CRESCO ENEA 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], G.Ponti[1], G.Santomauro[1]
[1] ENEA- Energy Technologies Department – ICT Division – HPC laboratory – Lungotevere Thaon de Ravel, Rome, Italy

P.Procacci[2]
[2] Dipartimento di Chimica, Universita di Firenze, Via della Lastruccia 3, I - 50019 Sesto Fiorentino, Italy

EXTENDED ABSTRACT

Abstract — ENEA is the Italian National Agency for New Technologies, Energy and Sustainable Economic Development. ENEA operates in many sectors among which the most important are: energy technologies, materials physics, life sciences and climate. In the framework of its institutional mission, the ICT Division provides computing and storage resources integrated into ENEAGRID/CRESCO, an infrastructure distributed over 6 sites, whose main facilities are the HPC CRESCO clusters. The bulk of all storage is based on IBM Spectrum Scale (GPFS) since many years. The access to data, even over WAN, is managed by GPFS clusters. In May 2018 the new cluster CRESCO6 was inaugurated. CRESCO6, a 1.4 Pflops based on Intel Xeon X86_64 SkyLake CPU ranked at 420th of TOP 500 Nov.2018 list. While the interconnection of CRESCO6 is based on Intel Omni-Path (OPA) (100 Gbps), the previous CRESCO4 and CRESCO5 clusters have a network based on InfniBand QDR Truescale fabric (40 Gbps). Hence, in order to provides storage to all CRESCO clusters a GPFS multifabric layout has been implemented after dedicated tests. The work describes the ENEAGRID/CRESCO infrastructure and in particular the solution adopted to implement the GPFS multifabric.

Keywords—HPC, Omni-Path, Infiniband, GPFS

I. INTRODUCTION

A new cluster named CRESCO6 has been installed at the ENEA Research Center of Portici since Oct. 2018. The new supercomputer refreshes the complex computational infrastructure of ENEA, granting a performance increase of a factor 10 compared with so-far available resources. At beginning the storage infrastructure of the CRESCO data center [1], mainly composed of DDN raid systems, was based on Infiniband Qlogic Truescale in order to provide the high performance filesystem GPFS also to the HPC clusters CRESCO4/5, they too on the same fabric. When the cluster CRESCO6 was designed, a new fabric technology, named Omni-Path, was available by Intel, so it arose the problem to integrate different fabrics in order to access the same storage infrastructure. This paper shows the technical specifications and benchmarks of the new cluster CRESCO6 including the technical details of how to implement a multi-fabric layout able for working Infiniband and Omni-Path on a single GPFS cluster using the same storage systems.

II. CRESCO6 TECHNICAL SPECIFICATION

CRESCO6 is a high performance computing system (HPC) consisting of 434 nodes for a total of 20832 cores. It is based on Lenovo ThinkSystem SD530 platform, an ultra-dense and economical two-socket server in a 0.5 U rack form factor inserted in a 2U four-mode enclosure. Each node is equipped with 2 Intel Xeon Platinum 8160 CPUs, RAM of 192 GB and a low-latency Intel Omni-Path 100 Series Single-port PCIe 3.0 x16 HFA network interface. The nodes are interconnected by an Intel Omni-Path network with 21 Intel Edge switches 100 series of 48 ports each, bandwidth equal to 100 GB/s, latency equal to 100ns. The connections between the nodes have 2 tier 2:1 no-blocking tapered fat-tree topology. Actually three HPC clusters are in operations: CRESCO4: 100 Tflops peak rate, CRESCO5: 25 Tflops peak rate, both with Infiniband QDR low latency network, and CRESCO6: 1.4 Pflops peak rate with Omni-Path. A special nodes section with large memory and accelerator cpus (Intel Xeon PHI KNC and Nvidia GPU Tesla K40) are available within the Infiniband QDR network. The storage systems are based on DDN S2A9900, and SFA7700/990 solutions with a network based on Infiniband. The high performance filesystem is based on GPFS Spectrun Scale with 6 NSD nodes able to provide parallel I/O at all.

III. BENCHMARKS

CRESCO6 has been deployed in two steps: the first one on early of 2018 with 700 Tflops peak perfomance, 216 Skylake nodes and a Omni-Path network topology based on 1:1 no-blocking tapered fat-tree; the second one it has doubled with 1.4 Pflops peak performance, 434 Skylake nodes and a Omni-Path network topology based on 2:1 no-blocking tapered fat-tree.

A. HPL benchmark

The HPL test has allowed CRESCO6 to ranks 420th in the TOP 500 Nov.2018 list. The benchmark was made with 434 nodes, 2 MPI tasks per node (one for each socket) showing better performance around 72% efficiency and up 1.012 Pflops. The Table I shows the main parameters of the HPL test with *N* as coefficients matrix size, *NB* as block size subdivision matrix, *TB* as matrix memory size in terabytes and $P \times Q$ is the size of the grid equal to the sockets number.

TABLE I. HPL TEST PARAMETERS

nodes	#MPI /# Thread	N / NB / TB / P×Q	Perf range
434	2 / 24 for node	300979 /384/73/28×31	1.012 Pflops

B. Molecular Dynamic simulations

An extensive test performance during the first step of deployment of CRESCO6 was made using the *ORAC6* molecular dynamics (MD) code [2], comparing with MARCONI A1 Broadwell CPU platform at CINECA. The *ORAC6* program is a hybrid OpenMP/MPI code specifically designed for running simulation of complex systems on multicores NUMA architectures. The parallelism of the MD computation is done on two layers, based on *i)* an intranode strong scaling algorithm for force decomposition implemented on a shared memory environment via OpenMP threading; *ii)* an intra and internode 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 obtained by two case

studies based on MPI/OpenMP hybrid codes modelling fast switching double annihilation [3] computations, namely: *REM*: a replica exchange simulation and *FNE*: A non equilibrium alchemical annihilation. Most of the parallelization in these production jobs is invested on the weak scaling (MPI) part (with basically ideal parallel efficiency) with few $N_{threads}$ OpenMP instances (from 6 to 9) running in the intranode shared memory environment. As shown in Table II, such an hybrid setup allows to easily produce a total simulation time of the order of the microseconds using from few hundreds (*REM*) to few thousands (*FNE*) of cores. The parallel setup for the two jobs reflects the different architecture of the CRESCO6 and MARCONI/A1 nodes with 48 and 36 cores, respectively. The data reported in the Table show that MARCONI is faster for the *REM* stage while it appears to be slower for the *FNE* computation. The reason for this outcome lies in the fact that, on CRESCO6, the *FNE* force routine involving the alchemical calculation, is fully vectorized with *AVX512* feature of the Skylake CPU while vectorization is incomplete for the more complex force *REM* routine.

TABLE II. MOLECULAR DYNAMIC BENCHMARKS: ORAC6 HYBRID MPI/OPENMPI CODE

	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**

τ_{sim} and τ_{tot} indicate, the simulation time in nanoseconds.

In conclusion by stating that a careful code optimization/vectorization (diagnosed using the *-qopt-report-phase=vec* Intel compiler option) is essential for best performances on the CRESCO6 cluster. For example, for the fully vectorized FNE stage, based on the data given in Table II, after threads re-normalization due to the architecture dependent hybrid parallelization, the CRESCO6 architecture appears to be about 30% faster than the MARCONI/A1 HPC.

IV. MULTIFABRIC GPFS LAYOUT

The main challenge, to deploy CRESCO6 HPC cluster into the CRESCO data center infrastructure, has been to design and implement a layout on which: the old Infiniband QDR (40 Gbps) fabric of the CRESCO 4/5 clusters, including high performance storages DDN equipped with Infiniband FDR (56 Gbps), the new Omni-Path (100 Gbps) fabric of CRESCO6 cluster and the high performance filesystem Spectrum Scale (GPFS), work all together into a single infrastructure. The main hardware components of the layout solution is as follow: the high performance storage systems based on DDN Infiniband QDR/FDR/EDR, 6 NSD servers for GPFS based on 2U Lenovo ThinkSystem SR650, 96 GB RAM, 2 Intel Xeon Gold 5518 CPU, 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 Infiniband FDR for linking the NSD servers to the storage DDN systems and finally 1U switch Qlogic 12300 Infiniband QDR for linking the NSD servers to the CRESCO 4/5 clusters. The main software stack including operating system: Linux Centos 7.4, Infiniband support ver. 13.7 shipped with Centos 7.4 and Omni-Path fabric 10.3.1.0-11, RDMA, SRP, Multipath packages: Device-mapper-multipath and Device-mapper-multipath-libs version: 0.49-111.3l7 and finally Spectrum Scale GPFS 4.2.3-2. The services to configure on 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 only 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. A last configuration set is on the storage side to define the *GUIDs* of the HCA interface of the all NSD servers into the DDN controllers. In our case the command: *ibstat mlx5 | grep GUID* find it for Mellanox HCA.

A. Spectrum Scale GPFS

The six NSD servers have to provide RDMA service on both fabric: Infiniband for the GPFS client nodes of CRESCO4/5 and Omni-Path for the GPFS client nodes of CRESCO6. To make that, a list of RDMA enabled devices, ports, and subnets must be defined into GPFS configuration for all the class 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*. It 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 to be defined with specific properties, for example: *nodeclass crescoOPA* with *verbsPorts hfi1_0/1/2* for all client nodes of CRESCO6 and *nodeclass crescoQDROPA* with both *verbsPorts qib0/1/1* and *hfi1_0/1/2* for the six NSD servers. The GPFS commands are: *mmcrnodeclass* and *mmchconfig verbsPorts*. These definitions can be made any time but they become active only when GPFS is restarted on the nodes. The I/O performance of the Spectrum Scale in multifabric layout shows transfer rates in R/W of ~1.5 GB/s for the IB/QDR client nodes and ~2.5 GB/s for Omni-Path client nodes.

V. CONCLUSIONS

The new HPC facility, CRESCO6, is in operation at ENEA data center. The high performance storage systems have been integrated in a multifabric layout mixing different technologies: Qlogic/TrueScale, Mellanox and Intel Omni-Path on which the high performance filesystem Spectrum Scale (GPFS) works with good I/O throughput.

Acknowledgment

The computing resources and the related technical support used for this work have been provided by CRESCO/ ENEAGRID High Performance Computing infrastructure and its staff [4].

REFERENCES

[1] S. Migliori, G. Bracco, et al. Architecture and performances of the CRESCO HPC system. International Supercomputing Conference, Dresden 17-19/6/2008.

[2] Piero Procacci. Hybrid MPI/OpenMP Implementation of the ORAC Molecular Dynamics Program for Generalized Ensemble and Fast Switching Alchemical Simulations. J. Chem. Inf. Model., 56(6):{1117,1121}, 2016.

[3] Piero Procacci. I. dissociation free energies of drug-receptor systems via non-equilibrium alchemical simulations: a theoretical framework. Phys. Chem. Chem. Phys., 18:{14991,15004}, 2016.

[4] G. Ponti et al., "The role of medium size facilities in the HPCecosystem: the case of the new CRESCO4 cluster integrated in the ENEAGRID infrastructure", Proceedings of the 2014.

[5] International Conference on High Performance Computing and Simulation, HPCS 2014, art.no. 6903807, 1030-1033.