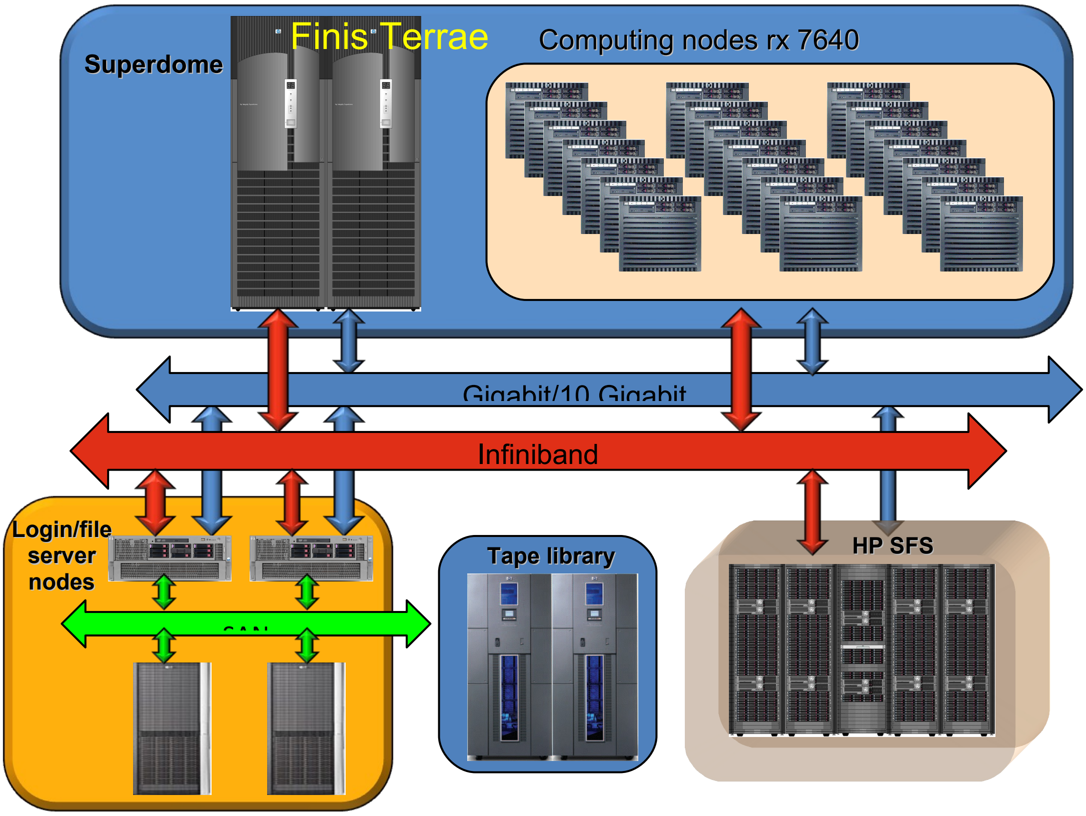
FINIS TERRAE: ARCHITECTURE

Finis Terrae is a complete solution intended to solve the problems of the community of researchers using CESGA. One fundamental aspect has been taken into account in its design, the **coexistence of great computational challenges**, which require thousands of processors and huge amounts of memory up to 20 TeraBytes, **with other kind of calculus, perhaps more daily, but important as well.**
At the same time, its **shared memory** architecture allows a scaling to a significant number of processors, which is necessary in order to deal with a high number of data.
As for blocks, FinisTerrae is made up of the following ones:

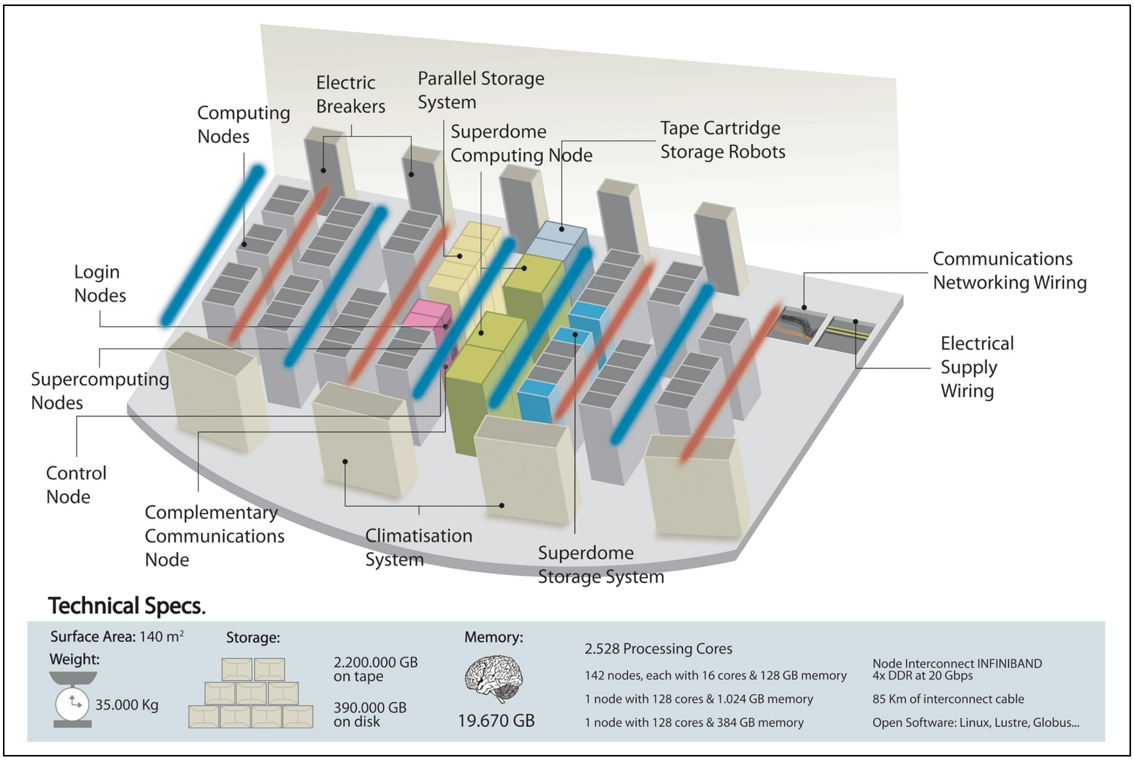


Computation Nodes rx7640
Computation Nodes Superdome
Interconnection Network Infiniband
HP-SFS storage system
Login nodes/file servers
Tape library

All the system components work using Linux operative system, in its SuSe version.

FINIS TERRAE lay out:

All these blocks are distributed in the Data Centre according to the disposition in the lay out below.


Computing Nodes
Electric Breakers
Parallel Storage System
Superdome Computing Node
Tape Cartridge Storage Robots
Communications Networking Wiring
Electrical Supply Wiring
Login Nodes
Supercomputing Nodes
Control Node
Complementary Communications Node
Climatisation System
Superdome Storage System
Technical Specs.
Surface Area: 140 m²
Weight:
35.000 Kg
Storage:
2.200.000 GB on tape
390.000 GB on disk
Memory:
19.670 GB
2.528 Processing Cores
142 nodes, each with 16 cores & 128 GB memory
1 node with 128 cores & 1.024 GB memory
1 node with 128 cores & 384 GB memory
Node Interconnect INFINIBAND 4x DDR at 20 Gbps
85 Km of interconnect cable
Open Software: Linux, Lustre, Globus...

Curiosities:

Area: The Finis Terrae takes up a 142 square metres area, the same as one family apartment.

Weight: 33,5t, equivalent to 35 medium-size private cars

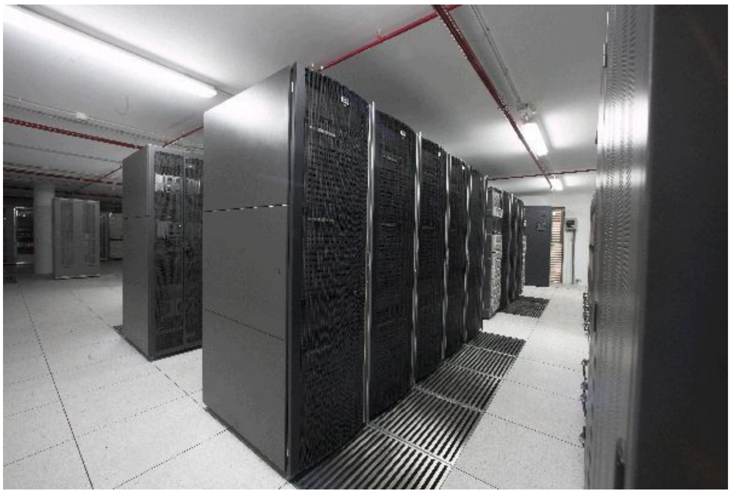
Hardware:

- It has 2,500 64-bit CPUs able to calculate at a speed equivalent to 10,000 last-generation PC working in parallel.
- 19,000 Gbytes RAM memory, where the 20 million books of the largest library in the world could be stored (US Congress).
- 390 Tbytes on-disk storage, where we could place more than half a million high-quality feature-length films.
- The Infiniband interconnection wiring measures the exact distance between the CESGA in Santiago de Compostela and the city of Finisterre (85 Km).
- The 20Gbps interconnection speed through last-generation optical fibre means transmitting the contents of 1,800 DVD in one hour's time.

Computation Nodes: rx7640 + Superdome

The computation nodes are the elements intended to solve the mathematic operations of the simulations.

There are two kinds of them in Finis Terrae: Rx7640 and Superdomes

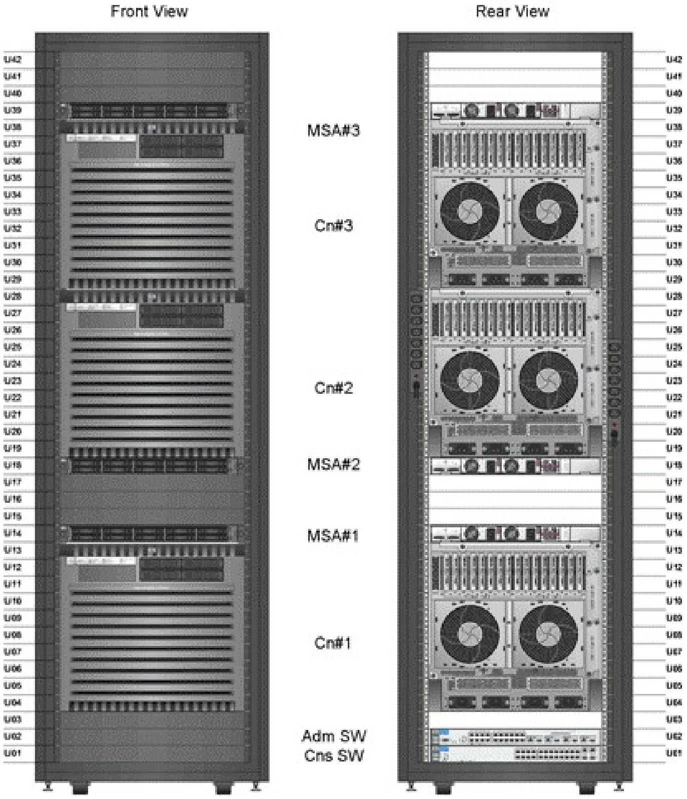


Both kinds of nodes have the same architecture and from the user's point of view, the only difference between them is the size of the node.

The **HP rx7640** are servers with **16 processor cores and 128 GB memory**, while **the main HP Superdome node has 128 processor cores and 1Tb memory**.

Rack View:

The computation nodes HP rx7640 are composed of 48 racks with the following components:



- In each rack:

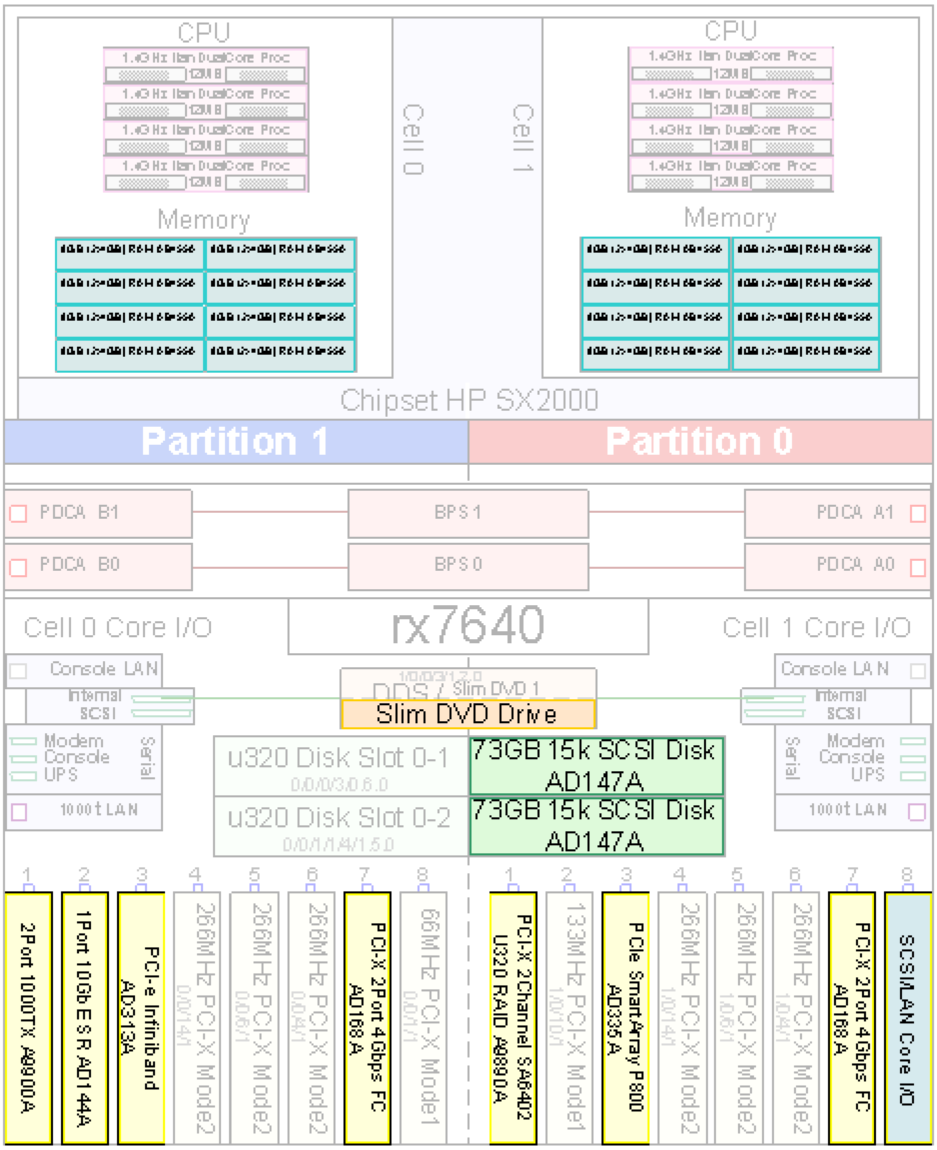
- **3 HP Integrity rx7640 servers** with 8 Intel dual core Itanium-2(Montvale) 1.6 Ghz/18 MB CPUs (16 cores), with 2 x 73GBs 15K rpm internal disks, with 8 x HP 8GB DDR2 Memory ModulesPair;
- **3 HP MSA50** 1U SFF SAS JBOD Enclosure with 8x 72GBs SAS 10K rpm disks,
- InfiniBand 4X PCI-E 2 Prt **Host Adapter,** InfiniBand 4X PCI-E 2 Prt Host Adapter and one P Smart Array P800 SAS **Controller**:

- There are **142 servers in all**, each one with:

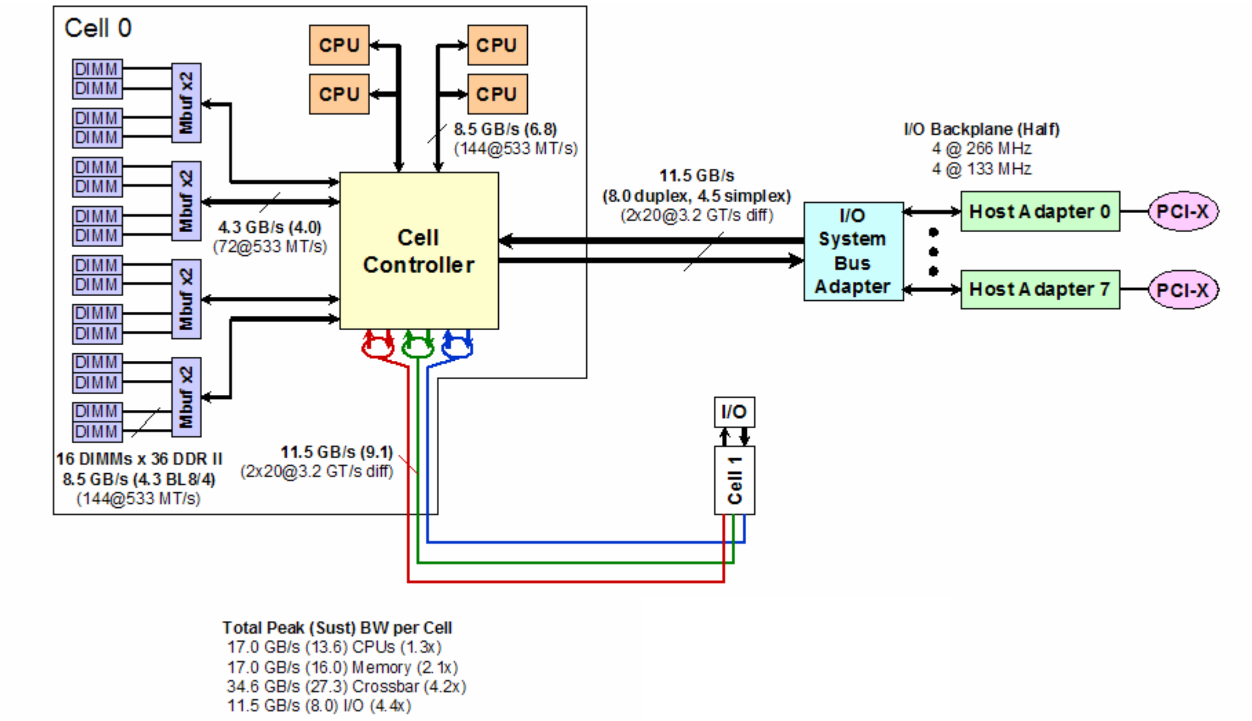
- 16 núcleos Itanium 2 Montvale @ 1.6 GHz , 18 MB cache
- 128 GB memory
- 6 146 GB SAS disks
- Suse Linux

Logical View:

There are **two cells** in the **rx7640 nodes**, while **the Superdome has 16**. **Each cell has 4** last-generation **1.6 GHz Itanium Montvale processors**, each of which incorporates **two processor cores and 18 MB cache**. The cell also provides a **64 GB memory** block.



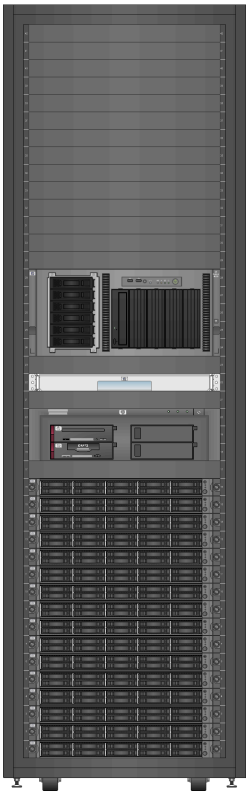
One-cell processors access the local memory of their cell through the 'cell controller' and there exists an interconnection of the different cell controllers that allows each processor to access directly the remote cell memory as well, practically at the same speed, but with a slightly greater latency.



The **architecture** of the system is thus **ccNUMA**: within each computation node, each processor accesses directly to all the memory, although the access time is slightly different depending on whether we are inside the local cell or we are accessing to a remote one. The operative system is in charge of arranging each processor's affinity to its cell's memory, trying to put the data accessed by each processor in their local cell.

Superdomes:

The system also includes other two Superdomes, which are also used as computation nodes.



- The **2 Superdomes** are distributed in 4 racks as the one in the image. Each Superdome has:

- **64 Processors /128 cores Itanium 2 Montvale @ 1.6 GHz, 18 MB cache that provide other 384 GB of main memory.**

- **1 TB memory in a single image**

- **128 72 GB SAS disks** (9.2 TB) for scratch

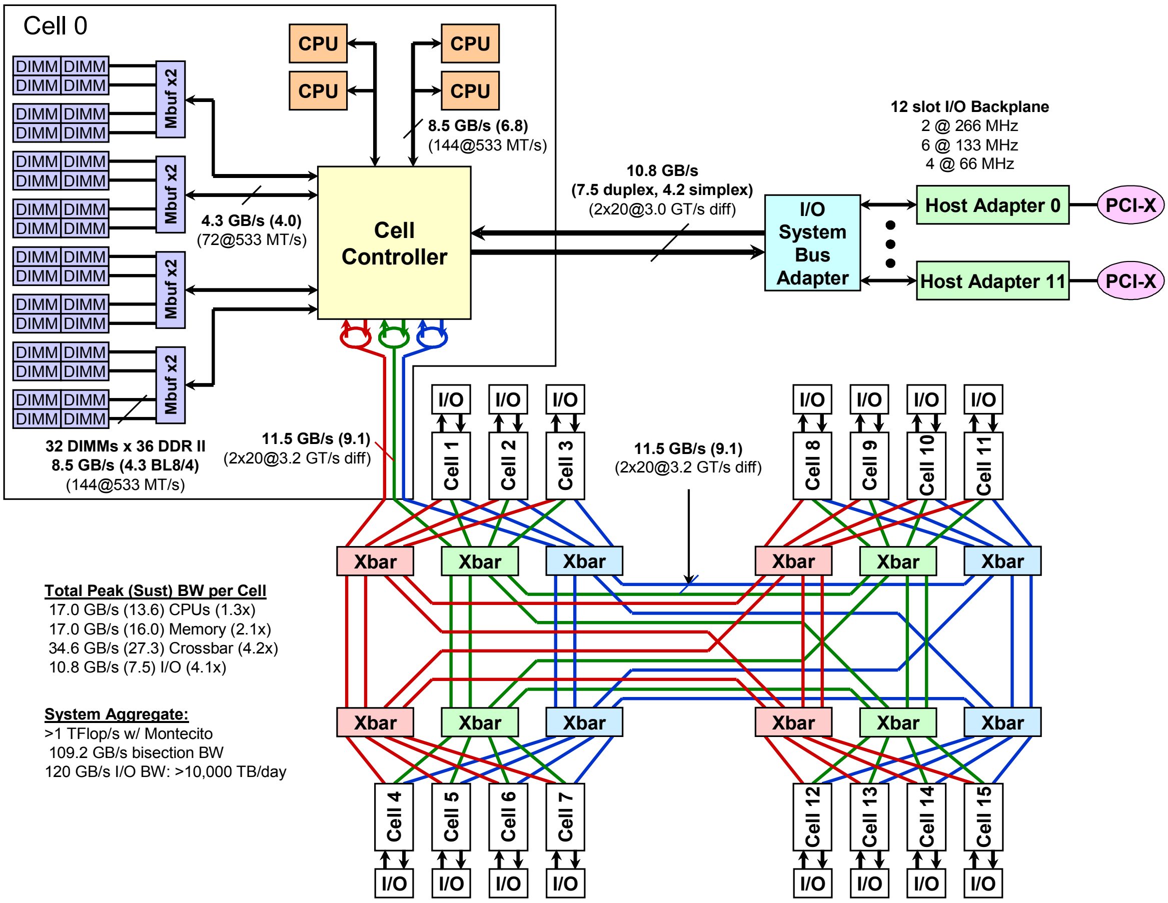
- **Suse Linux** SLES 10

- **High Availability characteristics: I**t has N+1 OLR ventilators, N+1 OLR power supplies, double current supply, OLAR for cells, OLAR for I/O cards, ECC in CPUs, memory and all data ways, Dynamic Processor Resilience, Dynamic Memory Resilience (**double** Chip Kill), and dual ways between the switches and the Cell Controller, the memory and the CPUs.

- **Partitioning** capacity: up to 16 physical partitions and up to 64 virtual partitions, PRM, WLM, IVM in HP-UX, gWLM multiSO.

▪ It has the *certificate of electrical failure tolerance*, issued by the Uptime Institute.

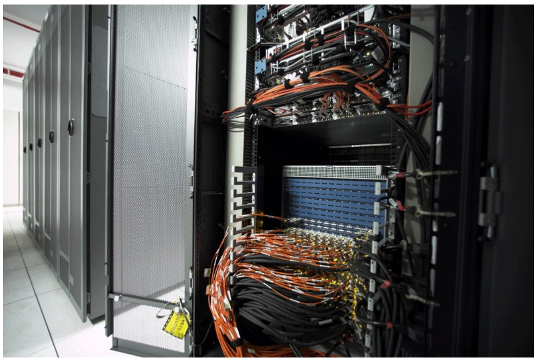
Internally, each node is composed of the union of cells as that in the image below.



The different cells are linked by three different ways. In the case of the Superdome, the 16 cells are interconnected by means of a network of type-Crossbar switches in a two-level architecture.

The different computation nodes can work with independent users' problems, that is, each user work is executed within one single node, but of course for work requiring the utilisation of more processors or more memory than that in one node, therefore the system has an Infiniband communications network.

Infiniband Network



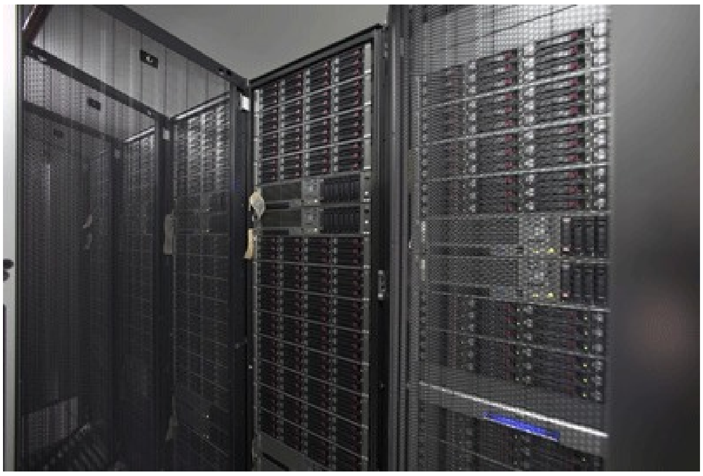
This network has as its most important characteristics the 16 Gbps high interconnection band width, as well as a very low latency, that for the user is 6 microseconds. Each of the computation and storing nodes is connected directly to a great central switch, so every pair of system elements can always communicate at the same speed.

All the elements (computation nodes/SFS/Login) connect to a single Voltaire ISR 9288 switch, which delivers up to 288 ports of 10 gigabits per second InfiniBand connectivity; leads the InfiniBand provide 10Gpbs InfiniBand connectivity and a wire-speed non-blocking switch capacity combined with latency of less than 450 nanoseconds. Also, the switch components are hot-swappable and redundant to allow for the highest availability.

Storage

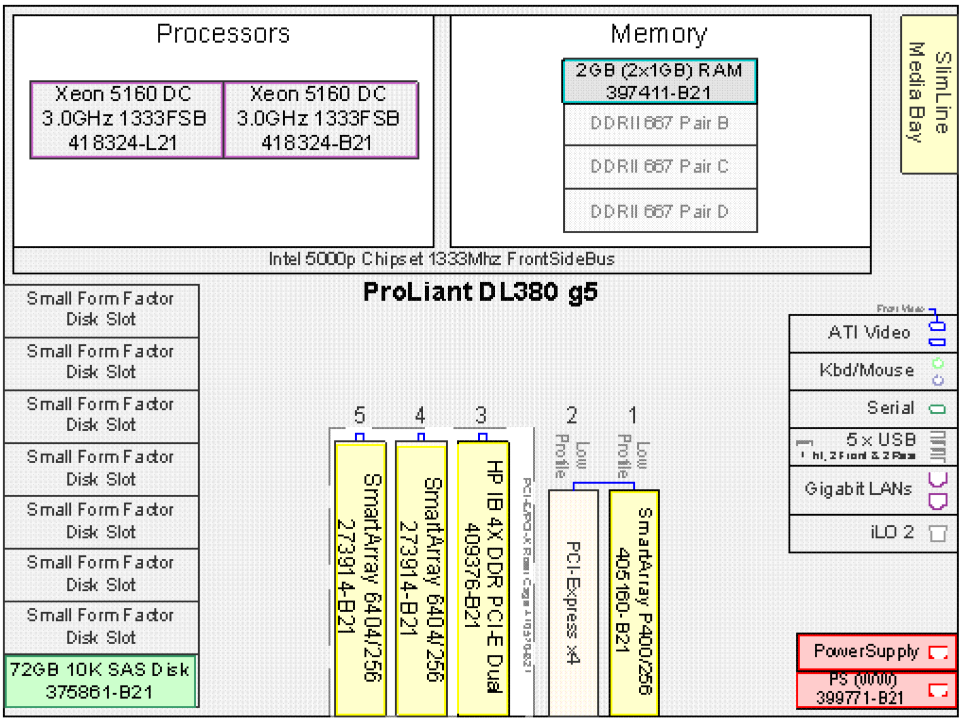
HP SFS

Another important element in the Finis Terrae solution is the **HP-SFS storage system.**



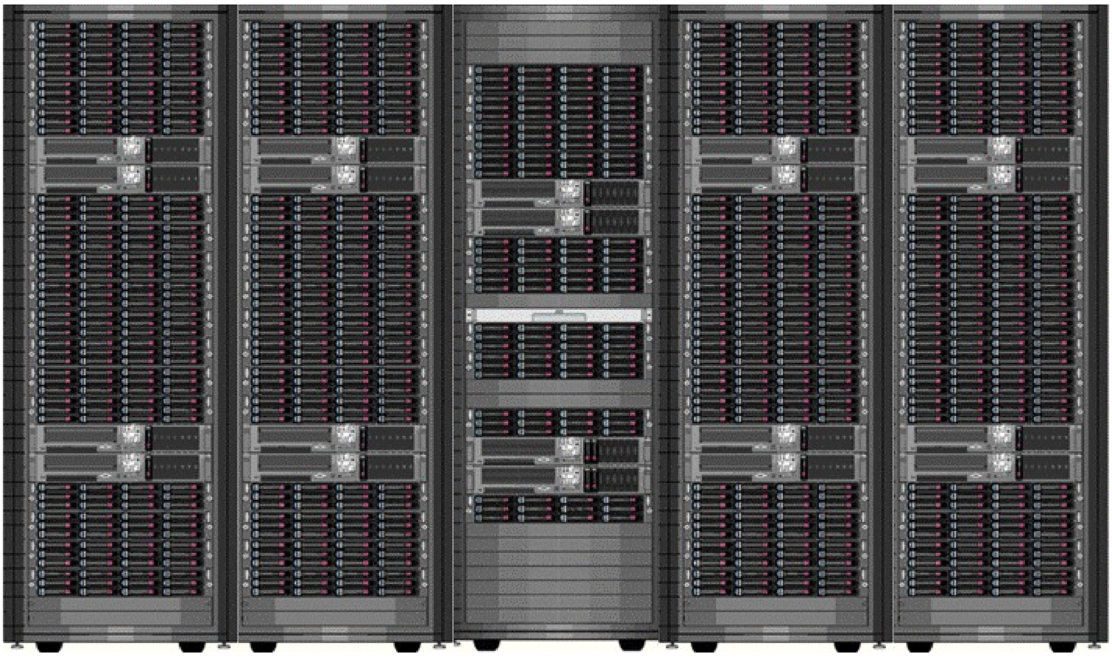
It is a **parallel storage** system, made up of a total amount of 20 servers that have **72 cabins connected to 864 SATA type disks**. The system acts as a great **216 Terabytes** hard disk, shared by all the computation nodes, in a way that the information of this system can be accessed from any of them and at a high speed. *Parallel supercomputing systems need to treat and store great amounts of information, and this is only possible to be made at a high speed with parallel storage systems such as this one*, based on **Lustre** free software.

The computation nodes are reserved for the users' problems execution, so the system has some connection servers, which are the point where the users connect and from which they can send and check the state of their works, etc. There exists a redunded configuration that also provides access to the rest of CESGA's storage cabins, where users' permanent data are normally located, available in the home directory.


Processors
Xeon 5160 DC 3.0GHz 1333FSB 418324-L21
Xeon 5160 DC 3.0GHz 1333FSB 418324-B21
Memory
2GB (2x1GB) RAM 397411-B21
DDRII 667 Pair B
DDRII 667 Pair C
DDRII 667 Pair D
SlimLine Media Bay
Intel 5000p Chipset 1333Mhz FrontSideBus
ProLiant DL380 g5
Small Form Factor Disk Slot
Small Form Factor Disk Slot
Small Form Factor Disk Slot
Small Form Factor Disk Slot
Small Form Factor Disk Slot
Small Form Factor Disk Slot
Small Form Factor Disk Slot
72GB 10K SAS Disk 375861-B21
5
4
3
2
1
Low Profile
Low Profile
SmartArray 6404/256 273914-B21
SmartArray 6404/256 273914-B21
HP IB 4X DDR PCI-E Dual 409376-B21
PCI-Express x4
SmartArray P400/256 405160-B21
ATI Video
Kbd/Mouse
Serial
5 x USB
Gigabit LANs
iLO 2
PowerSupply
PS (000) 399771-B21

The HP SFS storage solution is physically composed of 5 racks, such as the one shown in the image below:

In each rack, there are 16 **HP Storageworks SFS20 Enclosures** with 12 250 GBs SATA disks each, and 4 **HP Proliant DL 380R05** servers with 2 Proc./4cores, 2 GB FBD, 1 72 GB SAS disk , 2 SmartArray 6404, 2 redunded power supplies. So for each OSS Server DL380, there are 4 StorageWorks SFS20 y 48 250 SATA disks . In one of the racks there are also 2 HP ProCurve 2626 switches.



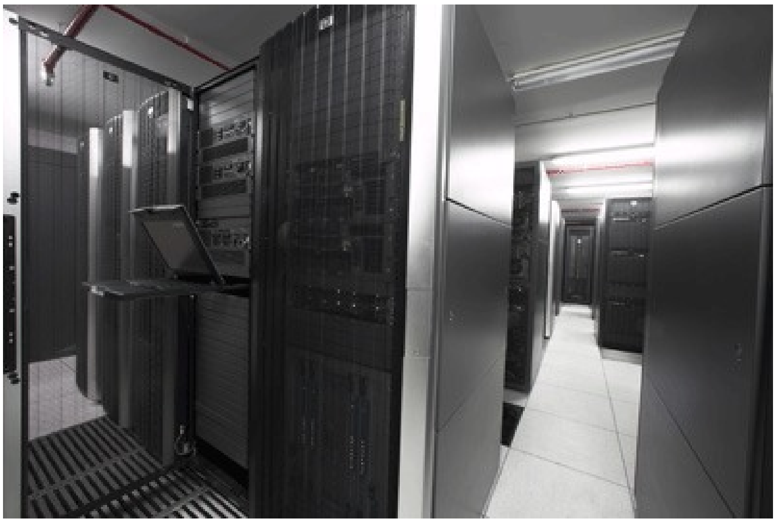
Thus, the result is an external storage system based on HP Scalable File Sharing (HP SFS) Cluster Filesystem of a great scalability and bandwith, with a 219TB raw capacity. The interconnection of the cells to each other and to the calculus system is made through backbone Infiniband. It includes: LINUX operative system, MPI libraries and HP and INTEL mathematics; compilers, purifiers and INTEL optimizers.

“Hendrix”

Lustre File System is the software used to manage. Lustre is a Project financed jointly by the US DOE (Department of Energy), Intel, CFS*, and HP through the “Tri-National Labs” (Livermore, Los Alamos and Sandia). Lustre is open-source technology implemented in equipments by different manufacturers.

- *Cluster File System, Inc (CFS) www.clusterfs.com is the main developer.*
  - Features: Great bandwith, using high-speed parallel interconnections
  - Efficiency: great and small I/O operations per second
  - Trustful: coherence as in the I/Os of an SMP
  - Simple: APIs Standard support, including MPI I/O & POSIX
  - Robust: uses very established file systems

Login nodes/File servers



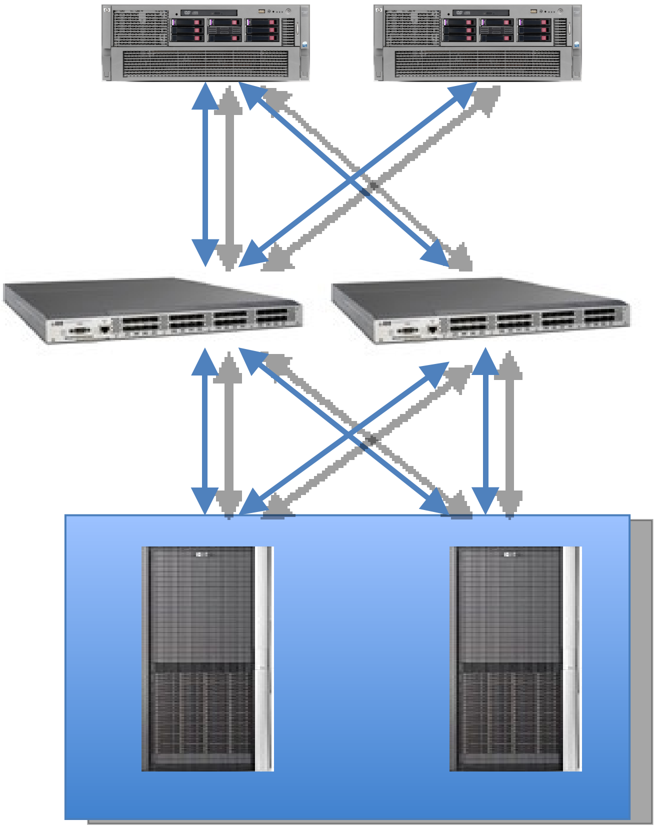
The computation nodes are reserved for the users’ problems execution, so the system has some connection servers, which are the point where the users connect and from which they can send and check the state of their works, etc. There exists a redunded configuration that provides access to the rest of CESGA’s storage cabins, where users’ permanent data are normally located, available in the home directory.

These login nodes or **redundant configuration** access servers have two main functions:

● login to the cluster:

- Check-up of the state of the works
- Sending of new works to queue
- Eliminate works from the queue
- Create and edit files
- Make file transferences to/from Finis Terrae
- Do some pre-processing and post-processing actions
- Access the interactive nodes (*compute* command)

● File systems /home (nfs)

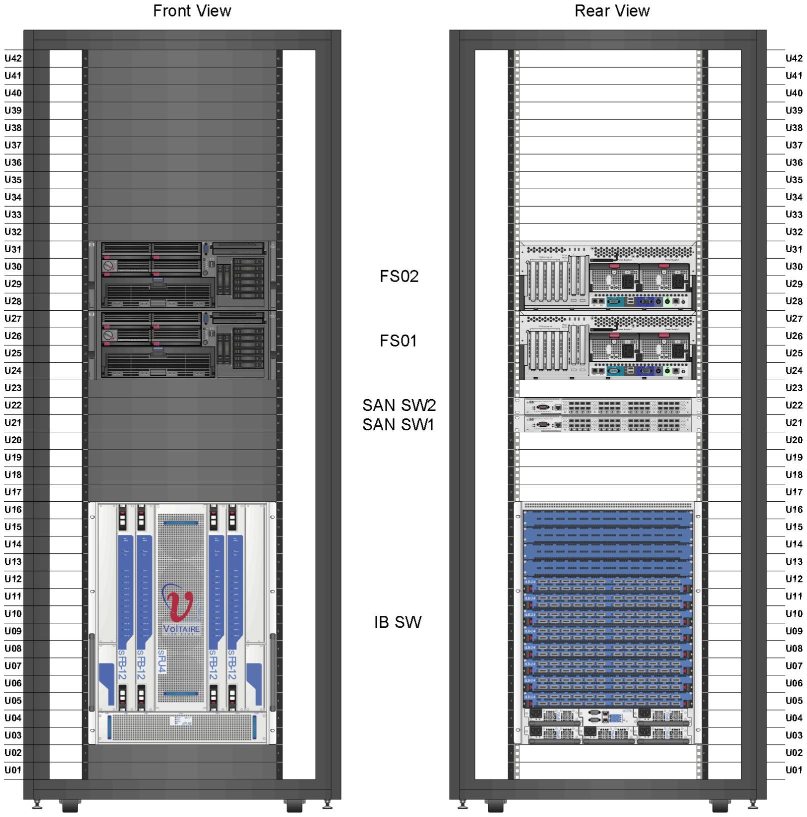


These access nodes are 2 DL 580 servers that are placed within the 5 HPSFS solution racks. The rack where they can be found is exactly shown in the image below and it has the following elements:

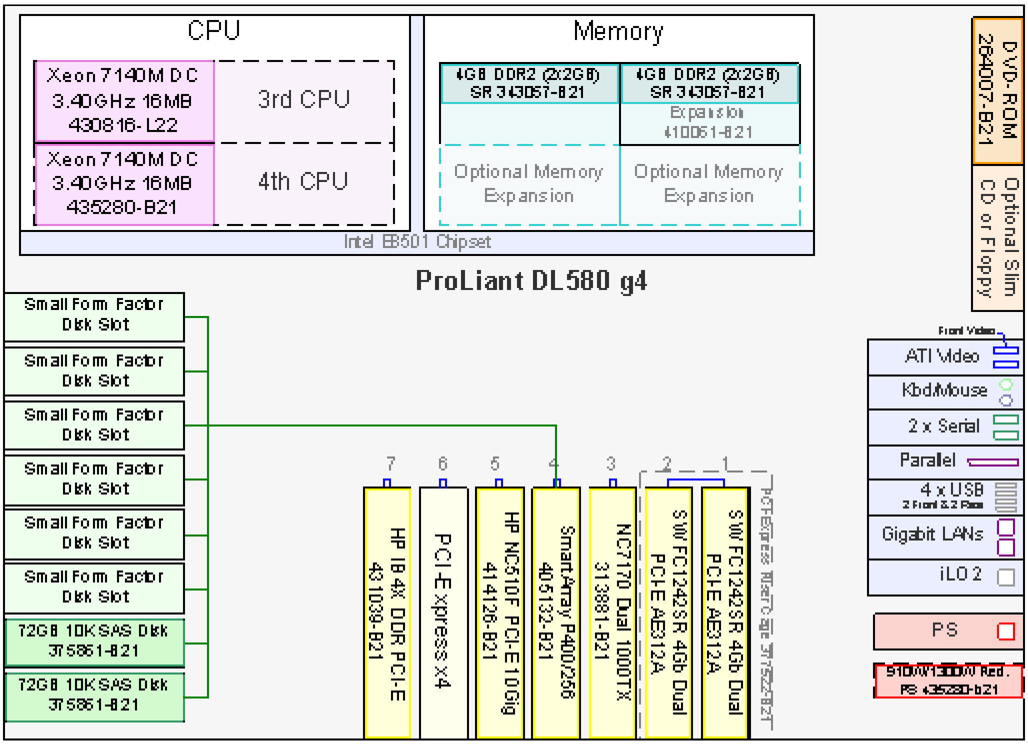
▪ 2 **HP Proliant DL580-G4** servers with 2 Processors (4cores), 8 GB FBD memory, 2 72 GB SAS disks, and one SmartArray P400. With the HP software Serviceguard for Linux.

▪ 2 4GBs SAN switches and 32 ports

▪ 1 Voltaire 288 ports IB switch.



The two access servers are configured in an identical way, as we see in their logical figure:


CPU
Xeon 7140M DC 3.40GHz 16MB 430816-L22
3rd CPU
Xeon 7140M DC 3.40GHz 16MB 435280-B21
4th CPU
Memory
4GB DDR2 (2x2GB) SR 343057-B21
4GB DDR2 (2x2GB) SR 343057-B21 Expansion 410061-B21
Optional Memory Expansion
Optional Memory Expansion
Intel E8501 Chipset
ProLiant DL580 g4
DVD-ROM 264007-B21
Optional Slim CD or Floppy
Small Form Factor Disk Slot
Small Form Factor Disk Slot
Small Form Factor Disk Slot
Small Form Factor Disk Slot
Small Form Factor Disk Slot
Small Form Factor Disk Slot
72GB 10K SAS Disk 375861-B21
72GB 10K SAS Disk 375861-B21
7
6
5
4
3
2
1
HP IB 4X DDR PCI-E 431039-B21
PCI-Express x4
HP NC510F PCI-E 10Gig 414126-B21
Smart Array P400/256 405132-B21
NC7170 Dual 1000TX 313881-B21
SW FC1242SR 4Gb Dual PCIE AE312A
SW FC1242SR 4Gb Dual PCIE AE312A
PCI-Express Riser Cage 375522-B21
Front Video
ATI Video
Kbd/Mouse
2 x Serial
Parallel
4 x USB
2 Front & 2 Rear
Gigabit LANs
iLO 2
PS
910W/1300W Red. PS 435280-b21

Tape libraries

The system has, at the same time, **a HP ESL 712e tape library that provides a storage capacity of 2,2 Petabytes of information, in its 1424 slots of type LTO4 tapes.** The library has 12 reading units that allow a high transference speed, of almost 3 Gigabytes/seg.

- 2 HP ESL 712e robotized libraries with pass-through
- 12 LTO4 drives with 800 GB capacity without compression
- 1.424 slots for tapes in all (1.140 TB without compression, > 2 PetaBytes with compression)
- 1.9 GB/s transference speed
- Connection to the SAN
- It is managed with the HP Data Protector backup software, HP Command View y HP Secure Manager