MareNostrum
Building and running the system

Lisbon, August 29th, 2005

Sergi Girona
Operations Head

History: Three Rivers Project

- IBM project
- Objective
  - Bring IBM Power Systems back into the Top5 list
  - Push forward Linux on Power
  - Scale-out
- Strategy
  - Find a **willing** partner to deploy bleeding edge technologies in an **open** collaborative environment
    - Research university preferred.
  - Integrate a complete supercluster architecture
    - optimized for **cost/performance**
    - using **latest available technologies** for interconnect, storage, and software
- Goals
  - Get system into Top500 list by SC2004 in Pittsburgh PA, hence the name.
  - Complete installation in 11/04 and system acceptance in 1H05

History: UPC

- CEPBA (1991 – 2004)
  - "Research and service center" within the Technical University of Catalonia (UPC)
  - Active in the European projects context
  - Research
  - Computer architecture
  - Basic HPC system software and tools
  - Data bases
- CIRI (2000 – 2004)
  - R&D partnership agreement between UPC and IBM
  - Research cooperation between CEPBA and IBM

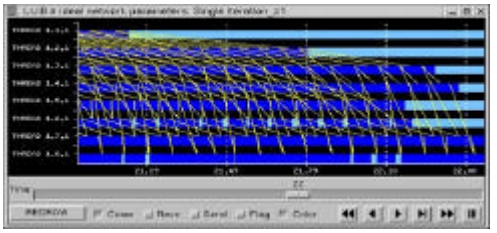
Index

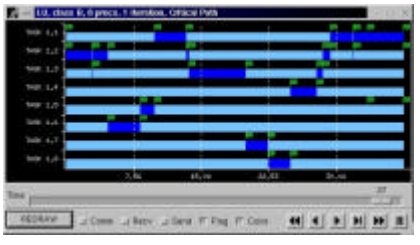
Barcelona Supercomputing Center

- Mission
  - Investigate, develop and manage technology to facilitate the advancement of science.
- Objectives
  - Operate national supercomputing facility
  - R&D in Supercomputing and Computer Architecture.
  - Collaborate in R&D e-Science
- Consortium
  - the Spanish Government (MEC)
  - the Catalonian Government (DURSI)
  - the Technical University of Catalonia (UPC)

IT research and development projects – Deep Computing

- Continuation of CEPBA (European Center for Parallelism in Barcelona) research lines in Deep Computing:
  - Tools for performance analysis.
  - Programming models.
  - Operating Systems.
  - Grid Computing and Clusters.
  - Complex Systems & e-Business.
  - Parallelization of applications.



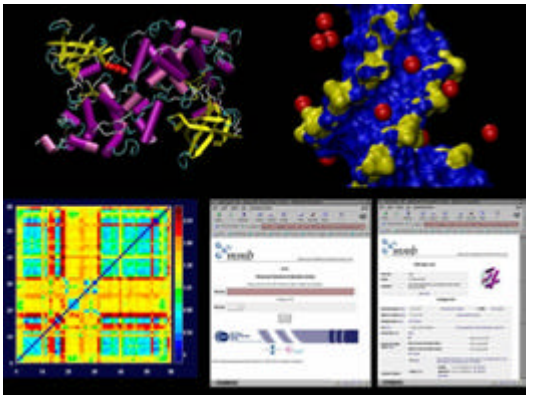


IT research and development projects – Computer Architecture

- Superscalar and VLIW processor scalability to exploit higher instruction level parallelism.
- Microarchitecture techniques to reduce power and energy consumption.
- Vector co-processors to exploit data level parallelism, and application specific co-processors.
- Quality of Service in multithreaded environments to exploit thread level parallelism.
- Profiling and optimization techniques to optimize the performance of existing applications.

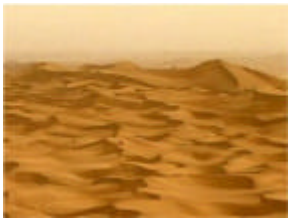
Life Science projects

- Genomic analysis.
- Data mining of biological databases.
- Systems biology.
- Prediction of protein fold.
- Study of molecular interactions and enzymatic mechanisms and drug design



Earth Science projects

- Forecasting of air quality and concentrations of gaseous photochemical pollutants (e.g. troposphere ozone) and particulate matter.
- Transport of Saharan dust (outbreaks) from North Africa toward the European continent and its contribution to PM levels.
- Modeling the climate change. This area of research is divided into:
  - Interaction of air quality and climate change issues (forcing of climate change).
  - Impact and consequences of climate change on a European scale



Services

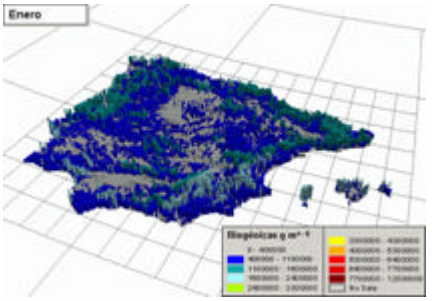
- Computational Services: Offering our parallel machines computational power.
- Training: Organizing technical seminars, conferences and focused courses.
- Technology Transfer: Carrying out projects for industry as well as to cover our academic research and internal service needs.

MareNostrum: Some current applications

Isabel Campos Plasencia
University of Zaragoza
- **Fusion Group**
- **Research of nuclear fusion materials**
- **Follow-up of crystal particles**

Javier Jiménez Sendín
Technical University of Madrid
- **Turbulent channel simulation with Reynold numbers of friction of 2000**

Modesto Orozco
National Institute Nacional of Bioinformatics
- **Molecular dynamics of all representative proteins**
- **DNA unfolding simulation**



Markus Uhlmann
CIEMAT
- **Direct Numerical Simulation of Turbulent Flow With Suspended Solid Particles**

Gustavo Yepes Alonso
Autonomous University of Madrid
- **Hydrodynamic simulations in Cosmology**
- **Simulation of a universe volume of 500 Mpc (1.500 millions light year)**

Opportunities

- Access Committee
  - Research groups from
    - Spain
    - Mechanism to promote cooperation with Europe, …
- European projects
  - Infrastructure: DEISA
  - Mobility: HPC-Europa
- Call for researchers

HPC-Europa
Pan-European Research Infrastructure on High Performance Computing

Distributed European Infrastructure for Supercomputing Applications

Index

History

Barcelona Supercomputing Center – Centro Nacional de Supercomputación

MareNostrum description

Building the infrastructure

Setting up the system

Running the system

MareNostrum

- 4.812 PowerPC 970 FX processors
  - 2406 2-way nodes
- 9.6 TB of Memory
  - 4 GB per node
- 236 TB Storage Capacity
- 3 networks
  - Myrinet
  - Gigabit
  - 10/100 Ethernet
- Operating System: Linux
  - Linux 2.6 (SuSE)

Peak Performance
42.35 TFlops

MareNostrum: Overall system description

29 Compute Racks (RC01-RC29)
- 171 BC chassis w/OPM and gigabit ether switch
- 2392 JS20+ nodes w/myrinet daughter card

4 Myrinet Racks (RM01-RM04)
- 10 clos256+256 myrinet switches
- 2 Myrinet spines 1280s

7 Storage Server Racks (RS01-RS07)
- 40 p615 storage servers 6/rack
- 20 FastT 100 3/rack
- 20 EXP100 3/rack

1 Operations Rack (RH01)
- 7316-TF3 display
- 2 p615 mgmt nodes
- 2 HMC model 7315-CR2
- 3 Remote Async Nodes
- 3 Cisco 3550
- 1 BC chassis (BCIO)

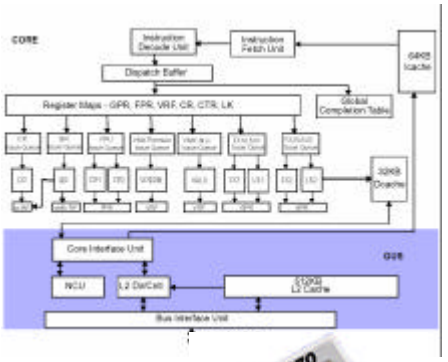
1 Gigabit Network Racks
- 1 Force10 E600 for Gb network
- 4 Cisco 3550 48-port for 10/100 network

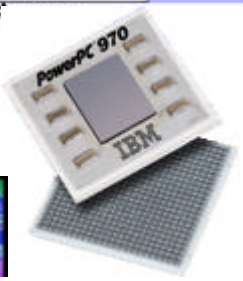
Environmental

	Individual	Frames	Composite
Compute	1200 Kg 21.6 KWatts 73699 BTUs/hr	29 172 * 7U chassis	34800 Kg 626,4 KWatts 2137271 BTUs/hr
Storage	440 Kg 6 KWatts 12000 BTUs/hr	7	3080 Kg 42 KWatts 84000 BTUs/hr
Management	420 Kg 1.5 KWatts 5050 BTUs/hr	1	420 Kg 1.5 KWatts 5050 BTUs/hr
Myrinet	40 Kg 1.4KWatts	4 12 * 14U chassis	480 Kg 16,8 Kwatts
Switch	128 Kg 5 KWatts 16,037BTU/hr	2	256 Kg 10 Kwatts 32 BTUs/h
TOTAL	Weight Power Heat AC Required Space		**39036 Kg** **696,7 Kwatts** **Over 2 million BTUs/hr** **180 Tons AC** **160 sq meters**

Hardware: PPC970FX

- PPC 970 FX @ 2.2 GHz:
  - 64 bit PowerPC implementation
  - 90 nm
  - 42W
  - + altivec VMX extensions
  - Featuring
    - 10 instr. issue
    - 10 pipelined functional units
    - L1: 64KB Instruction / 32KB data
    - L2 cache: 512KB
    - Support for large pages 16MB
      - **… leading to 8.8 Gflops peak**





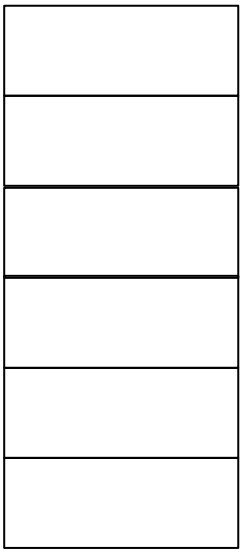
Blades, blade center and blade center racks

Blade Center
- **14 blades per chassis (7U)**
  - **28 processors**
  - **56GB memory**
- **Gigabit ethernet switch**

6 chassis in a rack (42U)
- **168 processors**
- **336GB memory**

29 bladecenter 1350 xSeries racks (RC01-RC29)

- Box Summary per rack
  - 6 Blade Center Chassis
- Cabling
  - External
    - 6 10/100 cat5 from MM
    - 6 Gb from ESM to E600
    - 84 LC cables to myrinet switch
  - Internal
    - 24 OPM cables to 84 LC cables

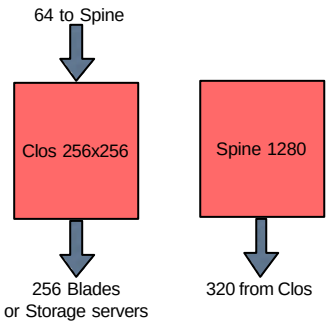


Myrinet racks

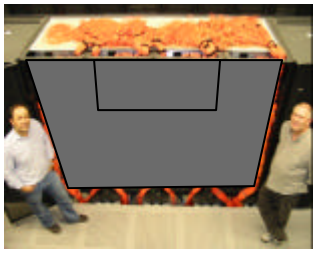
- 10 Clos 256x256 switches
  - Interconnect up to 256 Blades
  - Connect to Spine (64 ports)
- 2 Spine 1280
  - Interconnect up to 10 Clos 256x256 switches
  - Monitoring using 10/100 connection

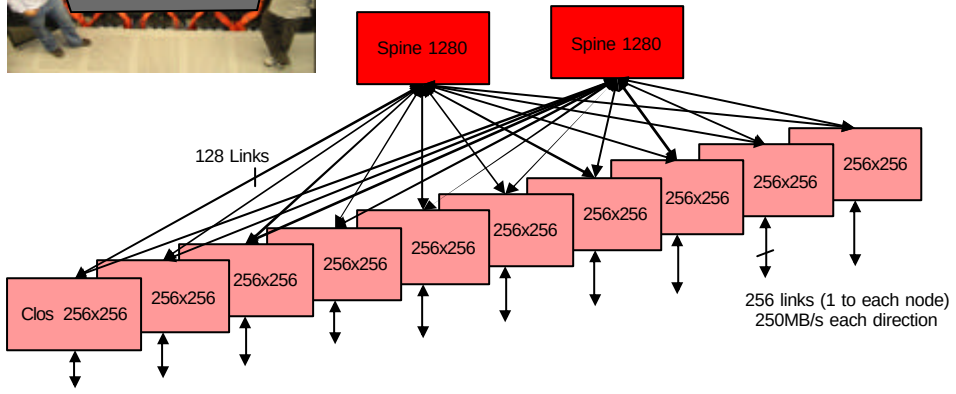
Myrinet racks

Clos 256x256	Spine 1280	Spine 1280	Clos 256x256
Clos 256x256	Clos 256x256	Clos 256x256	Clos 256x256
Clos 256x256	Clos 256x256	Clos 256x256	Clos 256x256



Myrinet

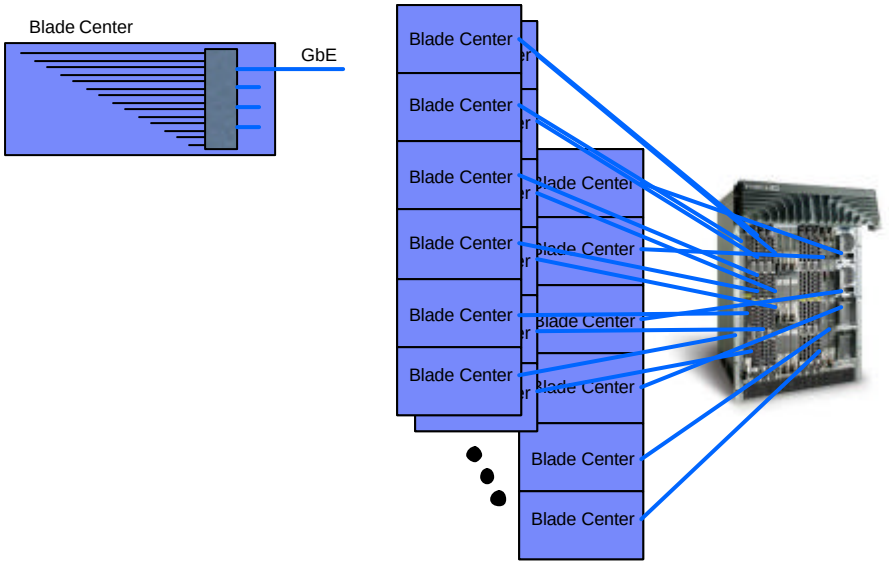




Gb Subsystem: Force 10 E600

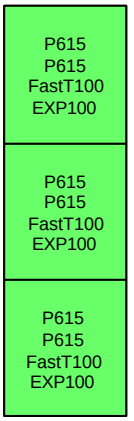
- Interconnection of Blade Centers
- Used for system boot of every blade center
- 212 internal network cables
  - 179 for blades
  - 42 for p615
- 67 connection available to external connection

Gb Ethernet

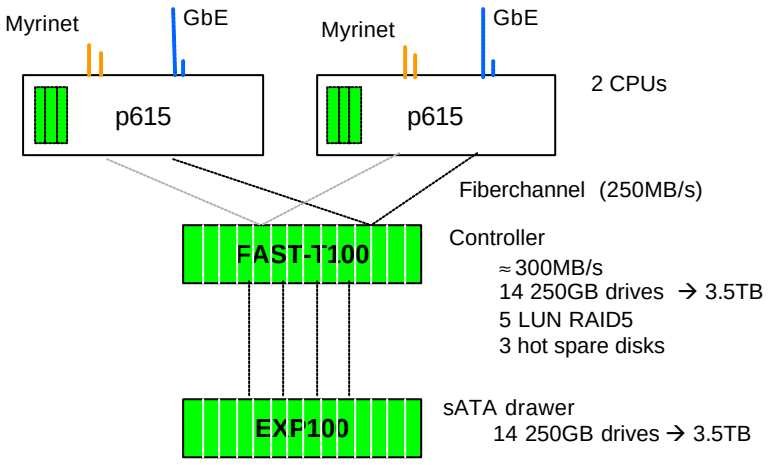


Storage nodes

- Total of 20 storage nodes, 20 x 7 TBytes
- Each storage node
  - 2xP615
  - FastT100
  - EXP100
- Cabling per node
  - 2 Myrinet
  - 2 Gb to Force10 E600
  - 2 10/100 cat5 to Cisco
  - 1 Serial



Storage node



Index

History

Barcelona Supercomputing Center – Centro Nacional de Supercomputación

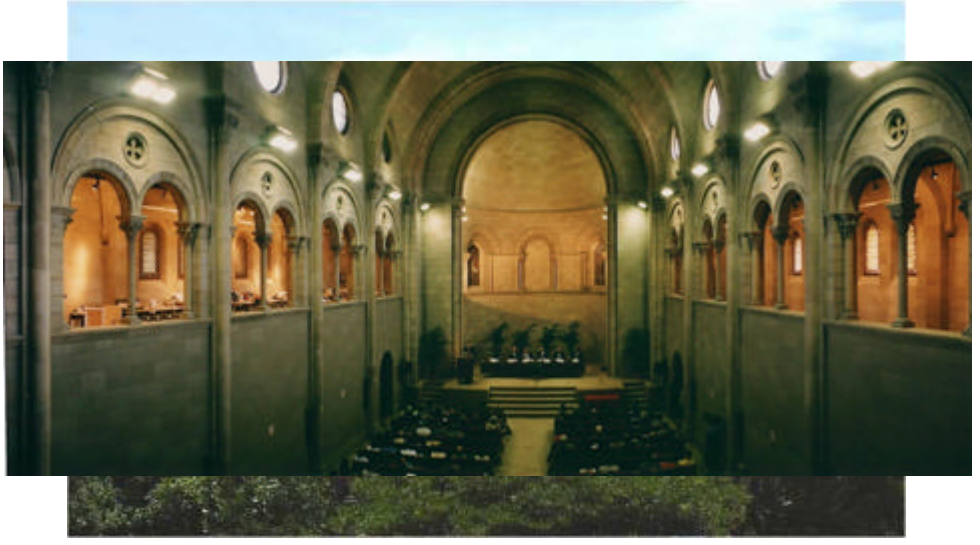
MareNostrum description

Building the infrastructure

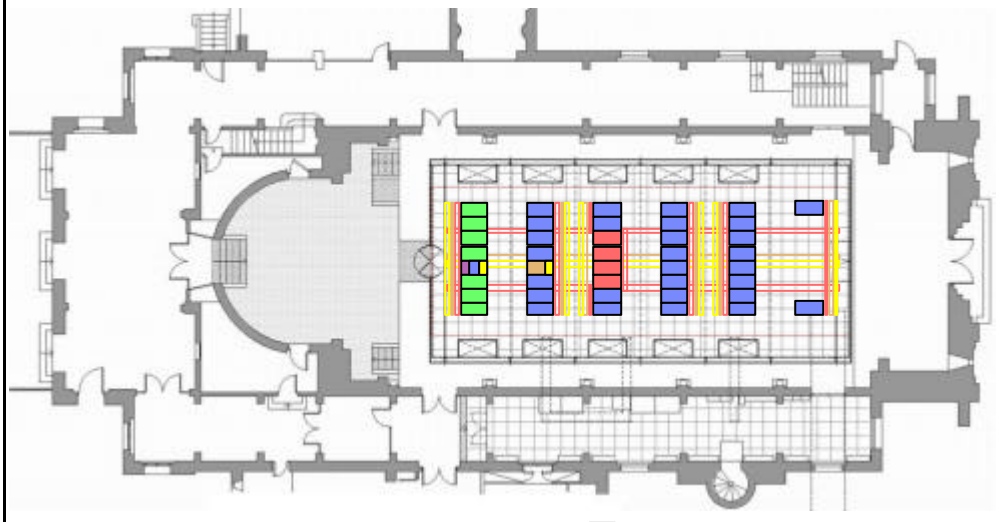
Setting up the system

Running the system

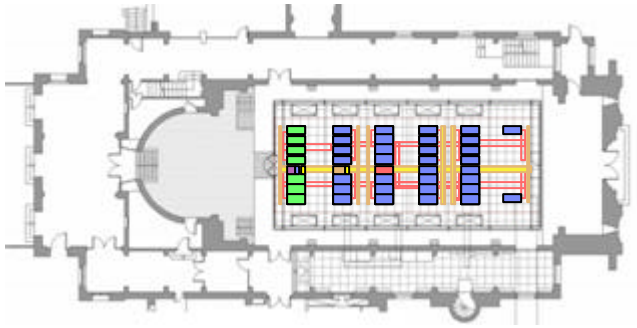
Location

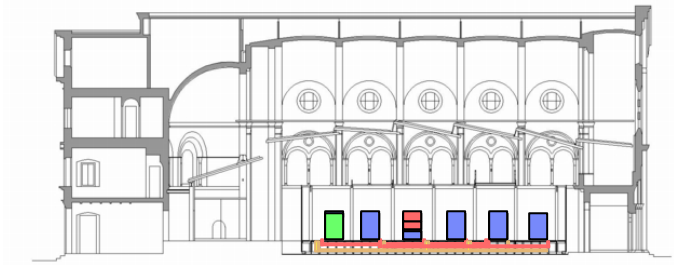


MareNostrum Floorplan

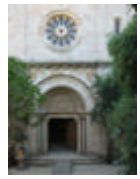


MareNostrum Floorplan





- Glass Box:
  - 18.74 x 9.04 x 4.97 m
  - False floor: 0.97 m
  - Area: 170m2
  - Volume: 660m3 + 170m3
  - Steel: 26 tons
  - Glass: 19 tons



Service

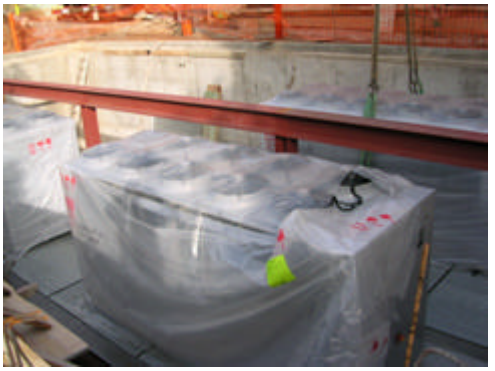
- The hole
  - 15.5 x 16 x 5.4 m
- Power
- External AC

Power

- 3 transformers from High to Low voltage
  - Machine
  - Air Conditioning + others
  - Redundant
- UPS
  - Disk servers + networking + some internal AC
- Generator (diesel)
  - Disk servers + networking + some AC

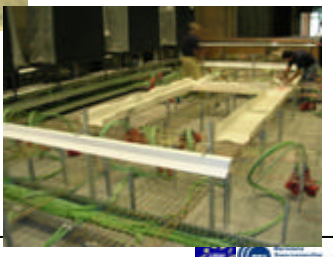
AC

- 4 External Units
  - 7ºC → 12ºC
- 2 water tanks
  - 25000 liters
  - 2 pumps (connected to generator)
- 10 Internal Units
  - 16ºC → 26ºC



Air conditioning, power, cabling, fire detection

Site preparation



The movie

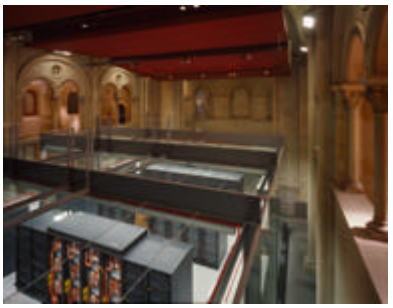
From July 7th to October 20th, 2005

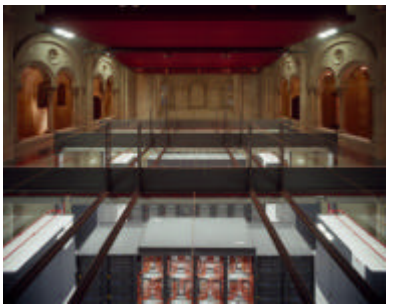
Index

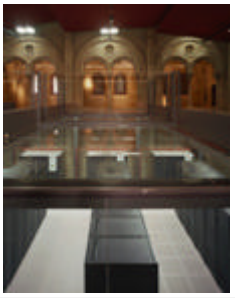
Mounting the system

From November 27th to December 7th, 2005

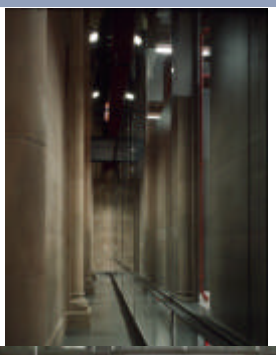
Site preparation

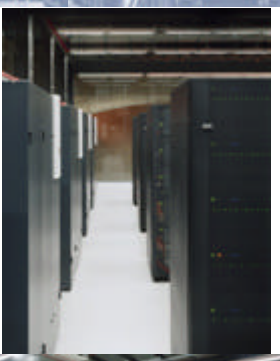




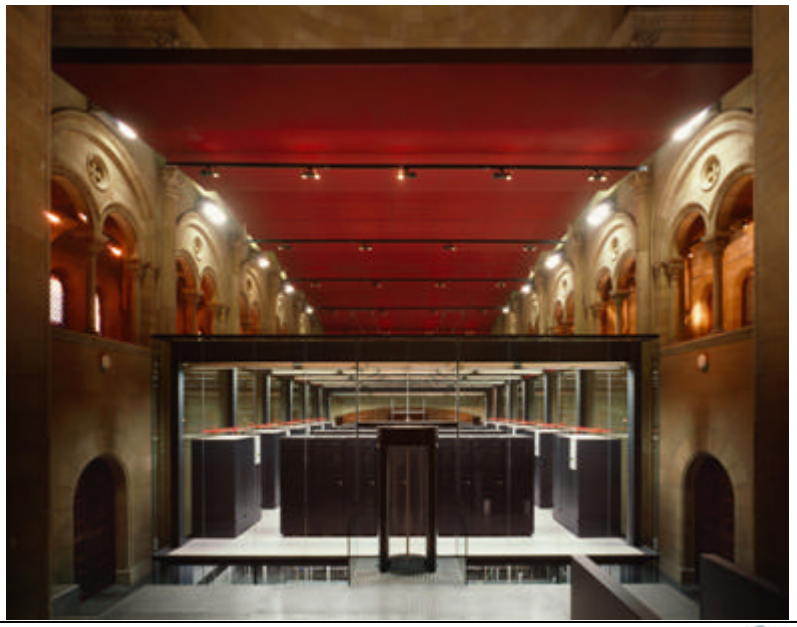


Site preparation



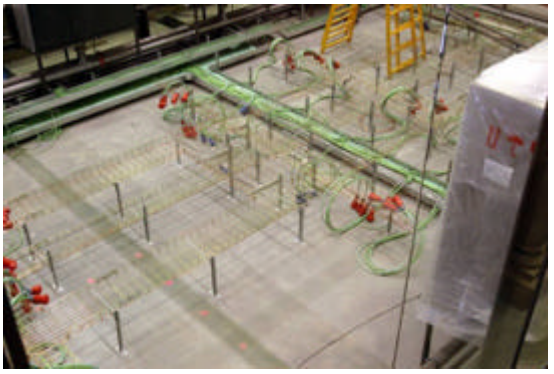


Site preparation



Cables

- Myrinet
  - 172*14 + 40 fibers (25 meters)
  - Near 61 km
- Gigabit and Ethernet (x2)
  - 212 cupper (25 meters)
  - 5,3 km
- Power
  - Blade Center rack: 4 * 29
  - Disk server rack: 3 * 7
  - Myrinet rack: 6 * 4



Index

Software

- Diskless boot
  - 2 mins. 1 node
  - ≈15 mins.
- Linux
  - 2.6 SuSE
- Each P615, using their SCSI disks, hosts via NFS
  - Root file system
  - Var file system
  - for 40 blades

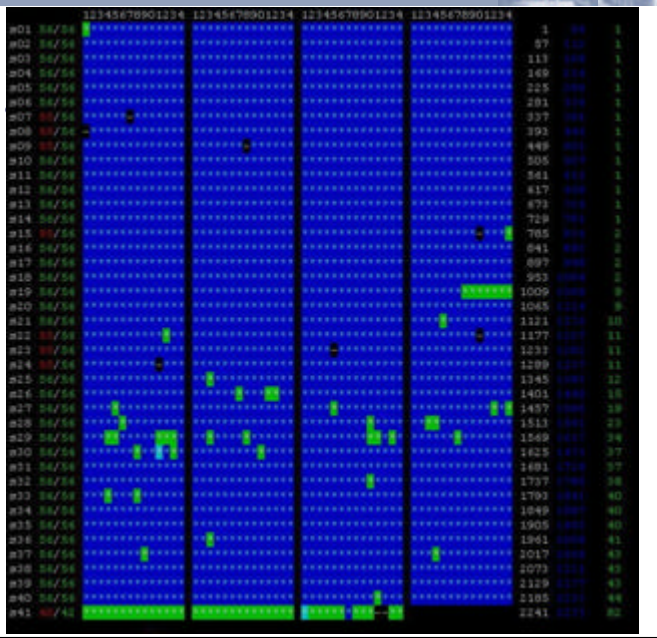
Software

- GPFS
  - Basic shared file system
    - Home, projects, scratch, apps
  - Scalability
    - Largest tested site ever: 1100 nodes
    - Testbed till 2406
  - Through GbE

Software

- LoadLeveler
  - Scalability:
    - Official: 1 job → 128 nodes
    - Tested: → 400
  - New version soon
  - Alternative: Slurm

Software



Thank you !