ENEA CRESCO HPC clusters

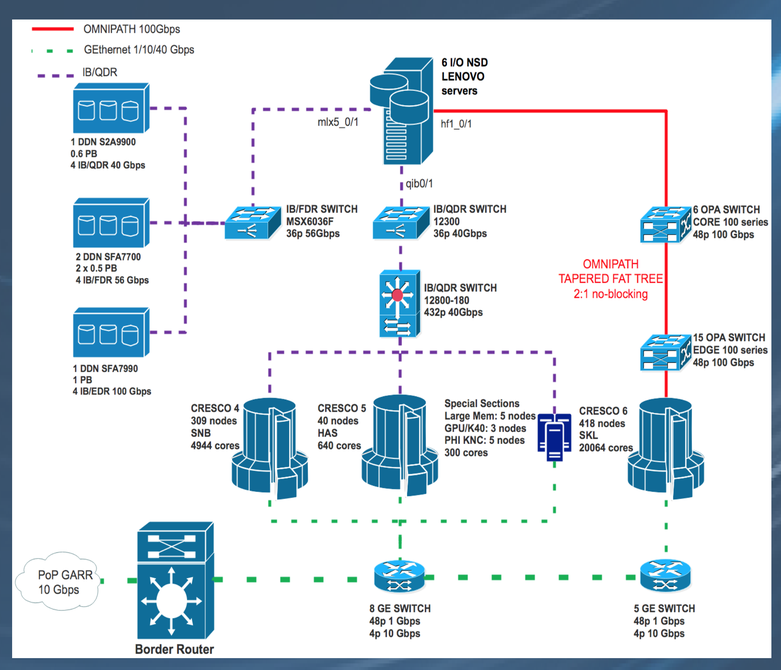
A working example of a multifabric GPFS Spectrum Scale layout

F.Iannone[1*], F. Ambrosino[1], G. Bracco[1], M. De Rosa[1], A. Funel[1], G. Guarnieri[1], S. Migliori[1], F. Palombi[1], P. Procacci[2] , G. Ponti[1], G. Santomauro[1]

[1] ENEA - Energy Technologies Department – ICT Division – HPC laboratory – Lungotevere Thaon de Ravel, Rome, Italy
[2] Dipartimento di Chimica, Universita di Firenze, Via della Lastruccia 3, I-50019 Sesto Fiorentino, Italy
[*] corresponding author: francesco.iannone@enea.it

PREMISE --- ENEA is the Italian National Agency for New Technologies, Energy and Sustainable Economic Development. We operate in a range of R&D sectors, including energy technologies, new materials, life and environmental sciences. In support of the institutional research activities, the ENEA IT division provides computing and storage resources, integrated into ENEAGRID, a computational infrastructure distributed over six geographical sites in Italy. The computing core of ENEAGRID is represented by the HPC CRESCO clusters.

In May 2018 a new PC cluster, **CRESCO6**, was inaugurated at ENEA Portici. CRESCO6 is a 1.4Pflop/s machine. It ranked 420th in the TOP500 list of Nov. 2018. CRESCO6 succeeded CRESCO4 and CRESCO5, two older HPC clusters already installed in the same data center (they are still in service), with a nominal computing power of 0.1 and 0.025 Pflop/s respectively.



CRESCO6 TECHNICAL DETAILS --- CRESCO6 is a high performance computing system consisting of 434 nodes for a total of 20,064 cores. It is based on the Lenovo ThinkSystem SD530 platform, an ultra-dense and economical two-socket server in a 1/2U rack form factor inserted into a 2U four-mode enclosure. Each node is equipped with:

- 2 Intel **Xeon Platinum 8160** CPUs, each hosting 24 cores with a clock frequency of 2.1GHz
- 192 GB RAM memory, corresponding to 4GB/core
- one low-latency **Intel Omni-Path 100** Series Single-port PCIe 3.0 x16 HFA network interface

Nodes are interconnected via an Intel Omni-Path network with 21 Intel Edge switches 100 series of 48 ports each, bandwidth equal to 100GB/s, latency equal to 100ns. Connections between the nodes have 2 tier 2:1 no-blocking tapered fat-tree topology. The consumption of electrical power during massive computing workloads amounts to 190kW.

DATA CENTER AT ENEA PORTICI --- The main data center is located at ENEA Portici, near Naples. The facility consists of two large rooms having an overall surface of about 90m^2. In each room an air cooled chiller with free-cooling technology has been installed. The rooms have been arranged for water free-cooling. Actually three HPC clusters are in operations:

Cluster name (Linux x86_64)	Network	Cores/Tflops
CRESCO4 INTEL/SNB	IB QDR	4944/~100
CRESCO5 INTEL/HAS	IB QDR	640/~25
CRESCO6 INTEL/SKL	OPA	10368/~700
Total		**~16624/~850**
CRESCO6+ INTEL/SKL	**OPA**	**10368/~700**

Storage Systems	Bandwidth GB/s	network	Capacity
1 x DDN S2A9900	4	IB QDR	0.6 PB
2 x DDN SFA7700	15	IB QDR	1 PB
2 x Dot-Hill 3730	8	FC/8Gpbs	0.5 PB
3 x RAID Servers	1.25	10 GEth	72 TB
1 x Dot-Hill 3000	1.25	FC/8Gbps	12 TB
1 x DS 4700	0.8	FC/4bps	26 TB
Total			**~2.2 PB**
1 x DDN SFA7790	**20**	**EDR**	**1 PB**

MULTIFABRIC GPFS LAYOUT --- The main challenge in deploying CRESCO6 in the ENEA data center at Portici, was to design and implement a layout such that the old Infiniband QDR (40Gbps) fabric serving CRESCO4 and CRESCO5, including the high performance DDN storage equipped with Infiniband FDR (56Gbps), the new Omni-Path (100Gbps) fabric serving CRESCO6 and the high performance Spectrum Scale (GPFS) filesystem work together within a single infrastructure. The main hardware components of the layout are as follow:

- High performance storage systems based on DDN Infiniband QDR/FDR/EDR
- 6 NSD servers for GPFS based on 2U Lenovo ThinkSystem SR650, 96GB RAM, 2 Intel Xeon Gold 5518 CPU @2.3 GHz 12 cores, 1 QLogic IBA7322 QDR Infiniband HCA (rev 02), 1 Mellanox MCX454A-FCAT DUAL FDR Infiniband HCA, 1 Intel Omni-Path HFI Silicon 100 Series.
- 1U switch Mellanox MSX6036F - 36 ports FDR Infiniband for linking the NSD servers to the storage DDN systems
- 1U switch Qlogic 12300 – 36 ports QDR Infiniband for linking the NSD servers to the CRESCO4 and CRESCO5 clusters.

The software stack includes the operating system (OS), the Infiniband and Omni-Path fabric plus the RDMA, SRP, Multipath packages and the Spectrum Scale GPFS.

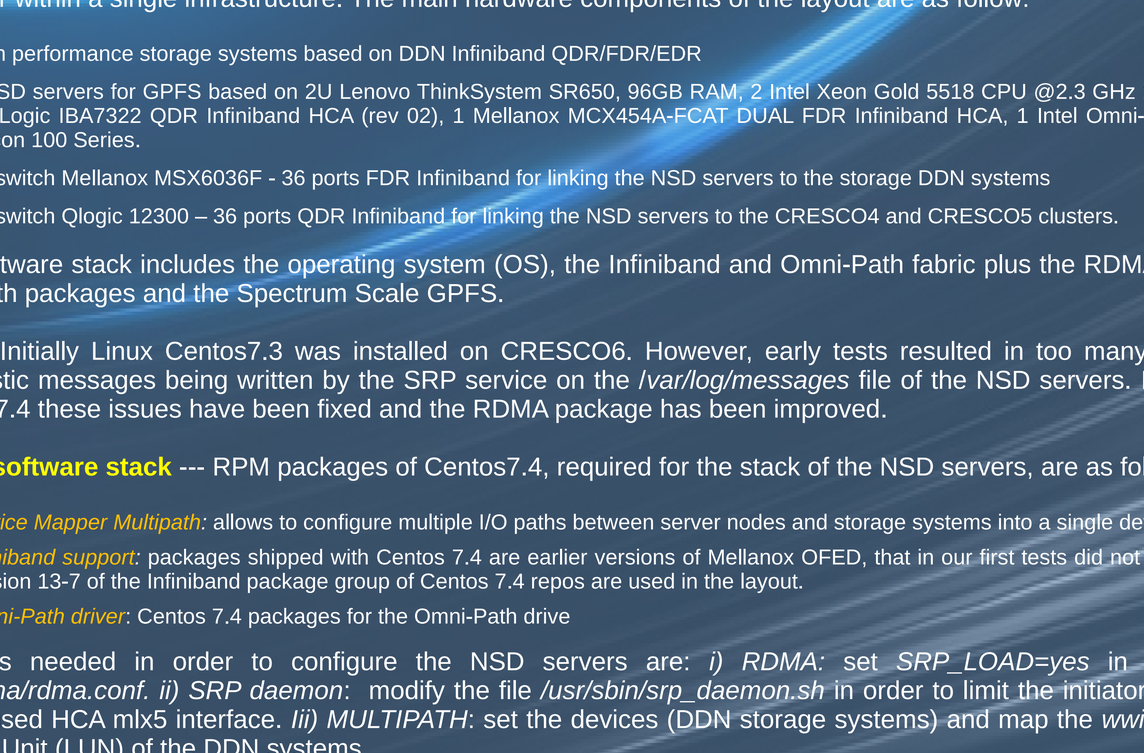
OS --- Initially Linux Centos7.3 was installed on CRESCO6. However, early tests resulted in too many critical diagnostic messages being written by the SRP service on the */var/log/messages* file of the NSD servers. In Linux Centos7.4 these issues have been fixed and the RDMA package has been improved.

RPMs software stack --- RPM packages of Centos7.4, required for the stack of the NSD servers, are as follow:

- *Device Mapper Multipath:* allows to configure multiple I/O paths between server nodes and storage systems into a single device.
- *Infiniband support:* packages shipped with Centos 7.4 are earlier versions of Mellanox OFED, that in our first tests did not work. So version 13-7 of the Infiniband package group of Centos 7.4 repos are used in the layout.
- *Omni-Path driver*: Centos 7.4 packages for the Omni-Path drive

Services needed in order to configure the NSD servers are: *i) RDMA:* set *SRP_LOAD=yes* in the file */etc/rdma/rdma.conf. ii) SRP daemon*: modify the file */usr/sbin/srp_daemon.sh* in order to limit the initiator search to the used HCA mlx5 interface. *Iii) MULTIPATH*: set the devices (DDN storage systems) and map the *wwid* of the Logical Unit (LUN) of the DDN systems.

Spectrum Scale GPFS --- The high performance filesystem installed on all CRESCO clusters is the Spectrum Scale (GPFS), version 4.2.3-2. The six NSD servers have to provide RDMA services on both fabric: Infiniband for the GPFS client nodes of CRESCO4/5 and Omni-Path for the GPFS client nodes of CRESCO6. To achieve this, a list of RDMA enabled devices, ports, and subnets has been defined in the GPFS configuration for the whole set of nodes. In a single fabric configuration, such as IB/QDR with Truescale, GPFS has no *nodeclass* definition and the default value of the variable *verbsPorts* is *qib0/1/1*. This means that only the GPFS client nodes with QDR HCA *qib0* can use RDMA to access NSD disks. To set the multifabric layout, two classes have been defined with specific properties.



The I/O performance of the Spectrum Scale in multifabric layout shows transfer rates in R/W mode on single process (*lmdd*) of ~1.5 GB/s on single process for the IB/QDR client nodes and ~2.5 GB/s for the Omni-Path client nodes.

BENCHMARKS --- CRESCO6 was deployed in two stages. In early 2018 a PC cluster was installed, consisting of 216 Skylake nodes and Omni-Path network topology based on 1:1 no-blocking tapered fat-tree. At the end of 2018, 218 nodes were added, for a total of 434 Skylake nodes with Omni-Path network topology based on 2:1 no-blocking tapered fat-tree. Several tests to assess the performance of Omni-Path, including OSU and IMB, were performed during the first stage. Latency and bandwith tests resulted in very similar performances on both 1:1 and 2:1 topologies. We report below a list of benchmarks performed on the machine.

MOLECULAR DYNAMICS SIMULATIONS --- Extensive tests, based on the *ORAC6* molecular dynamics (MD) code*, were performed during the first stage of deployment of CRESCO6, to compare our cluster with MARCONI A1 (Broadwell CPU) at CINECA, . The *ORAC6* program is a hybrid OpenMP/MPI code, specifically designed for running simulations of complex systems on multicores NUMA architectures. The code is written in FORTRAN and can be freely downloaded from the website *http://www.chim.uni.it/orac*. Parallelism of the MD computation is implemented on two layers, based on *i)* an intra-node strong scaling algorithm for force decomposition on shared memory via OpenMP; *ii)* an intra- and inter-node weak scaling parallelization for replica exchange and concurrent simulation of driven non equilibrium (NE) trajectories implemented via MPI. The total number of cores requested by a typical orac job is hence given by $N_{cores}=N_{MPI} \times N_{treadhs}$. The benchmarks were obtained in two case studies, based on MPI/OpenMP hybrid codes, modelling fast switching double annihilation** computations, namely:

- *REM*: A replica exchange simulation using 8 threads on the OpenMP layer and 96 MPI instances for a total of 768 cores on CRESCO6, and 9 threads on the OpenMP layer and 64 MPI instances for a total of 576 cores on Marconi/A1. The system in both cases is the *s45a* mutant of streptavidin in complex with biotin (protein data bank code 1df8) for a total of 12830 atoms.
- *FNE*: A non equilibrium alchemical annihilation for 720 concurrent trajectories on the MPI layer each running on 8 OpenMP threads for a total of 5760 cores on CRESCO6 and for 810 MPI trajectories each with 6 OpenMP threads on Marconi/A1 for a total of 4860 cores. The system in both cases is the structure of ricin *A* chain bound with $C_{17}H_{17}N_7O_4$ (protein data bank code 4mx5) for a total of 25280 atoms.

	REM stage: code 1df8 – 12830 atoms					
platform	$N_{threads}$	N_{MPI}	N_{cores}	τ_{sim}	τ_{tot}	*Elapsed/hours*
CRESCO6	8	96	768	3.6	345.6	**23.8**
MARCONI/A1	9	64	576	3.6	230.4	**19.2**
	FNE stage: code 4mx5 – 25280 atoms					
platform	$N_{threads}$	N_{MPI}	N_{cores}	τ_{sim}	τ_{tot}	*Elapsed/hours*
CRESCO6	8	720	5760	0.72	518.6	**7.2**
MARCONI/A1	6	810	4860	0.72	583.4	**11.4**

* P. Procacci, J. Chem. Inf. Model, 56(6):1117-1121, 2016

** P. Procacci,Phys. Chem. Chem. Phys., 18:14991-15004, 2016

N-BODY SIMULATIONS --- A Particle-Particle (PP) method based on a brute-force approach was used to simulate the *n-body* problem. The *n-body* simulation integrates Newton's equation with gravitational interactions numerically. A case study of the *N-body* simulation was designed to compare a hybrid MPI+OpenMP algorithm running on CRESCO6 and on various CPU platforms at CINECA, i.e. MARCONI *A1* with Broadwell, MARCONI *A2* with Knigths Landing and MARCONI *A3* with *Skylake*. A hybrid *master-only* model was used with one MPI rank per node and OpenMP threads scattered on the cores of the node, no MPI calls inside MP parallel regions. Since the *n-body* PP method has a complexity of $O(n^2)$, tests were performed by measuring the execution time per time-step with two choices of the number of particles, namely 10^6 and 10^7. Tests were made on *8*, *16* and *32* nodes on MARCONI A1/A2/A3 and up to *128* nodes on CRESCO6. Results show that the algorithm scales very well with the number of nodes. Moreover, the performance on the Skylake nodes is better then on the other CPUs.