CESGA – Finis Terrae Computational Science Conference
Santiago de Compostela, 2008

Finis Terrae User Guide v1.5

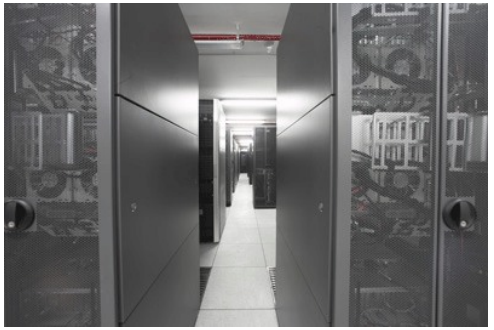
Driving science with productivity computing

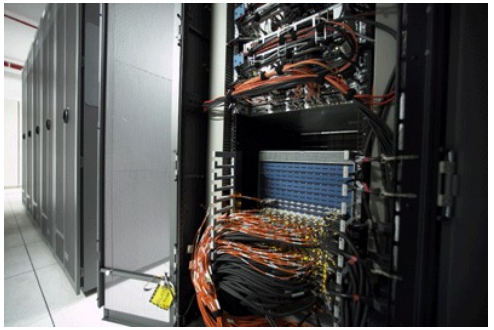
Carlos Fernández Sánchez
(HPC & Grid Systems Manager, Galicia Supercomputing Centre)

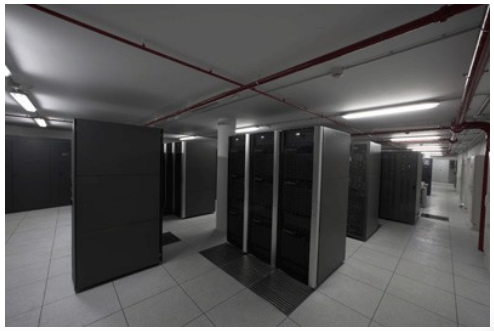
CESGA
XUNTA DE GALICIA
CSIC CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS
ERDF European Regional Development Fund

Agenda

- Supercomputing facilities
- Connecting
- Storage
- Batch system
- Special resources
- Support
- Enhacements roadmap







Supercomputing facilities

FT

Capability computing

Parallel jobs (>4 ... 1024 cores)

Huge memory (>4... 1024GB)

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

SVGD

Throughput and Capacity computing

Sequential jobs (up to 8 cores)

Low-medium memory (<8GB)

Low-single node scratch (<50GB)

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

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

Connecting

- ssh -X ft.cesga.es (putty windows client)
- x86 system (login node) functionalities:
  - send, receive files to the system
  - edit files
  - submit, check jobs
  - **Incompatible** with Finis Terrae compute nodes
- To compile and check small jobs:
  - compute mem memory (GB)

Storage

- home ($HOME):
  - use carefully
  - no top performance,
  - 10GB/user, smaller files
- SFS ($HOMESFS):
  - top performance, mandatory for MPI jobs (parallel scratch)
  - No backup. Backup on-demand
  - Storage of results partial results
  - Up to 1TB, 1000 files (big files, >1MB)
- Scratch ($TMDIR):
  - job-lifetime, single slot, up to 500GB
- Compartido & SVG ($COMPARTIDO, $HOMESVG):
  - share data with SVG

Batch system

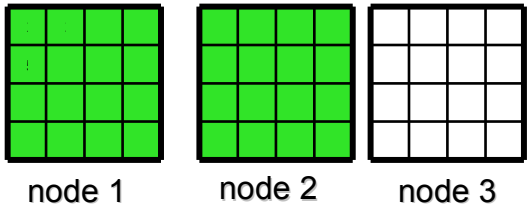
- How to submit:
  - qsub l num_proc,s_rt,s_vmem,h_fsize pe mpi job.sh
- Limits (possibility of special resources):
  - Processors: 160
  - s_vmem: 112GB/slot
  - h_fsize: 500GB/slot
  - s_rt: depends on the number of processors
- Checking jobs (qstat & web)
- Deleteting jobs (qdel)
- Jobs analysis (post-morten): qacct JOBID
- Estimating resources (qacct & time)

Parallel jobs allocation

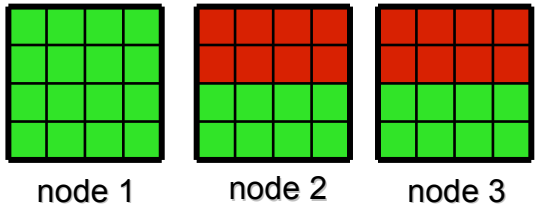
- OpenMP: single node (num_proc)
- MPI: single node or multiple node (slots)
  - num_proc=1, PE mpi slots
  - num_proc=1, PE mpi_rr slots
- OpenMP+MPI:
  - num_proc & PE mpi
- Special allocation: ask for it
- Total number of cores=num_proc x slots

Parallel jobs allocation Examples

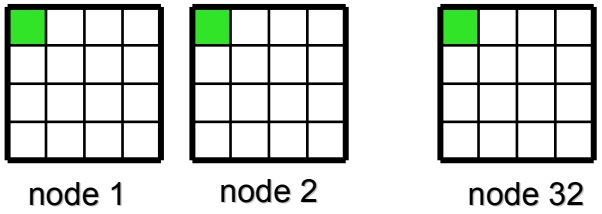
- num_proc=1, PE mpi 32



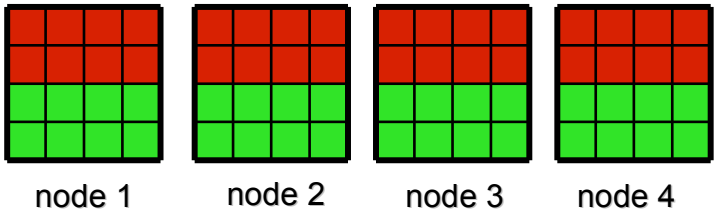
many options



- num_proc=1, PE mpi_rr 32

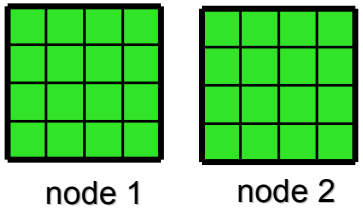


many options



- num_proc=16, PE mpi 2 (num_proc=16, PE mpi_rr 2)

no options



Batch Priorities Assignation

- System for parallel, memory-demanding jobs
- Split in 3 parts:
  - small: up to 4 cores (10%-20% system)
  - medium: 4 up to 16 cores (40% system)
  - large: more than 16 cores (40% system)
- Priorities assignation:
  - Based on waiting time and past use

Special resources

- FLEXIBILITY!!!!
- Aim to solve problems
- Ask for special resources if:
  - Limits don´t let you solve your problem
  - Urgent jobs
- Form on the web, fill & submit
- Technical committe approval (allow 1-2 weeks)

Support

- Send mail: sistemas@cesga.es
  - Prefered method and ticket-based
- Phone: +34-981-569814
- Avoid personal mailings, if possible
- Biggest improvement in Finis Terrae:
  - 4 System administrators

Roadmap

- Upgrades:
  - SFS2.3, OFED1.3, legacy Superdomes integration, all fiber
- Checking nodes/resources before job runs
- Selection of processor/nodes allocation (benchmarking)
- Better isolation of jobs (cpusets)
- Online monitoring job performance
- Intelligent qsub (application aware)
- Digital Certificates for communications
- SFS on SVG compute nodes
- SVG-FT shared frontend
- User based job priorization (user decides)

sistemas@cesga.es

Cesga-Finisterrae
Computational Science Conference

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

Reservas

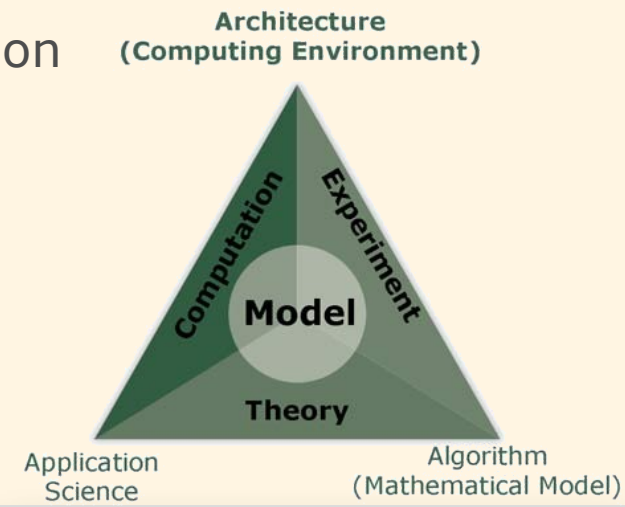
PAST, PRESENT, FUTURE (in 2 minutes)

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

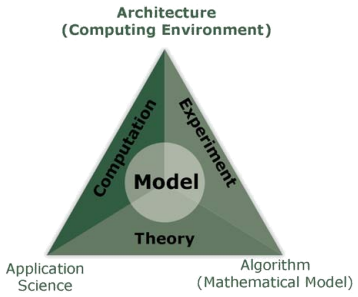
Services: High performance computing, storage and communication resources (RedIris PoP)

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

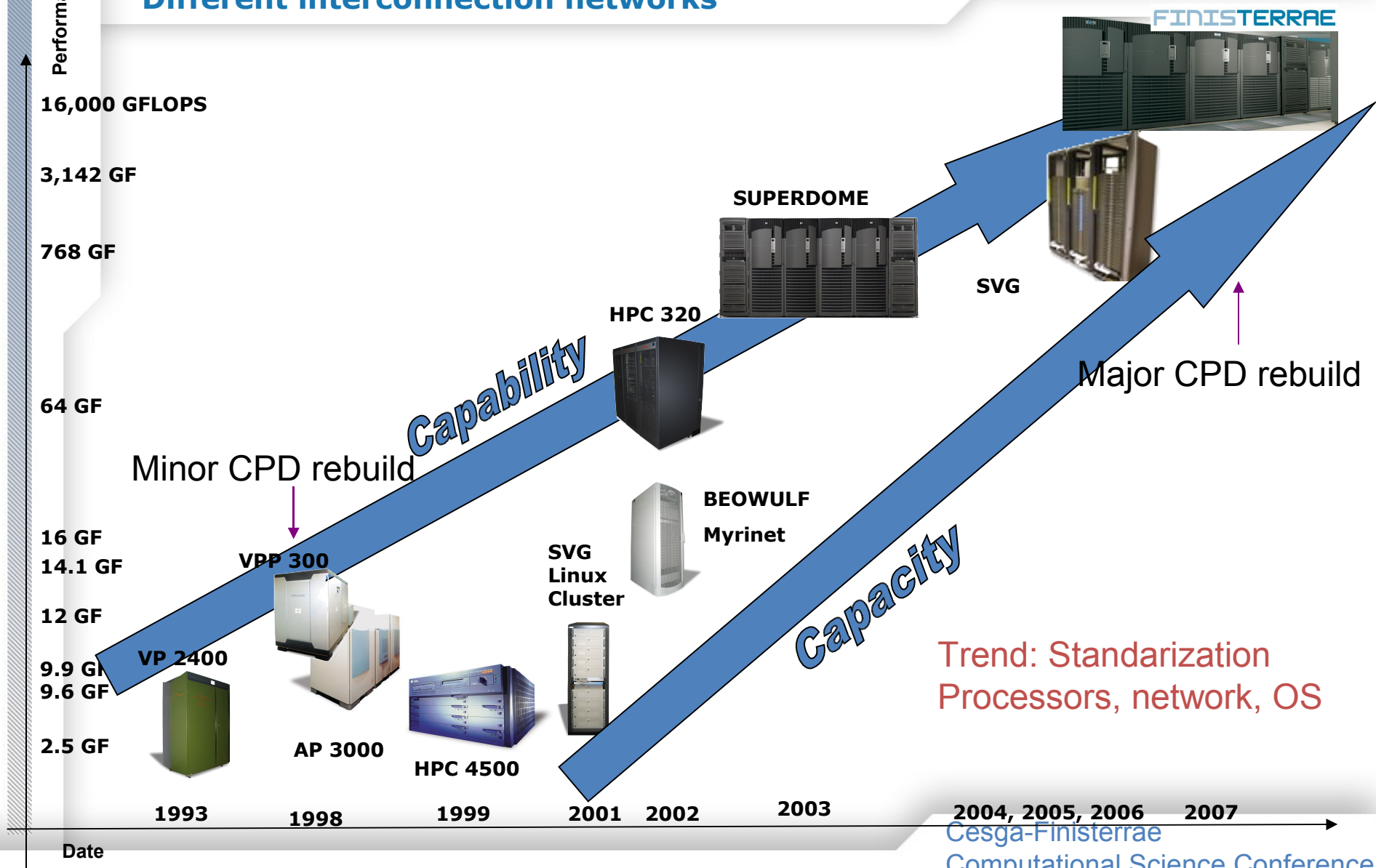
R&D Galician Plan 2006-2010

- 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

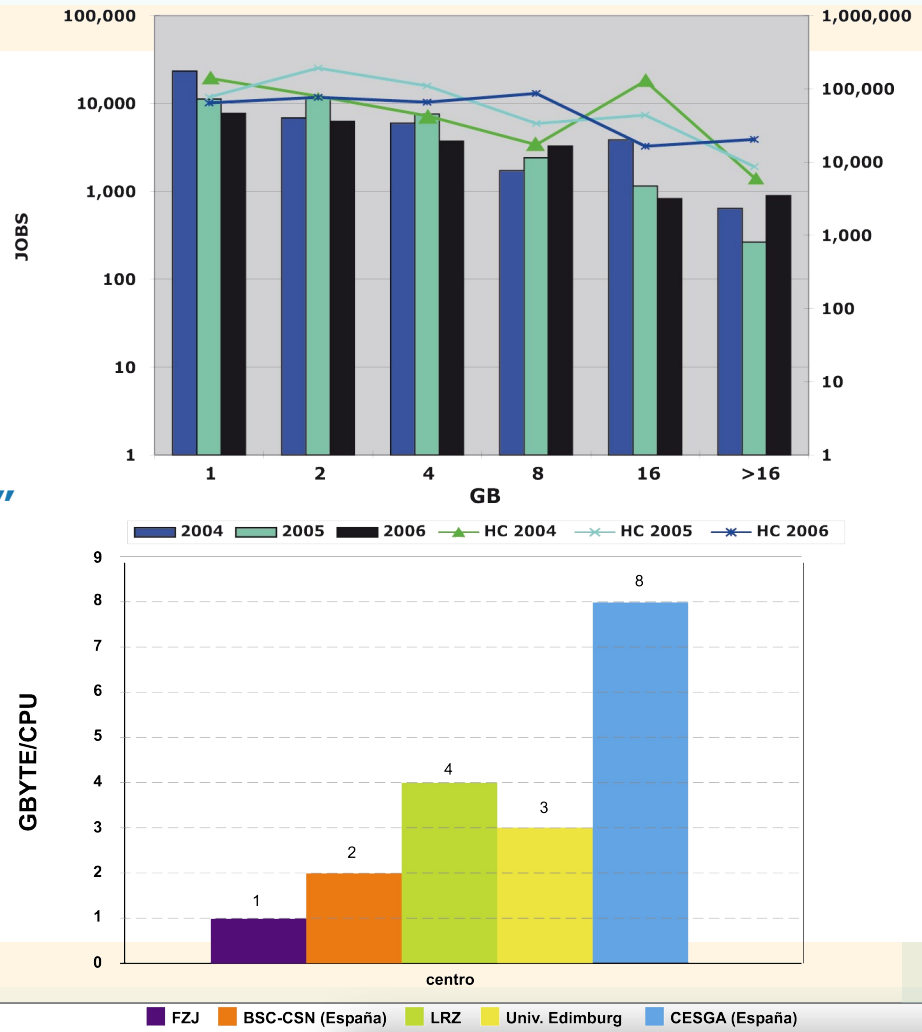
COMPUTING CHALLENGES
& "EVERYDAY" RESEARCH

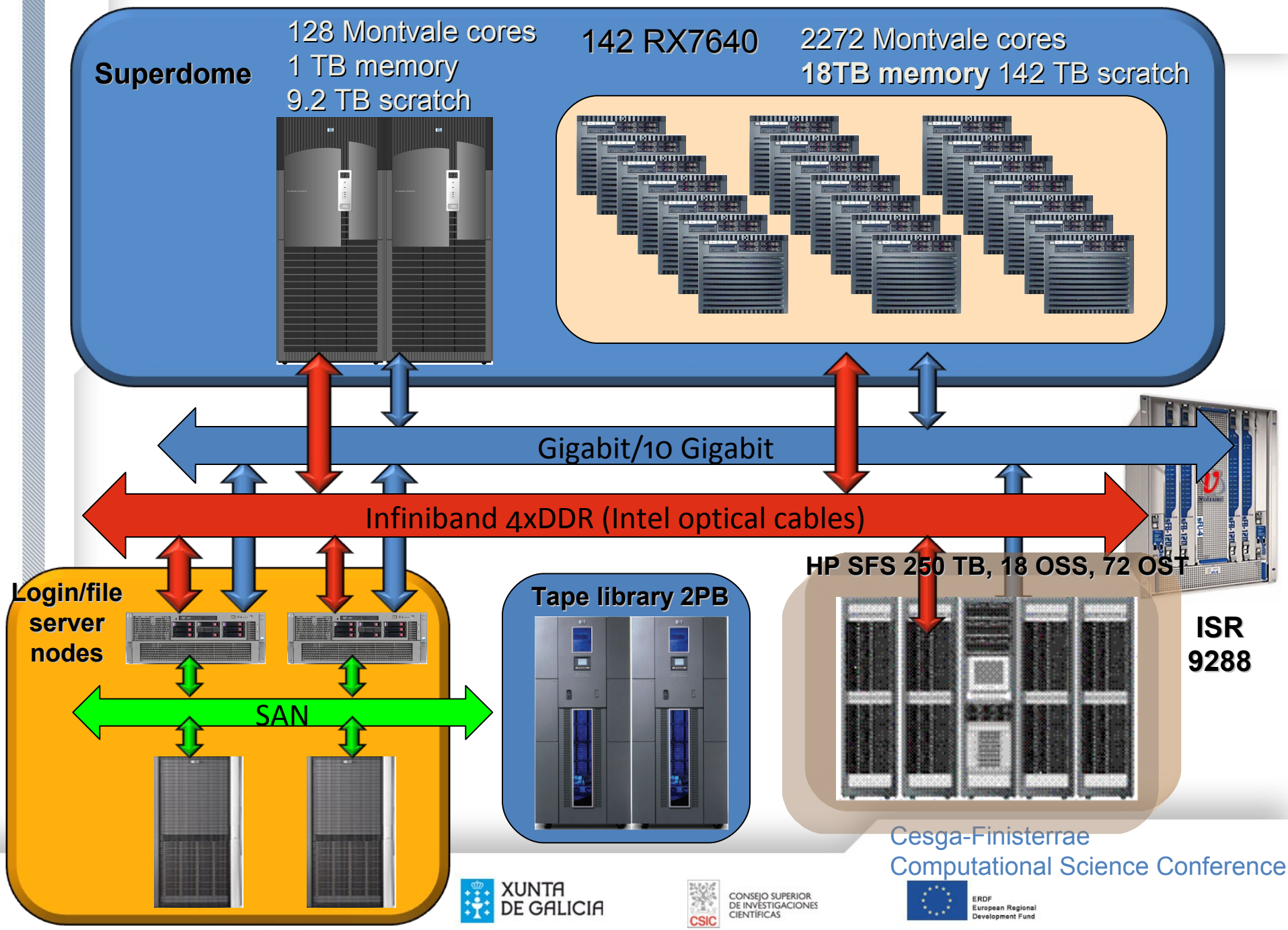
"CAPABILITY" AND "CAPACITY"

MINIMIZES "TIME TO SOLUTION"
(SIMPLER PROGRAMMING)

TAKES ADVANTAGE OF OLD, THOUGH EFFICIENT CODES

Highest GB/core in Europe
No Top500/Green500 focus





Cesga-Finisterrae
Computational Science Conference

XUNTA DE GALICIA

CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

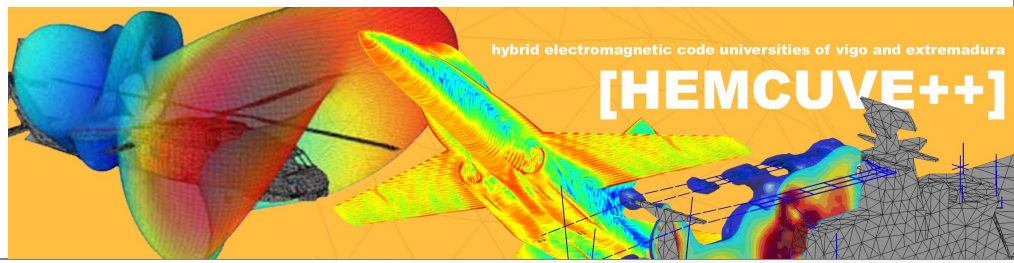
ERDF European Regional Development Fund

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