MareNostrum 4 (2017) System Architecture

MareNostrum is a supercomputer based on Intel Xeon Platinum processors, Lenovo SD530 Compute Racks, a Linux Operating System and an Intel Omni-Path interconnection.

See below a summary of the general purpose cluster system:

- Peak Performance of 11.15 Petaflops
- 384.75 TB of main memory
- 3,456 nodes:
  - 2x Intel Xeon Platinum 8160 24C at 2.1 GHz
  - 216 nodes with 12x32 GB DDR4-2667 DIMMS (8GB/core)
  - 3240 nodes with 12x8 GB DDR4-2667 DIMMS (2GB/core)
- Interconnection networks:
  - 100Gb Intel Omni-Path Full-Fat Tree
  - 10Gb Ethernet
- Operating System: SUSE Linux Enterprise Server 12 SP2

MareNostrum 4 user documentation

The current usage of the General purpose machine, is shown in the following graph, where the total cores available, the cores idle and the allocated ones are shown in red, yellow and green respectively.

Compute racks

MareNostrum 4 has 48 racks dedicated to calculations. These racks have a total of 165,888 Intel Xeon Platinum cores with a frequency of 2.1 GHz and 384.75 TB of total memory.

In total, each rack has 3,456 cores and 6,912 GB of memory.

The peak performance per rack is 226.80 Tflops, and a peak power consumption of 33.7 kW.

Each Lenovo SD530 compute rack is composed of:

- 72 Lenovo Stark compute nodes
- 2 Lenovo G8272 Switches
- 3 Intel OPA 48 port edge switch
- 4 32A 3 phase switched and monitored PDUs

Compute node

The compute nodes are based on the last generation Intel Xeon Platinum technology, and they offer high performance, flexibility and power efficiency. You can see below a description of one node:

Intel Omni-Path rack

The 3,456 compute nodes are interconnected through a high speed interconnection network: Intel Omni-Path (OPA). The different nodes are interconnected via fibre optic cables and Intel Omni-Path Director Class Switches 100 Series.

Six of the racks in MareNostrum are dedicated to network elements which allow to interconnect the different nodes connected to the OPA network.

The main features for the Omni-Path Director Class Switch are:

- Up to 786 x 100GbE ports in 20U (+1U Shelf)
- 12 x hot swap PSUs (N+N)
- Hot swap fan modules
- 2 x Management modules
- 8 x Double spine modules (non-blocking)
- Up to 24 x 32 port leaf modules (20 occupied – 640 ports)
- Each leaf modules contain 2 ASICs
- 9.4kW power consumption

CTE-POWER System Architecture

CTE-POWER is a cluster based on IBM Power9 processors, with a Linux Operating System and an Infiniband interconnection. Its main characteristic is the availability of 4 GPUs per node, making it an ideal cluster for GPU accelerated applications.

It has the following configuration:

- 2 login node and 52 compute nodes, each of them:
  - 2 x IBM Power9 8335-GTH @ 2.4GHz (3.0GHz on turbo, 20 cores and 4 threads/core, total 160 threads per node)
  - 512GB of main memory distributed in 16 dimms x 32GB @ 2666MHz
  - 2 x SSD 1.9TB as local storage
  - 2 x 3.2TB NVME
  - 4 x GPU NVIDIA V100 (Volta) with 16GB HBM2.
  - Single Port Mellanox EDR
  - GPFS via one fiber link 10 GBit

The operating system is Red Hat Enterprise Linux Server 7.5 (Maipo).

Barcelona Supercomputing Center - Centro Nacional de Supercomputación

Source URL (retrieved on 27 Jul 2020 - 18:35):
https://www.bsc.es/marenostrum/marenostrum/technical-information