MareNostrum 5

MareNostrum 5 is a pre-exascale EuroHPC supercomputer hosted at BSC-CNS. The system is supplied by Bull SAS combining Bull Sequana XH3000 and Lenovo ThinkSystem architectures and it has a total peak computational power of 314PFlops. The system will provide 4 partitions with different technical characteristics that jointly can fulfill the requirements of any HPC user.

The MareNostrum 5 partitions are divided as follows :

1. General Purpose partition based on Intel Sapphire Rapids, 6408 nodes + 72 HBM nodes, 45 PFlops peak
2. Accelerated Partition based on Intel Sapphire Rapids and Nvidia Hopper GPUs, 1120 nodes with 4 Hopper GPUs each one, 230 PFlops peak
3. General Purpose - Next Generation Partition, , based on Nvidia GRACE CPU
4. Accelerated - Next Generation Partition , not fully defined. more information in the following months.

System Overview

The 2 main partitions of MareNostrum 5 have these technical characteristics :

MareNostrum 5 GPP (General Purpose Partition)

The machine has 6408 nodes based in Intel Sapphire rapids. Each is configures as:

- 2x Intel Sapphire Rapids 8480+ at 2Ghz and 56c each (112 cores per node)
- 256 GB of Main memory, using DDR5 (with 216 nodes with 1024GB)
- 960GB on NVMe storage (/scratch)
- 1x NDR200 shared by 2 nodes (SharedIO) (BW per node 100Gb/s)

In addition to the 6408 standard nodes, the machine has 72 HBM nodes based in Intel Sapphire Rapids 03H-LC with 112 cores per node at 1.7Ghz and 128GB HBM memory. This small sub-system will provide a high memory BW of 2TB/s per node.

The full machine provides a Peak Performance of 45.9 PFlops.

The network topology used is fat-tree , with islands with full fat tree without contention of 2160 nodes, and with a contention between islands of 2/3.

MareNostrum 5 ACC (Accelerated Partition)

The machine has 1120 nodes based in Intel Sapphire rapids and Nvidia Hopper GPUs. Each is configures as:

- 2x Intel Sapphire Rapids 8460Y+ at 2.3Ghz and 32c each (64 cores node)
- 512 GB of Main memory, using DDR5
- 4x Nvidia Hopper GPUs with 64 HBM2 memory
- 460GB on NVMe storage (/scratch)
- 4x NDR200 (BW per node 800Gb/s)

The full machine provides a Peak Performance of 260 PFlops

The network topology used is fat-tree , with islands with full fat tree without contention of 160 nodes, and with a contention between islands of 1/2.

Storage system and long term Archive

MareNostrum provides a top class storage system of 248PB net capacity based on SSD/Flash and hard disks, with an aggregated performance of 1.2TB/s on writes and 1.6TB/s on reads. Long-term archive storage solution based on tapes will provide 402PB additional capacity. IBM Storage Scale and Archive will be used as parallel filesystem and tearing solution respectively.

Software available for both partitions

- Red Hat Enterprise Server
- Intel OneAPI
    - C/C++/Fortran Compilers
    - MKL
    - Intel MPI
    - Intel Trace Analyzer and Collector
- DDT parallel debugger
- BSC performance tools
- EAR: Energy management framework for HPC
- NVIDIA HPC SDK
- NVIDIA CUDA Toolkit
- OpenMPI

- Slurm batch scheduler
