MareNostrum 3

In August 2012, an agreement was signed between the Spanish government and IBM in order to update the supercomputer.

MareNostrum 3 was a supercomputer based on Intel SandyBridge processors, iDataPlex Compute Racks, a Linux Operating System and an Infiniband interconnection. See below a summary of the system:

- Peak Performance of 1,1 Petaflops
- 115.5 TB of main memory
- 3,056 nodes
- Main Nodes
  - 2x Intel SandyBridge-EP E5-2670/1600 20M 8-core at 2.6 GH
  - 128 nodes with 8x16 GB DDR3-1600 DIMMS (8GB/core)
  - 128 nodes with 16x4 GB DDR3-1600 DIMMS (4GB/core)
  - 2752 nodes with 8x4 GB DDR3-1600 DIMMS (2GB/core)
- 3 PB of disk storage

- Xeon Phi Nodes 42 heterogeneous compute nodes
  - 2x Intel SandyBridge-EP E5-2670/1600 20M 8-core at 2.6 GHz
  - 2x Xeon Phi 5110 P
  - 8x8GB DDR3-1600 DIMMS (4GB/core)
- Interconnection networks:
  - Infiniband FDR10

    - Gigabit Ethernet
- Operating System: Linux - SuSe Distribution
- MareNostrum has 52 racks and takes up a space of 120m2.

Specific technical characteristics of the main components of the machine:

- MareNostrum III had 36 racks dedicated to calculations. These racks had a total of 48,448 Intel SandyBridge cores with a frequency of 2.6 GHz and 94.625 TB of total memory.
- In total, each rack had a total of 1,344 cores and 2,688 GB of memory.
- The peak performance per rack was 27.95 Tflops, and a peak power consumption of 28.04 kW.

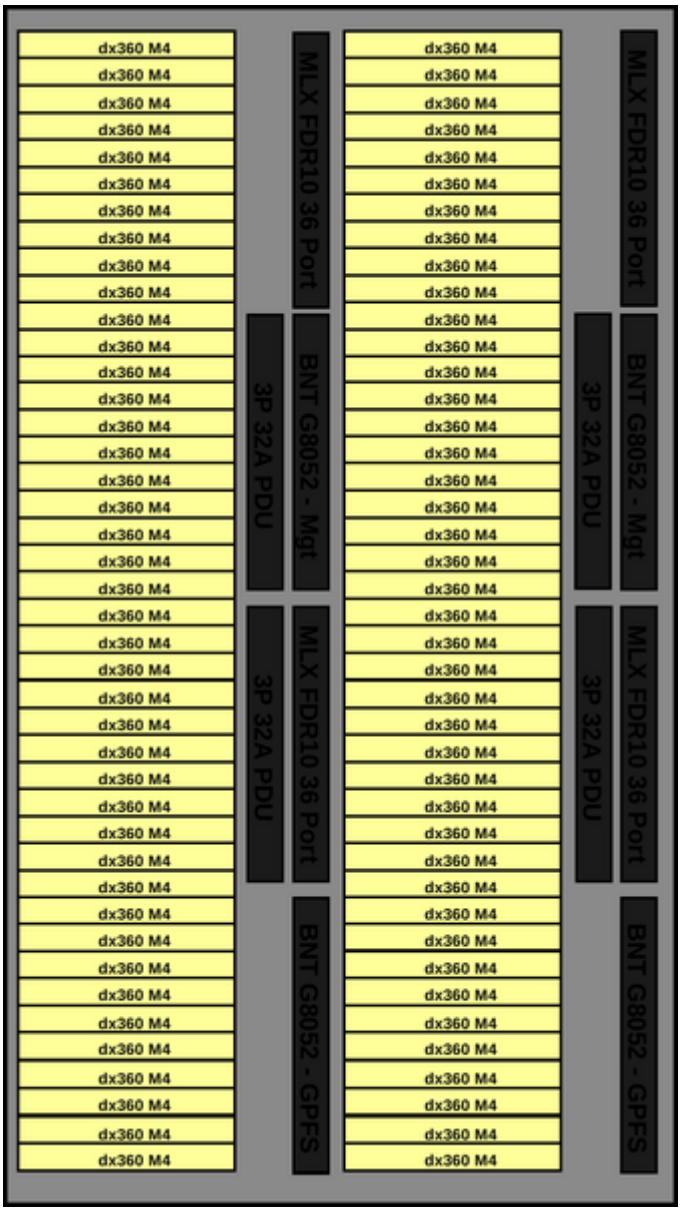
Compute Rack

Each IBM iDataPlex Compute rack was composed of:

- 84 IBM dx360 M4 compute nodes
- 4 Mellanox 36-port Managed FDR10 IB Switches
- 2 BNT RackSwitch G8052F (Management Network)
- 2 BNT RackSwitch G8052F (GPFS Network)
- 4 Power Distrbution Units

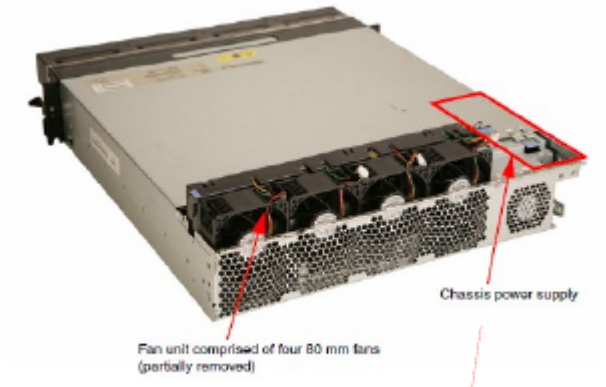
According to the performance per rack:

- 2.60 GHz x 8 flops/cycle (AVX) = 20.8 Gflops/core
- 16 core x 20.8 Gflops/core = 332.8 Gflops/node
- 84 nodes x 298.64 Gflops/node = 27.95 Tflops/rack

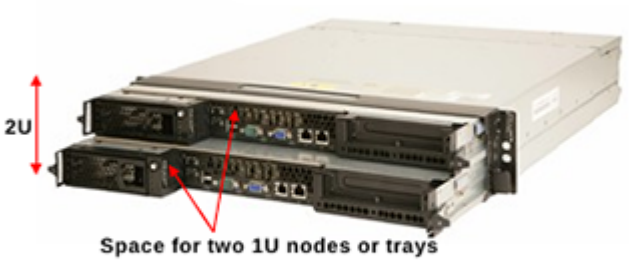


Chassis

dx360 M4 compute nodes were grouped into a 2U Chassis, having two columns of 42 2U Chassis. Each 2U Chassis had two 900W independent power supplies for redundancy purposes. In this way, if one of the supplies failed there was the other one still working.



Rear of Chassis



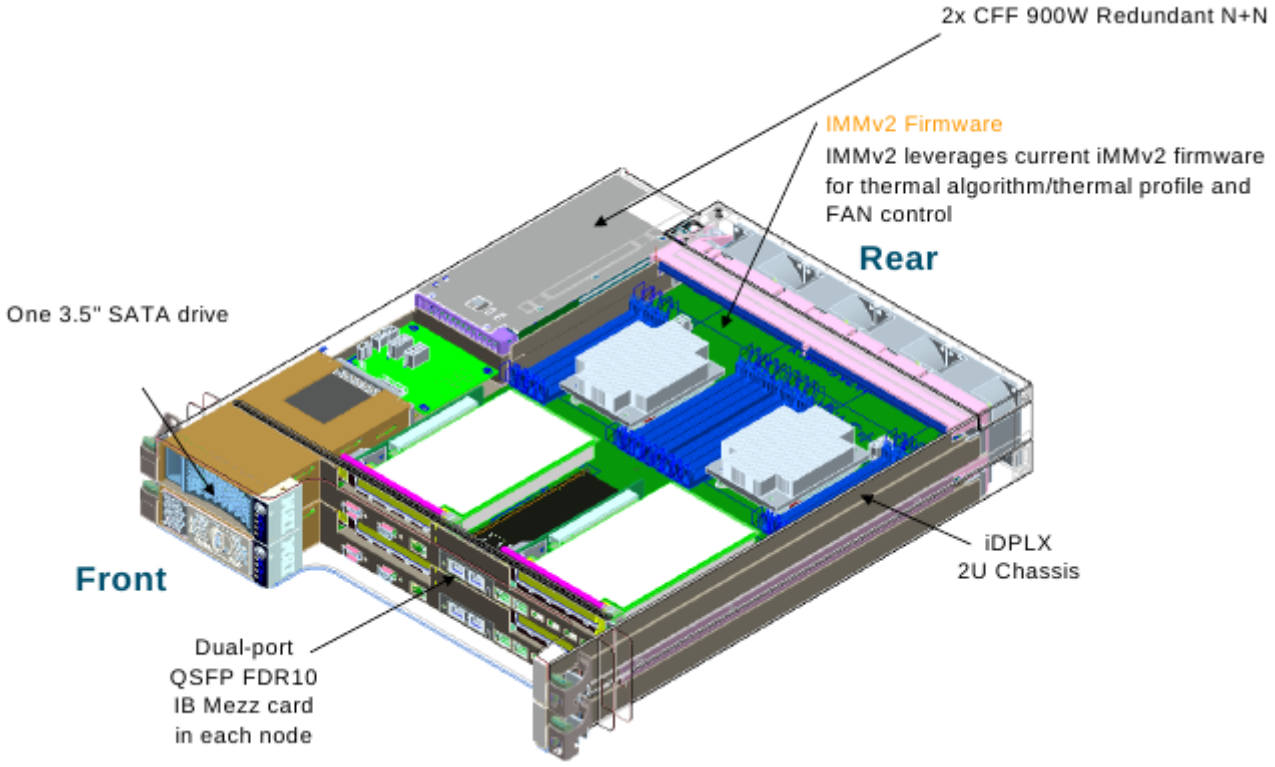
Front of Chassis

Compute Node

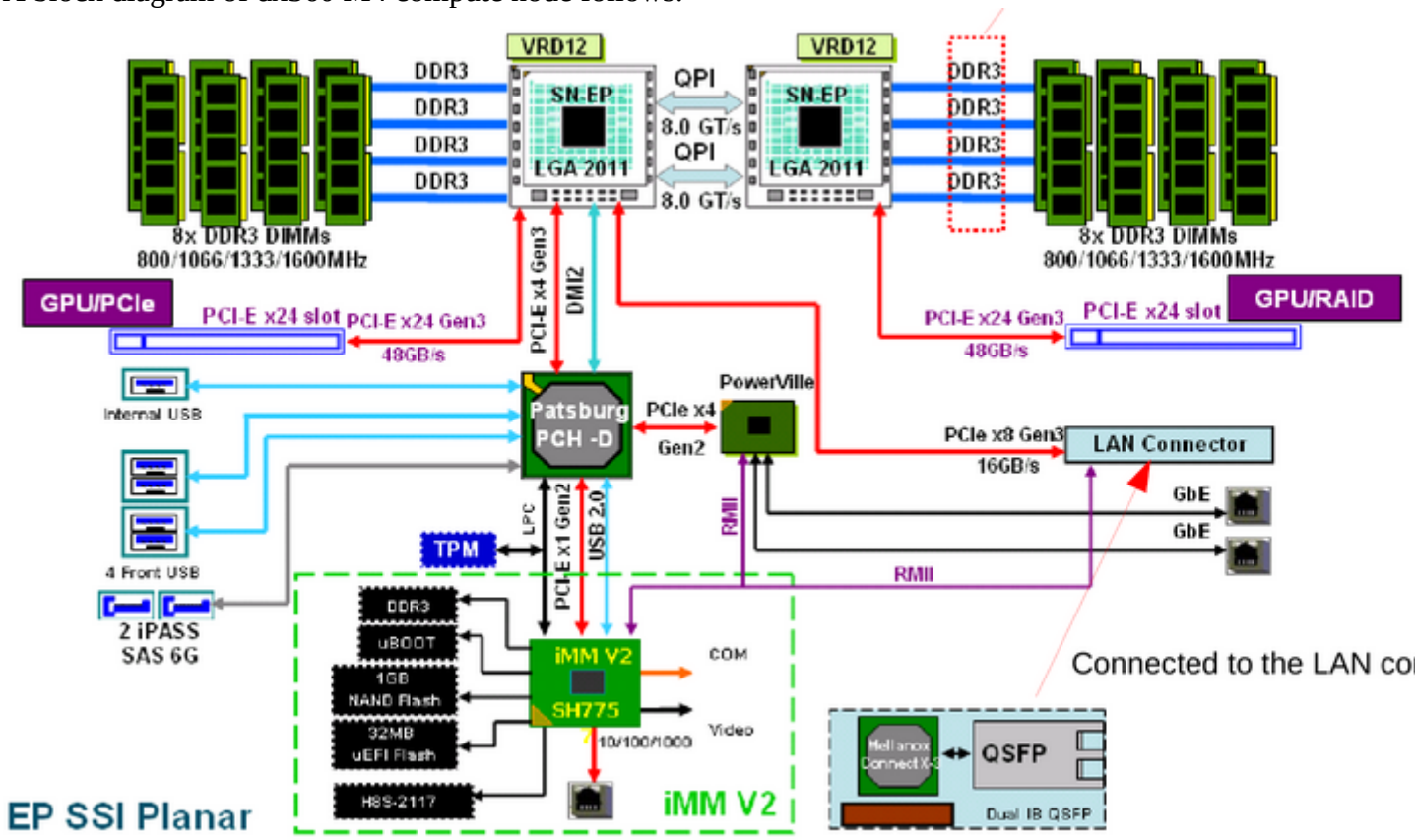
The compute nodes were the last generation of IBM System X servers: iDataPlex dx360 M4. These nodes are based on Intel Xeon (R) technology, and they offer high performance, flexibility and power efficiency.

Each compute node were composed of:

- Two 8-core Intel Xeon processors E5-2670 at 2.6 GHz, 20 MB cache memory, with a peak performance of 332.8 Gflops per node.
- Eight 4 GB DIMM's, 1.5V DDR3 @ 1600 MHz. Having 32 GB per node and 2 GB per core. In terms of memory bandwidth, this is the best configuration for memory access for Intel E5-26xx EP processors (4 memory channels per chip).
- Local hard drive: IBM 500 GB 7.2K 6Gbps NL SATA 3.5.
- MPI network card: Mellanox ConnectX-3 Dual Port QDR/FDR10 Mezz Card.
- 2 Gigabit Ethernet network cards (management network and GPFS).



A block diagram of dx360 M4 compute node follows:



Infiniband Racks

The 3,028 compute nodes were interconnected through a high speed interconnection network: Infiniband FDR10. The different nodes were interconnected via fibre optic cables and Mellanox 648-port FDR10 Infiniband Core Switches.

Four of the racks in MareNostrum were dedicated to network elements which allowed to interconnect the different nodes connected to the Infiniband network.

The main features for the Mellanox 648-port Infiniband Core Switch were:

- Hardware-based routing
- Congestion control
- Quality of Service enforcement
- Temperature sensors and voltage monitors
- Port speed auto-negotiation
- Full bisectional bandwidth to all ports
- Hot-swappable fan trays
- Port and system status LED indicators
- RoHS-5 compliant

Barcelona Supercomputing Center - Centro Nacional de Supercomputación

---