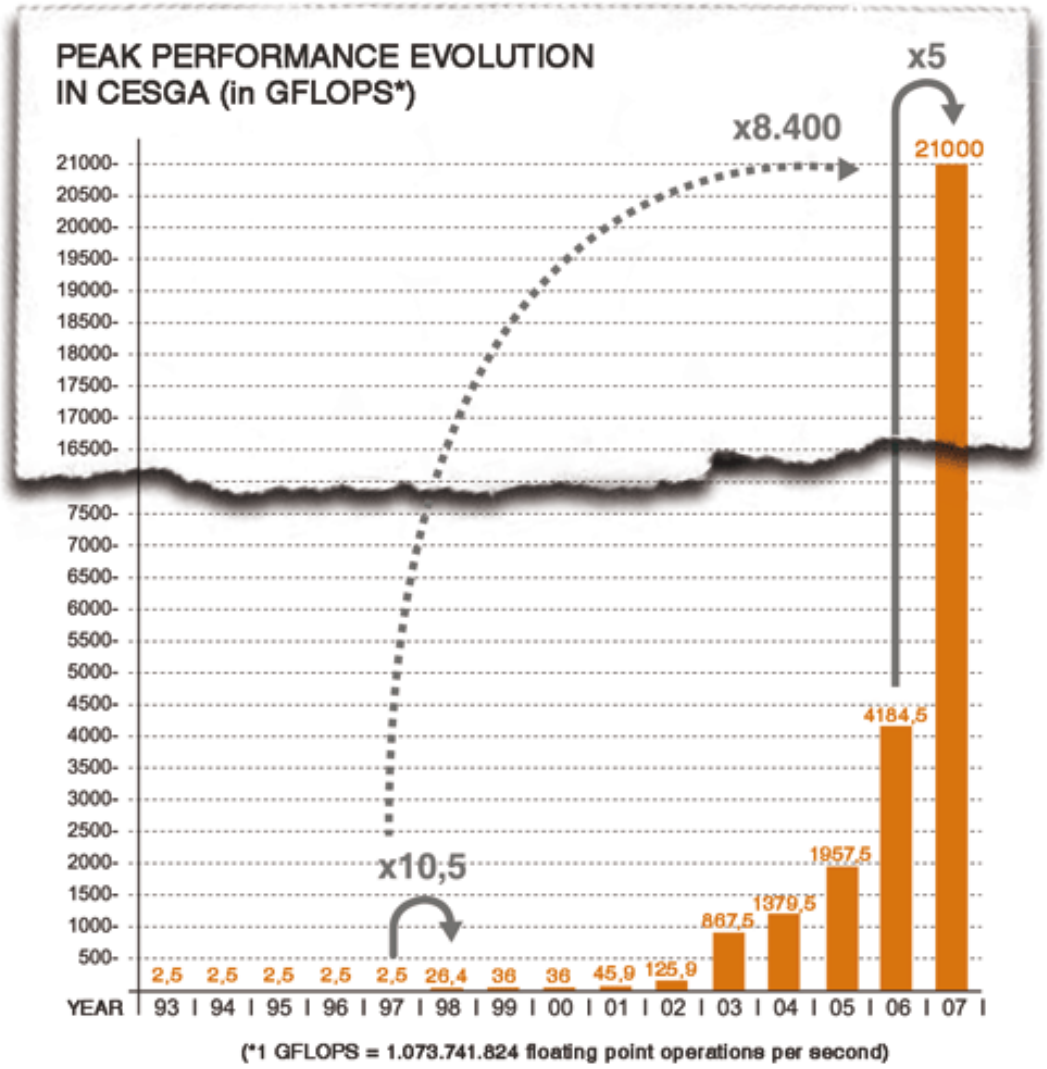
ERDF European Regional Development Fund

ANALYSIS OF THE DEMANDS FOR COMPUTING RESOURCES AVAILABLE AT CESGA

JOB WAITING PERIOD EVOLUTION (01/2001 – 12/2006) IN CAPABILITY SERVERS AND COMPARISON WITH THE MEAN FOR ALL CESGA'S SERVERS



CESGA's PEAK PERFORMANCE EVOLUTION



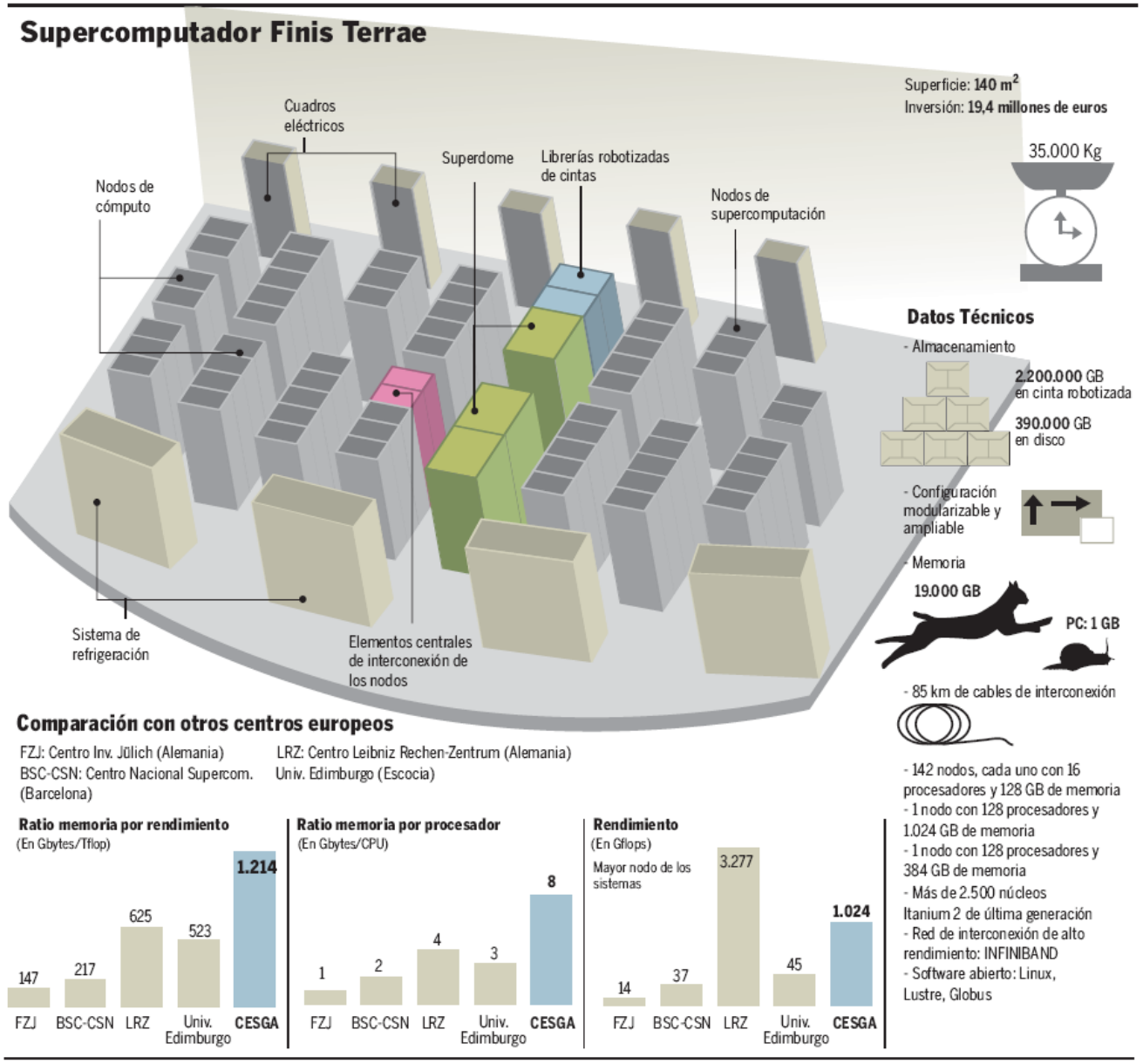
TECHNICAL INFRASTRUCTURE

Los grandes números

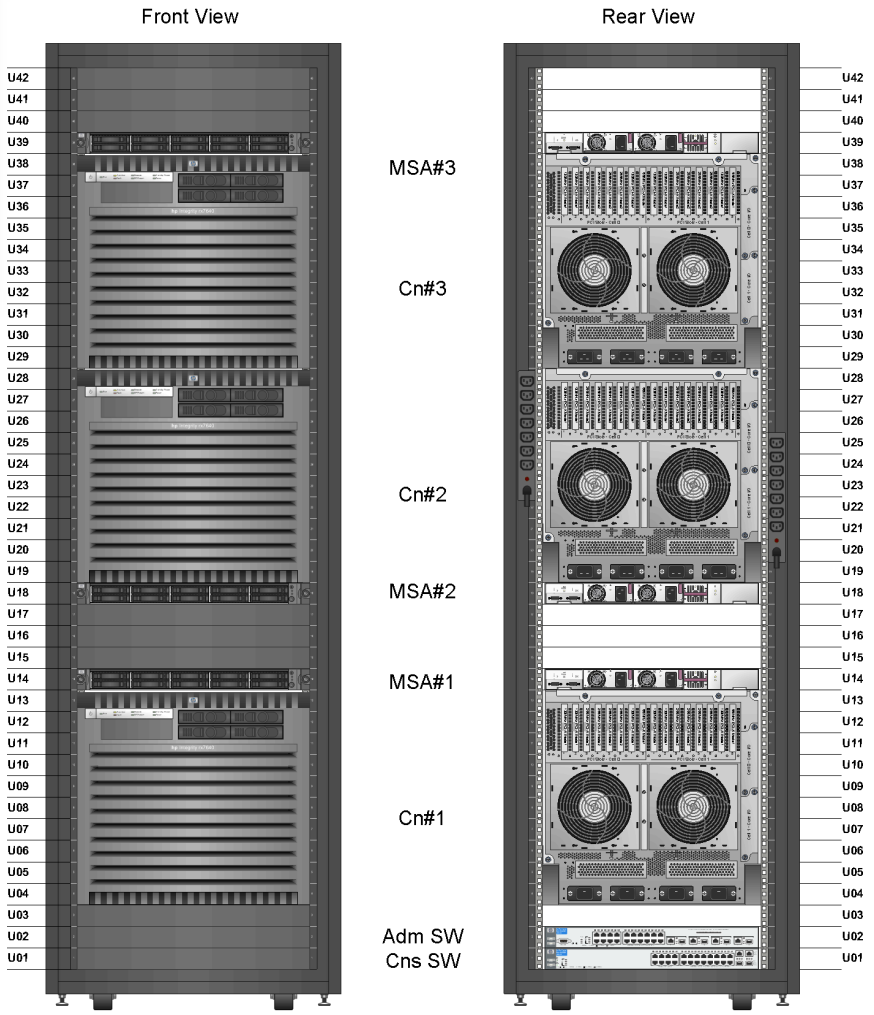
- 2400 núcleos de procesador Itanium 2 Montvale a 1.6 GHz, 18 MB caché/procesador
- 19 TB de memoria en arquitectura compartida
- Interconexión Infiniband
- SUSE Linux
- 16 TFLOPS pico (≈ 90% real)
- TOP100 (Nov07 list)

El cluster Finisterrae

Fuente: El Correo Gallego

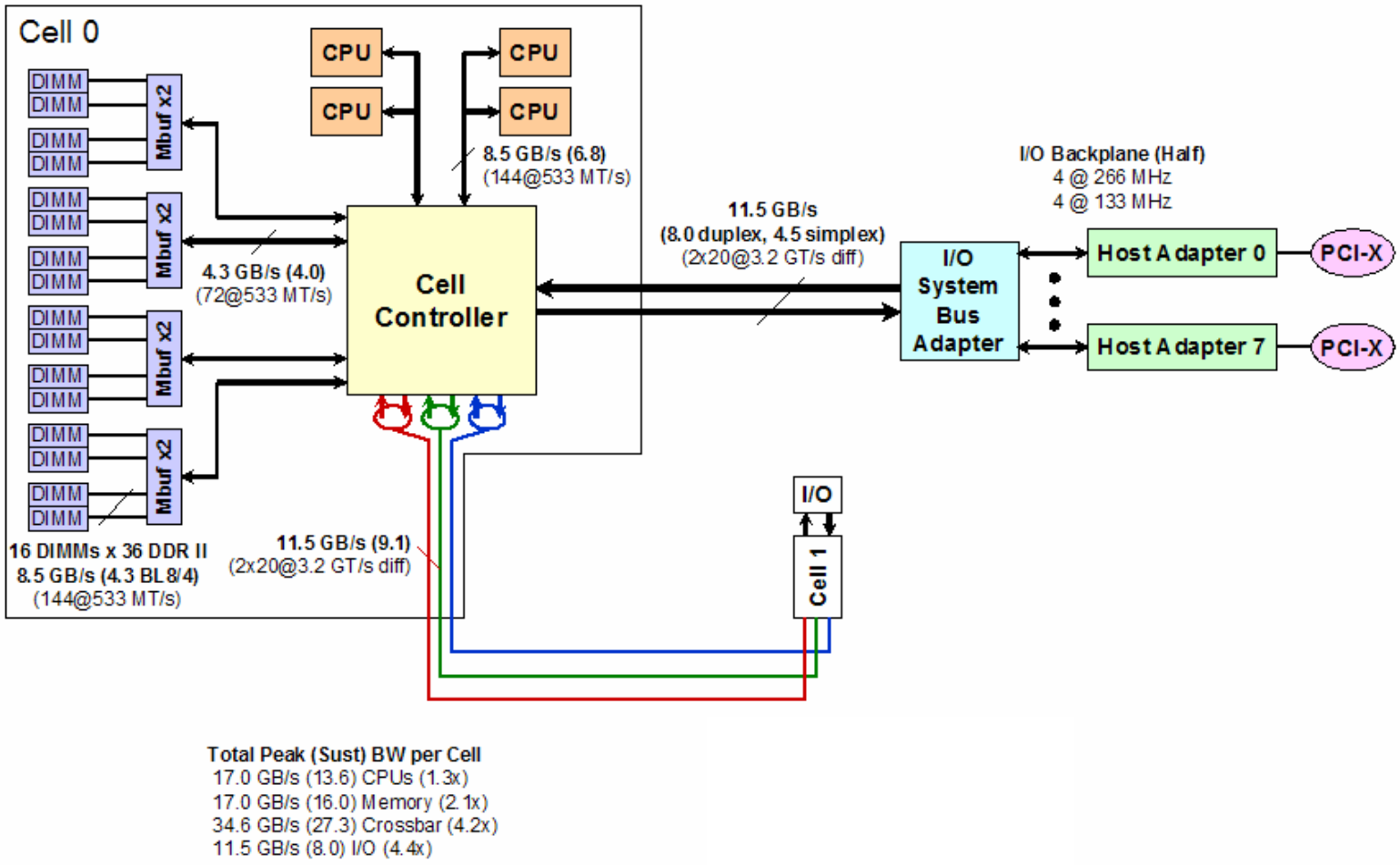


Racks de computación: rx 7640



- 48 racks
- Cada rack 3 servidores HP Rx 7640 y 3 cabinas de discos (6 discos de 146 GB por servidor)
- 142 servidores en total, cada uno con:
  - 16 núcleos Itanium 2 Montvale @ 1.6 GHz , 18 MB de caché
  - 128 GB de memoria
  - 6 discos SAS de 146 GB
  - Suse Linux

HP RX7640: Arquitectura

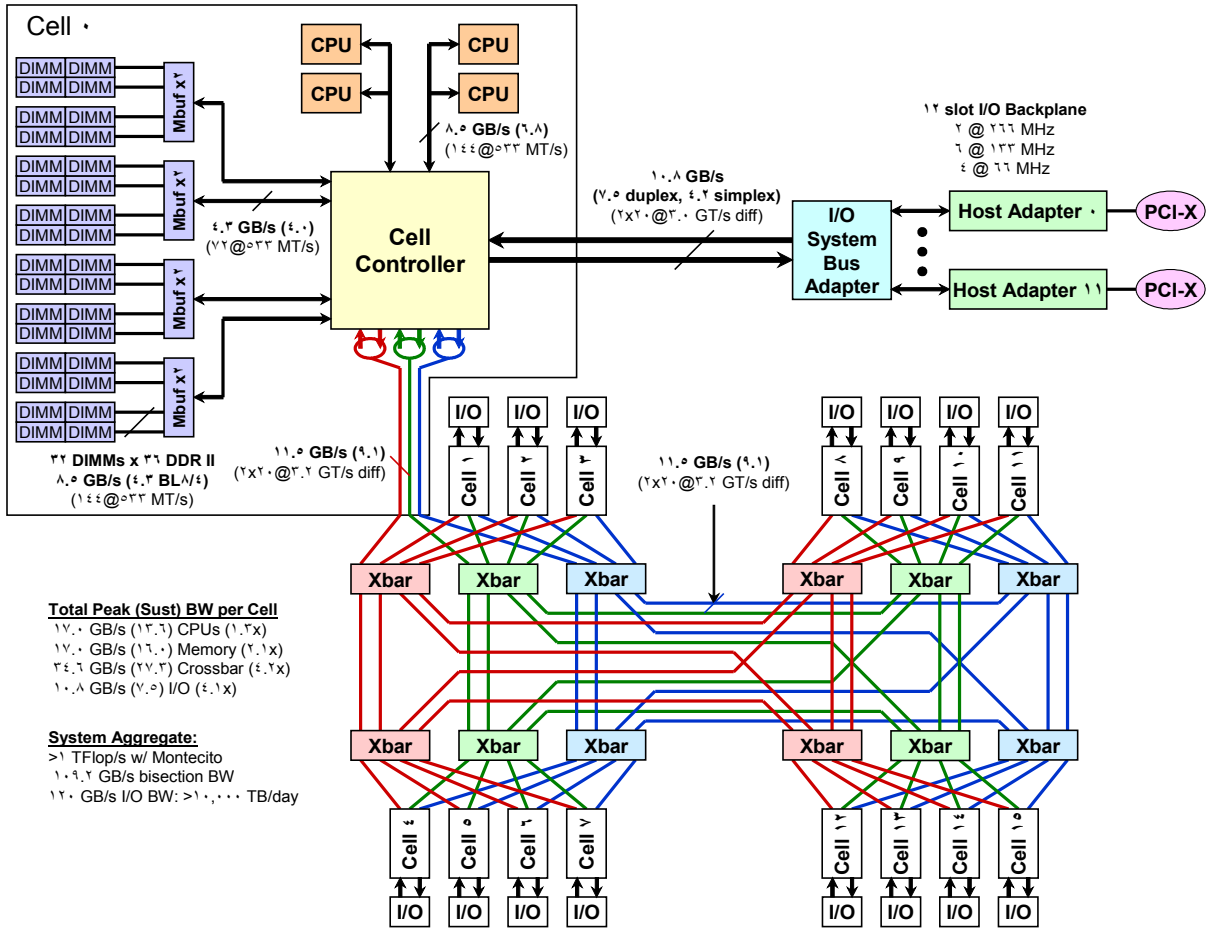


Total Peak (Sust) BW per Cell
17.0 GB/s (13.6) CPUs (1.3x)
17.0 GB/s (16.0) Memory (2.1x)
34.6 GB/s (27.3) Crossbar (4.2x)
11.5 GB/s (8.0) I/O (4.4x)

Nuevo Superdome

- 64 Procesadores/128 núcleos Itanium 2 Montvale @ 1.6 GHz, 18 MB de caché
- 1 TB de memoria en imagen única
- 128 discos de 72 GB SAS (9.2 TB) para scratch
- Suse Linux SLES 10

Superdome: Arquitectura



Red Infiniband

- Estándar de interconexión de redes de baja latencia
- Switch Voltaire ISR 9288
- 4X DDR, 20 (16) Gbps non blocking por puerto.
- Baja Latencia MPI (≈7μs real)
- Todos los elementos (nodos de computación/ SFS /Login) se conectan a un único switch

HP-SFS

- Sistema de almacenamiento paralelo basado en Lustre
- Capacidad total de 216 Terabytes, obtenida mediante 864 discos SATA de 250 GB
- 18 celdas (servidores Proliand DL380) y 72 cabinas de discos HP SFS-20.
- Proporciona hasta 10 GB/s en lectura y 6 GB/s en escritura (real)
- Se accede mediante Infiniband.
- Es visto desde todos los nodos de computación como un sistema de ficheros convencional (/sfs)

Librería de cintas

- 2 librerias robotizadas HP ESL 712e con pass-through
- 12 drives LTO4 con capacidad de 800 GB sin compresión
- En total 1.424 slots para cintas (1.140 TB sin compresión, > 2 PetaBytes con compresión)
- Velocidad de transferencia de 1.9 GB/s
- Conexión a la SAN
- Backup HP Data Protector

¿Cómo usar el CESGA?

¿Por dónde empezar?

- Web: http://www.cesga.es
  - Registro de usuarios
  - Descripción y guía de uso de los servicios
  - Dudas más frecuentes
- Correo electrónico:
  - sistemas@cesga.es
  - aplicacions@cesga.es
- Teléfono: 981 569 810
  - Depto. Sistemas
  - Depto. Aplicaciones y Proyectos

Registro de usuarios CSIC

USUARIOS DE CENTROS DE INVESTIGACIÓN DEL CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS (CSIC)

- Para acceder a los recursos de Cálculo y Almacenamiento instalados en el CESGA es necesario:
  - Que el Grupo de Investigación al que usted pertenece se encuentre registrado en el CESGA. De no ser así, necesita rellenar el registro de grupo (Descargar formulario). Esta solicitud la realizará el Director del Grupo de Investigación una única vez.
  - Que usted solicite ser registrado como Usuario (Registro de Usuario).
  - Que usted solicite los servicios a los que desea acceder. Por favor, adjunte una solicitud por cada registro de Usuario (Petición de Servicios).

- Por último estos formularios tienen que ser enviados por correo ordinario a las siguientes direcciones
  - CTI
  - CESGA

¿Cómo se utilizan? Conexión

- Nociones básicas sobre UNIX
- Conexión mediante cliente SSH
  - Unix, Linux, etc& : OpenSSH
  - Windows: Putty, Cygwin
- Posibilidad de utilizar ventanas
  - Unix: ssh X &
  - Windows: Utilizar un cliente X (X-Win32,&) o Cygwin
- Transferencias de ficheros
  - Unix: scp o sftp
  - Windows: Winscp

Modos de trabajo

- Modo Interactivo
  - Recursos limitados y compartidos
  - Compilaciones, edición de ficheros, pruebas
- Modo Batch (basado en colas de ejecución)
  - Acceso a los recursos de modo exclusivo y reservado
  - Estimar recursos necesarios (máximo):
    - Número de procesadores
    - Tempo de ejecución
    - Memoria
    - Espacio en disco (scratch o temporal)
  - Tiempos de espera&

FT: Límites iniciales

- Máximo de procesadores: 160
- Tiempo de ejecución (s_rt):
  - Trabajos secuenciales hasta 1000 horas.
  - Trabajos paralelos 2 a 8 procesadores: 100 horas.
  - De 9 a 16 procesadores: 48 horas.
  - De 17 a 64 procesadores: 24 horas.
  - De 65 a 128 procesadores: 10 horas.
  - De 129 a 160 procesadores: 8 horas.
- Memoria: 112 GB (por nodo)
- H_fsize: 500GB (scratch local)

¿ … y si necesito más?

- Solicitar acceso a recursos especiales
- Procedimiento
  - descrito en www.cesga.es (computacion -> Rec. Especiales)
  - dirigiéndose a sistemas@cesga.es
- En proceso de estudio de mecanismo vía comité de acceso (Requerimiento ICTS para 20% MEC)

Servicio de Almacenamiento

- Servicio de almacenamiento masivo de datos
- Almacenamiento en cinta para archivado de información
- Diferentes recursos para cada tipo de información
- Clasificación de los tipos de información
  - Tipo 1: Almacenamiento temporal o scratch (solo durante la ejecución de un cálculo)
  - Tipo 2: Incrementar el almacenamiento en el directorio home y el nº de ficheros
  - Tipo 3: Almacenamiento masivo de datos (bases de datos, repositorios, etc...)
  - Tipo 4: Copias de seguridad a disco
  - Otro tipo (adaptado a las necesidades descritas por el investigador)
- Formulario de almacenamiento (http://www.cesga.es/ga/Almacenamento)

¡ Muchas Gracias !

Carlos Fernández Sánchez
(sistemas@cesga.es)