Journal of large-scale research facilities

Journal of large-scale research facilities, 8, A186 (2024) http://dx.doi.org/https://doi.org/10.17815/jlsrf-8-186

Published: 15.01.2024

LEONARDO: A Pan-European Pre-Exascale Supercomputer for HPC and AI applications

CINECA Supercomputing Centre,
SuperComputing Applications and Innovation (SCAI) Department*

Instrument Scientists:
- CINECA, SCAI Department, Via Magnanelli 6/3, Casalecchio di Reno 40033, Bologna, Italy
  phone + 39 051 6171411, email: superc@cineca.it
- CINECA, SCAI Department, Via dei Tizii 6, Roma 00185, Italy
  phone +39 06 444861, email: superc@cineca.it

Abstract: A new pre-exascale computer cluster has been designed to foster scientific progress and competitive innovation across European research systems, it is called LEONARDO. This paper describes the general architecture of the system and focuses on the technologies adopted for its GPU-accelerated partition. High density processing elements, fast data movement capabilities and mature software stack collections allow the machine to run intensive workloads in a flexible and scalable way. Scientific applications from traditional High Performance Computing (HPC) as well as emerging Artificial Intelligence (AI) domains can benefit from this large apparatus in terms of time and energy to solution.

*Cite article as: CINECA Supercomputing Centre, SuperComputing Applications and Innovation Department. (2024). LEONARDO: A Pan-European Pre-Exascale Supercomputer for HPC and AI applications. *Journal of large-scale research facilities, 8*, A186. http://dx.doi.org/https://doi.org/10.17815/jlsrf-8-186

CC BY

1 Introduction

LEONARDO is a new European computer cluster with pre-exascale computing capabiliy, at the level of 0.2×10^{18} floating point operations per second (FLOPS). The project has been conceived by LEONARDO *Consortium*, a group of six signatory countries[1] of the European declaration on High Performance Computing (Declaration, 2018) whose purpose is to foster scientific and technological federative innovation across the European Union. LEONARDO is owned by the European High Performance Computing Joint Undertaking initiative (EuroHPC JU, 2018) and is hosted by CINECA interuniversity consortium (CINECA, 2023) at the Tecnopolo Manifattura Data Valley Hub in Bologna, Italy (Tecnopolo, 2023).

The foreseen operational lifetime of the machine is 5 years. In this period it is going to serve as a research facility for a broad class of scientific investigations, due to a complete set of state-of-the art hardware and software technologies that are presented in this paper. The most relevant are a massive amount of computational power available at single node (i.e. a peak performance of 78 teraFLOPS), a fast access storage (over a TB/s bandwidth) and a flexible scalability for multi-node computations. With LEONARDO, researchers from academia and industry can tackle many challenges in different crucial fields, like Digital Twins applications, e.g. DTGEO (2023), Data-driven projects, e.g. GEOIN (2023) and Urgent-Computing, e.g. CHEESE2 (2023) to name a few.

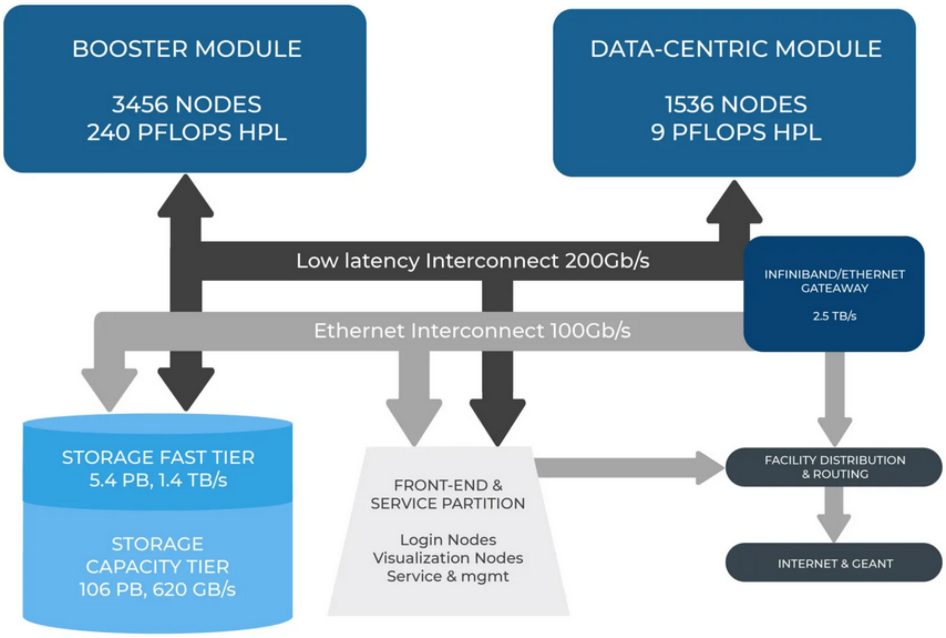

Figure 1: Architectural overview

The supercomputer has been designed by Eviden company with its technology partners and is composed by two compute partitions that are coupled with interconnects, storage and service subsystems. A general purpose Data-Centric (DC) partition is intended to fulfill a vast range of traditional HPC applications by leveraging the latest central processing unit (CPU) technologies. It measures 1536 compute nodes, based on the Intel's 4th generation of Xeon Scalable processor, codenamed *SapphireRapids*. The CPU model is the 56-core 8480+ that features several hardware accelerators to support Single Instruction Multiple Data extension (SIMD) on top of x86 instruction set; accelerated functionalities include cryptography and vector algebra (Intel, 2023). The other compute partition is a heterogeneous module called Booster which is dedicated to applications that can benefit from the parallelism offered by general purpose graphical processing unit (GPU). The Booster consists in 3456 nodes configured with single-

[1]The countries are: Italy (project leader), Austria, Greece, Hungary, Slovakia and Slovenia.

socket host Intel *Ice Lake* CPU (Intel, 2019) and four NVIDIA A100 *Tensor Core* GPU chips (NVIDIA, 2020). The internal network, used for inter-node communication, relies on 200 Gbps Mellanox's InfiniBand High Data Rate (HDR) technologies (NVIDIA, 2020c) and is organized in a *dragonfly+* topology. The storage is composed of a mix of high-speed and high-capacity appliances to accommodate the requirements of modern Big Data and AI applications, including Cloud services and Interactive computing. The infrastructure is completed by an operational 100 Gbps Ethernet network, 11 management nodes and 32 frontend servers where users can land, develop codes, submit jobs, and analyze results. Figure 1 presents a schematic overview of LEONARDO. All subsystems are shared between the two compute partitions. A set of four Ethernet/InfiniBand gateways allows the cluster to be connected to external networks.
This paper describes the overall architecture of LEONARDO and focuses on Booster's technologies. Section 2.1 presents the Booster's node, including computing elements and organization. Details on interconnect and storage subsystems can be found in 2.2 and 2.3. This is followed by paragraph 2.4 on frontend and service resources. Software tools and libraries are listed in 2.5. Finally the power supply and the cooling systems are presented in 2.6. Additionally, some benchmark results are reported in Appendix A and the list of hardware components can be found in Appendix B. The DC module will be detailed in a separate article.

2 System details

LEONARDO is a quite large apparatus consisting of 155 racks, 2 tons each. The compute partitions are made up of 138 racks based on the Eviden BullSequana XH2000 cabinet, a platform that offers high level integration density and Direct Liquid Cooling capabilities (Eviden, 2020). Table 1 shows how compute racks are organized in cells and how blade servers and node units compose each rack. One cell encompasses both Booster and DC type nodes and is called *Hybrid* cell. An additional cell (the twenty-third) houses storage and service equipment. This includes 12 racks equipped with DDN's appliances and 5 further Eviden racks dedicated to management and frontend servers.

Type	Cell	$\frac{Rack}{Cell}$	$\frac{Blade}{Rack}$	$\frac{Nodes}{Blade}$	Rack	CPU nodes	GPU nodes
Booster	19	6	30	1	114	-	3420
DC	2	8	26	3	16	1248	-
Hybrid	1	2	18	1	2	-	36
		6	16	3	6	288	-
Total	22	-	-	-	138	1536	3456

Table 1: Organization of the compute partitions

2.1 Booster partition

The Booster is the first LEONARDO's compute partition to go in full production in 3Q 2023. It consists of 3456 heterogeneous nodes designed to create a significant speedup in traditional HPC and new AI applications. In facts, this supercomputer is one of the top level facility in the world to support scientific investigation in many fields: with 238.7 petaFLOPS of sustained Linpack performance, it reached the 4th spot in the TOP500 ranking in June 2023 (Top500, 2023), being at the same time the largest supercomputer based on NVIDIA Ampere architecture, with about 14k GPUs. However, the improvement brought by LEONARDO is not only a matter of pure performance, instead, the design of the machine has been intended to accompany the evolution of computing architectures towards hardware specializations and to extend the support for workloads related with the training and the usage of large AI models.

2.1.1 GPU accelerator device

The A100 *Tensor Core* GPU is an accelerator device introduced by NVIDIA in 2020 based on the *Ampere* micro-architecture (NVIDIA, 2020). In the fast changing accelerators market, it represented a breakthrough in terms of flexibility, computational power (+24% floating point, FP) and communication speed (+73% memory bandwidth) in comparison to its predecessor, the V100 *Tensor Core* GPU based on the *Volta* architecture (NVIDIA, 2017).
The *Ampere* offers an upgraded compute unit structure (third-generation Tensor Core, TC) that extends hardware support for tensor math to a wider set of datatypes including both floating point and integer numerical formats. Concerning floating point computation in double precision (FP64), the device offers an impressive peak performance of about 20 teraFLOPS for tensor operations and around 10 teraFLOPS for non-tensor math. Together with the single precision performance treated below, this is particularly engaging for HPC communities that rely on very high precision representations in their models. In fact, moving from V100 to A100, a speedup between x1.5 and x2.1 has been measured in HPC benchmarks spanning from molecular dynamics to geo-sciences (Krashinsky et al., 2020). On the AI side, a new numerical format called *Tensor Float 32* (TF32) definitively enables the use of TC to accelerate the training of a vast number of neural network models. The TF32 is a custom floating point format with 8-bit range (as in FP32) and 10-bit precision (as in FP16). The halved precision does not affect the accuracy of the computations in the AI context and brings a significant speed up instead. The FP32 data path has been kept for I/O operations and the TF32 is the default choice for computation, so the speedup benefit is transparent to the user (no code change). For maximum speed in training, the supported tensor math includes the standard FP16 datatype (inherited from the previous generation TC) and the new AI dedicated BF16 datatype (8-bit range, 7-bit precision) which allow a factor x2 in throughput respect to TF32 and a factor x20 compared to non-tensor operations. Integer arithmetic is supported as well, for example 8-bit operations have a peak performance of 624 teraOPS and INT4 and binary even more.

	Ampere A100 (custom)	Ampere A100	Volta V100
FP64 [teraFLOPS]	11.2	9.7	7.8
FP32 [teraFLOPS]	22.4	19.5	15.7
FP64 TC [teraFLOPS]	22.4	19.5	n.a.
TF32 TC [teraFLOPS]	179	156	n.a.
FP16 TC [teraFLOPS]	358	312	n.a.
INT8 TC [teraOPS]	716	624	n.a.
INT4 TC [teraOPS]	1432	1248	n.a.
SM [#]	124	108	80
CUDA FP64 core [#]	3968	3456	2560
CUDA FP32 core [#]	7936	6912	5120
CUDA Tensor core [#]	496	432	640
Max Clock [MHz]	1395	1410	1530
L2 Cache [MB]	32	40	6
Memory [GB]	64	40	16
Memory BW [GB/s]	1640	1555	900
TDP [W]	440	400	300

Table 2: Comparison of GPU chips specifications

Table 2 displays the main specifications of the two generations (*Ampere* and *Volta*) and presents the characteristic of the A100 variant installed in LEONARDO. The latter is a *custom* model consisting in a 97% implementation of the full A100 GPU design (124 vs 128 Streaming Multiprocessors, SM), while the *standard* A100 uses 84% of it (104 SM).

In addition, the A100 offers an instructions set called *Sparse Tensor Core* that double the TC performance reported in Table 2 when working with AI applications. With this approach, which is referred to as *Structural Sparsity* by the vendor (NVIDIA, 2020b), the pruning of the weights matrix is structurally constrained by zeroing two elements out of four in a row. At inference time, an efficient use of hardware resources allows to gain a clean factor two in throughput.

2.1.2 GPU blade

The Booster's blade is a single node blade, based on the latest high-end GPU server board by Eviden company (BullSequana X2135). The blade is called *Da Vinci* and a picture of it is shown in Figure 2. The entire blade is liquid-cooled, so there are no fans onboard.

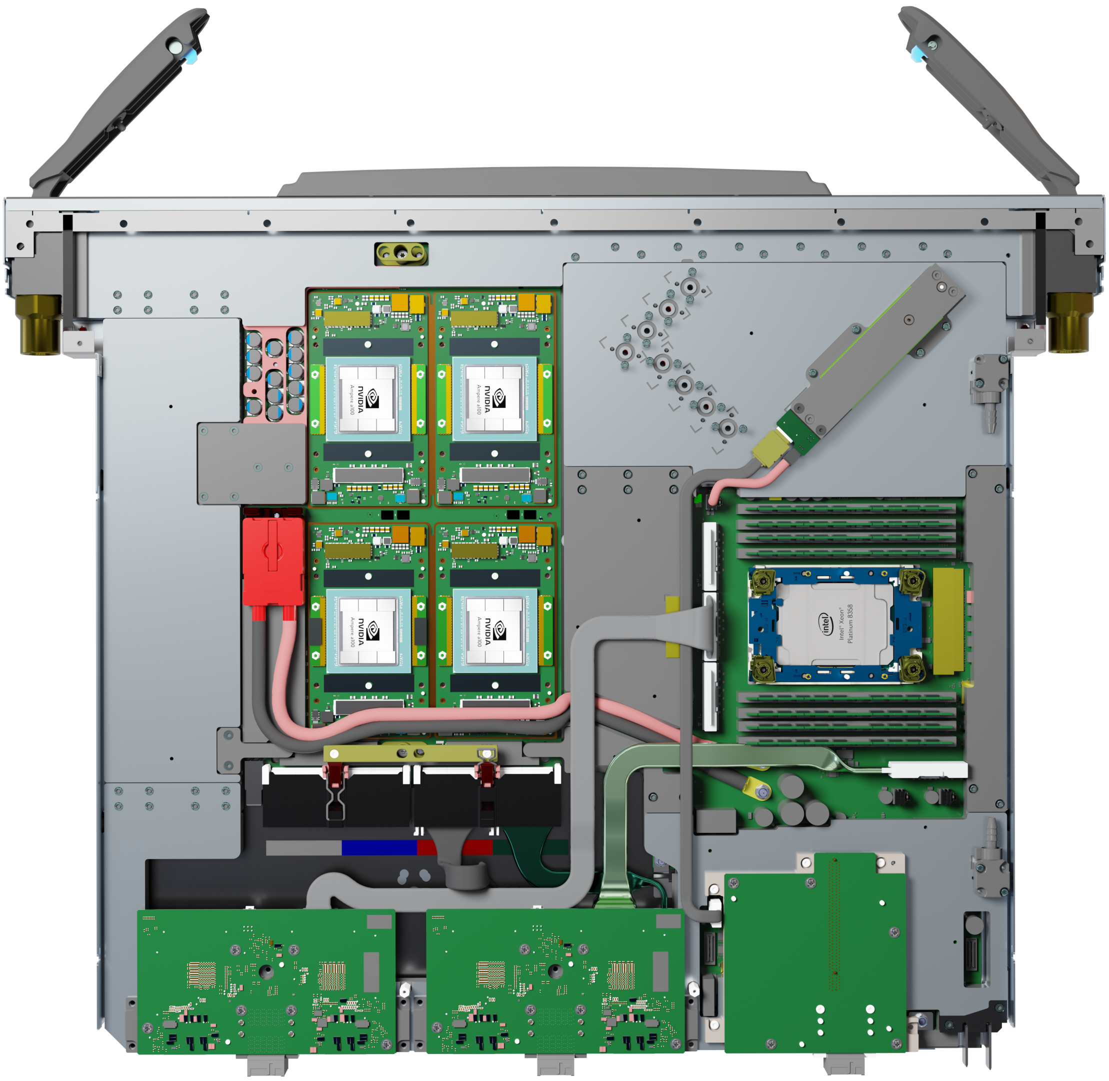

Figure 2: GPU blade top view

The host processor is a single socket Intel Xeon Platinum 8385 CPU (Intel, 2019) with 32 cores and 48 MB cache (codenamed *Ice Lake*). The IceLake CPU is AVX-512 capable. Each core contains two AVX-512 execution units which results in a 1024 operations per clock cycle and a peak performance of 2.6 teraFLOPS per core at the nominal frequency of 2.6 GHz. The memory subsystem is DDR4, clocked at 3200 MHz (6400 MT/s). There are eight 32-bit memory controllers. Each one is capable of 25 GB/s for a total maximum bandwidth of 200GB/s for CPU-RAM communication. The corresponding eight DIMM slots are equipped with 64 GB capacity modules, so the total RAM available on node is 512 GB. Four A100 GPUs (see 2.1.1) in SXM4 form factor are integrated in the blade. The local memory subsystem of the GPU is placed in the same physical chip of the processing element, thus offering high density and performance. This relies on the second generation High Bandwidth Memory express interface technology (HBM2e). Each GPU has 64 GB of addressable memory that is organized into four 16 GB HBM2e stacks. Each stack is controlled by two 512-bit memory controllers capable of 3200 MT/s. Overall, more than a terabyte per second can be delivered by each GPU, namely 1638 GB/s. In total, the local storage for GPU computation is 320 GB in capacity and can be accessed with an impressive 6.5 TB/s aggregated bandwidth.

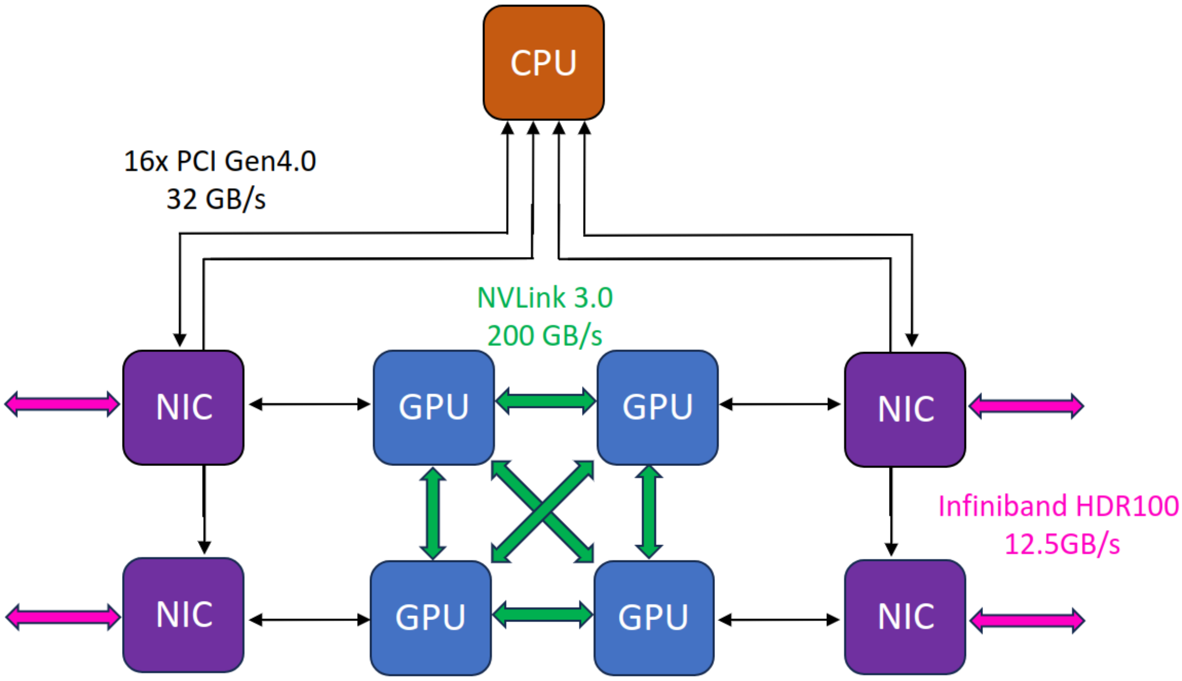

Figure 3: Booster blade intra-node communication pattern (logic view)

Intra-node communication pattern is depicted in Figure 3. The CPU utilizes four bundles of PCIe lanes to communicate independently with individual GPU. A bundle consists of 16 PCIe Gen 4.0 lanes for a total of 32 GB/s bandwidth per CPU-GPU communication. Total bandwidth available along the 64 lanes of the CPU is 128 GB/s. Multi-GPU systems are supported by a proprietary fast high speed interconnect (NVIDIA NVLink 3.0) that provides 200 GB/s bidirectional bandwidth per GPU pair, 600 GB/s in total. The A100 blade is equipped with 2 dual-port Mellanox HDR100 ConnectX-6 InfiniBand network interface cards (NIC) for inter-node communication. They provide an aggregated 400 Gbps bandwidth as well as CPU offloading features that are described in the next sections.

2.2 Network system

The internal network of a cluster connects compute nodes together and offers access to storage spaces. LEONARDO's network follows a scalable hierarchical cell-based architecture, with a cell being a collection of server nodes. At top level, there are 23 cells fully connected in a *dragonfly* topology (Kim et al., 2008) as shown in Figure 4.

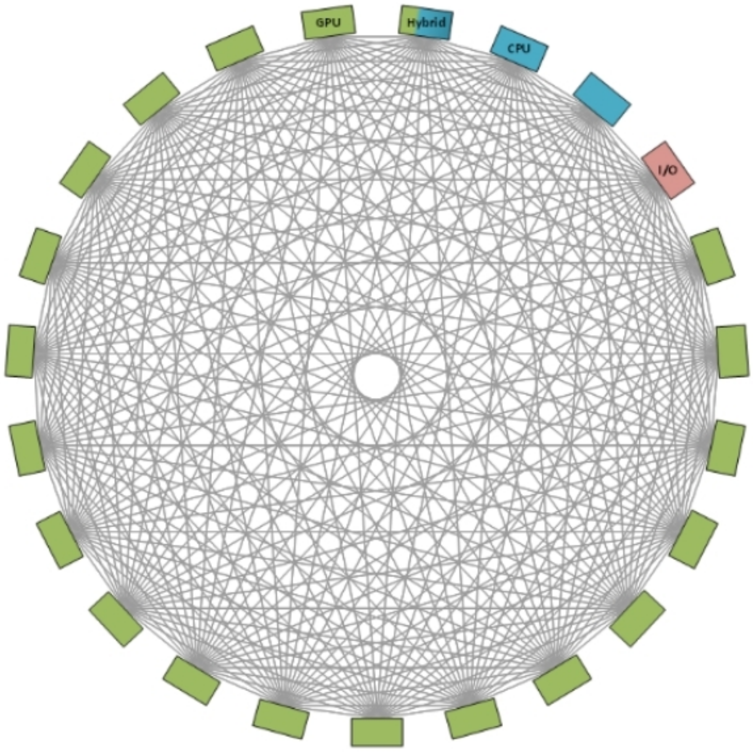

Figure 4: *Dragonfly* topology of the internal network. Colors indicate the technology of the underlying nodes. Green is used for Booster cells, blue for DC cells, pink for the I/O. See text for details.

Locally, intra-cell routers are organized in a bipartite graph in which a first tier is directly connected to servers (*leaf* routers) and a second tier (*spine* routers) is equally provisioned with down-links. Such scheme, called *dragonfly+*, allows twice the group size and a factor four in scalability, when compared to the standard *dragonfly* topology (Shpiner et al., 2017), it is denser and request less switches.

Nodes in the cluster are tighly coupled using 200 Gbps InfiniBand Mellanox's High Data Rate (HDR) technology components (NVIDIA, 2020c). The switch model is the QM8700, offering a latency of 90 nanoseconds port-to-port and up to 390 million messages delivery per second per port (NVIDIA, 2020d). The total number of QM8700 switches is 823. Two configurations are used: 40 ports at 200 Gbps or 80 ports at 100 Gbps bandwidth, with the latter widely adopted at *leaf* level and referred to as HDR100.

Depending on the underlying node's technology, the two tiers of routers have different arrangement:

- *Spine* switches are always 18 per cell, regardless the cell type. They are configured in 200G 40-port mode with 22 up-links and 18 down-links (pruning factor 0.82).
- *Leaf* switches organization depends on the cell type and is always HDR100, except for the *Fast* storage subsystem where each link uses the full 200G HDR bandwidth (see Section 2.3). Booster's *leave* switches receive 20 down-links each (pruning factor 1.11) with a single node spreading its connections over two different physical switches. Differently, in the Data-Centric partition each node is connected to a single *leaf* switch, i.e. 16 HDR ports on a switch serve 32 CPU nodes. For the hybrid cell, the two arrangements just described are combined, namely 6 out of 8 racks of the cell are DC style and the remaining 2 are Booster style. The number of *leaf* switches per cell is 18 for Booster and Hybrid type and 16 for the DC type. The I/O cell uses 13 *leaf* switches.

At node level, the network adapter is the ConnectX-6 card (CX6) which can sustain up 200 millions of messages per second with a latency of 600 ns (NVIDIA, 2020a). The CX6 support PCIe Gen4 communication on 32 lanes including pass-through functionality. Applications that do not require the entire bandwidth can benefit from the integrated PCIe switch that allows to serve up to 8 virtual machines

BY

on host[2]. In addition, the CX6 comes with acceleration engines that provide CPU offloading for important HPC and AI tasks like: Remote Direct Memory Access (RDMA) for direct data movement from storage infrastructure to GPU local memory, transport operations like adaptive routing and congestion management, MPI collectives and tag matching, encryption based on personal user key.
Considering the latencies of the switch and of the NIC mentioned above and the following lengths of optical fiber - 1 meter from NIC to leaf, 5 meters from leaf to spine and 20 meters between the spines - the maximum latency between two nodes located at opposite side of the cluster is 3 microseconds. In general, inter-node communication latency is dominated by the sending and receiving NICs that introduce 1.2 microseconds delay, independently from the destination.
Finally, four gateways routers are used to interface the cluster with external networks. Each of these units provides eight 200 Gbps Ethernet-InfiniBand protocol translators for a total bandwidth per unit of 1.6 Tbps and 6.4 Tbps aggregated (NVIDIA, 2021). In addition, an Ethernet administrative network is used for management, with dedicated switches at rack level and single port adapters on each node.

2.3 Storage system

A 12-rack subsystem provides storage functionality to the whole computer cluster. The subsystem is based on DDN's appliances and consists of two tiers, named *Fast* and *Capacity*, to accommodate all the requirements of modern HPC and AI diverse data access patterns:

- *Fast Tier* provides 5.7 PB of raw memory for IOPS eager applications and offers burst buffer capability for *hot* data generally. It is composed by 31x ES400NVX2 appliances configured with $\simeq$ 150 TB of solid state drives (SSD) using Non-Volatile Memory Express (NVMe) technologies.
- *Capacity Tier* is a 137.6 PB raw capacity storage partition using SAS Hard Disk Drive (HDD) components. It consists of 31 modules composed by a controller head (ES7990X) and two SAS expansion enclosures (SS9012) housing a total of 246 by 18 TB HDD, providing 4.4 PB of storage capacity per module. Metadata is handled by four additional flash-based ES400NVX units.

Overall, the storage subsystem consists of 66 DDN's appliances together with the related software stack that is essential for high speed data movement in the different computing scenarios that LEONARDO can serve, in addition to standard fault tolerance and security functions. Table 3 shows the mapping of the three global namespaces to the hardware resources just described, together with related net size and measured bandwidth.
The filesystem is based on Lustre (Lustre, 2023) and supports encryption and multi-tenancy. The first is a feature for security and isolation that allows to access selected portion of the storage namespace to authenticated users only. This is based on CryptoFS (CryptoFS, 2023). The second feature allows multiple access (multiple client) to files. Of prominent importance for AI workloads, *GPUDirect* technology is also supported by the storage subsystem, it can directly use the GPU memory for I/O, avoiding the use of system memory (RAM) as bounce buffer. With objects striped across multiple disks, Lustre provides also parallel access to large files at near-wire speed.

Work area	ES7990X #	ES400NVX2 #	ES400NV #	NetSize PiB	Bandwidth GB/s
`/home`	-	4	-	0.5	240
`/archive`	18	-	2	53.9	360
`/scratch`	13	27	2	42.4	1300

Table 3: Filesystem organization and characteristcs

[2]Known as MultiHost.

CC BY

2.4 Frontend and service partitions

The Frontend partition provides the user with access to the system. Typical operations on frontend nodes encompass software development, code compilation, data management, interface to other systems on site, job submission, data pre-processing, data post-processing and results visualization. In LEONARDO the frontend nodes see both compute partitions and the global filesystems as well. The number of frontend servers is 32, each equipped with a dual socket Xeon Scalable processor (2x 32 cores, Intel 3rd Gen, the same model of the Booster's compute node) and 16 DDR4-3200 channels per socket. Sixteen nodes are dedicated to login and are configured with a local 6 TB disk space in RAID-1 configuration. The other 16 nodes are specialized for post-processing visualization and are equipped with NVMe disks (6.4 TB total local capacity) and two NVIDIA PCIe Quadro RTX8000 by 48 GB RAM each.
In order to deploy, manage and monitor LEONARDO cluster that accounts for about 5000 compute nodes, 11 tailored servers are used, called Operational Managment Nodes (OMN). OMNs feature a single AMD EPYC *Rome* CPU with 64 cores and 3 three dual ports NICs supporting 10 GbE, 50 GbE and HDR100.
The complete list of hardware components is reported in Appendix B.

2.5 Software ecosystem

LEONARDO runs Red Hat Enterprise Linux 8 operating system on all nodes and uses SLURM as workload manager (SLURM, 2023). Two architecture-specific suites are installed, namely Intel OneAPI and NVIDIA HPC SDK. The latter includes a complete software stack to build AI applications with highly optimized libraries such as cuDNN for deep neural networks and NCCL for multi-GPU communication. The GNU compiler collection is also installed.
Software management is done using Spack (Gamblin et al., 2015) and Environment Modules (Environment Modules, 2023). A large set of HPC programming tools is available for developers, based both on closed and open source products. The software for scientific production is organized on a category basis, serving each research community with dedicated pre-installed tools e.g. chemestry-physics, deep learning, life sciences and meteo. Further details and updates can be found in the user guide available on the website of CINECA (LEONARDO UserGuide, 2023), while baseline tools are listed below.

- Parallel profilers and debuggers
  - GNU debugger (GDB)
  - Intel debugger (IDB) and VTune profiler
  - NVIDIA Nsight profiler (System and Compute) and CUDA-GDB
  - Valgrind
- Communication libraries
  - OpenMPI,
  - Intel MPI
- Numerical application libraries
  - Intel Math Kernel Library,
  - GNU scientific library,
  - Math and Python libraries
- Containerization is supported through several different tools:
  - Syslab Singularity Enterprise edition
  - NVIDIA Container Framework and Pyxis Slurm plugin
  - ParTec Parastation also supports the execution of containerized applications, improving the flexibility of a pure Singularity approach;
- Monitoring is operated via Eviden SMC xScale suite, based on Prometheus, and using Grafana as frontend. Detection and tracking of issues is performed by Parastation HealthChecker.

2.6 Power consumption, cooling and management

LEONARDO is hosted by CINECA in its new data center at the Big Data Technopole (Tecnopolo, 2023) in Bologna, Italy. The room floor has been designed with a plan in two steps to support the current pre-exascale and a future exascale machine. Presently, the data center features 10 MW of IT load with 1240 m^2 of computing floor space and 900 m^2 of ancillary space. The second step considers an increased power support up to 20 MW IT load and 2600 m^2 additional computing floor.
All major components of LEONARDO are cooled down using warm water-cooling technology, including the power supplies. The inlet water temperature is 37 Celsius degrees and the total Direct Liquid Cooling capacity is 8 MW. The system is pretty efficient with a 1.1 value of Power Usage Effectiveness (PUE), this means that the overhead needed to cool down LEONARDO is the 10% of the power used to feed it.
Energy consumption of the cluster is controlled by means of various tools including two Eviden proprietary software products (Bull Energy Optimizer and Bull Dynamic Power Optimizer). One allows to log time profiles of energy and temperature via IPMI and SNMP protocols and to cap the clock frequency of the CPUs depending on the total power consumption. The other is used to find the best workpoint in terms of energy consumption and performance of a running application i.e. reducing the power absorption by adjusting the clock frequency with limited performance degradation. Concerning GPU, a vendor specific manager tool (NVIDIA Data Center GPU Manager) is used to limit device clock when a configurable energy threshold is surpassed.

3 Access to LEONARDO

LEONARDO is a EuroHPC-JU system that is hosted and operated by CINECA supercomputing center. Researchers from academia, research institutes, public authorities, and industry can apply for access to computing time. The access is mainly based on *Calls for Proposal* from EuroHPC (50%) and CINECA (50%) via its ISCRA program (Italian SuperComputing Resource Allocation). Submitted proposals are peer reviewed for scientific merit and undergo a technical assessment for suitability to perform on LEONARDO architectures, in order to ensure the highest scientific reach of the selected project. Detailed information are available at (EuroHPC JU, 2023) and (ISCRA, 2023) webpages.

Acknowledgement

The acquisition and operation of LEONARDO supercomputer is funded jointly by the Italian Ministry of University and Research and by the EuroHPC Joint Undertaking (EuroHPC JU) under grant agreement *N. cnect.ddg1.c.2(2019)8804531 - LEONARDO supercomputer* through the European Union's Connecting Europe Facility and the Horizon 2020 research and innovation programme. The EuroHPC JU is a legal and funding entity created in 2018 to enable the European Union and EuroHPC participating countries to coordinate their efforts and pool their resources with the objective of making Europe a world leader in supercomputing.
The authors thank Eviden for the solution provided and all the key technology partners NVIDIA, Intel, DDN, for their support during design, construction, delivery and testing.

References

Amati, G., Succi, S., Fanelli, P., Krastev, V. K., and Falcucci, G. (2021). Projecting LBM performance on Exascale class Architectures: A tentative outlook. *Journal of Computational Science*, 55:101447.

CHEESE2 (2023). Centre of Excellence for Exascale in Solid Earth. https://cheese2.eu/.

CINECA (2023). Italian interuniversity consortium. https://www.hpc.cineca.it/.

CryptoFS (2023). CryptoFS repository. https://reboot.github.io/cryptofs/.

Declaration (2018). European declaration on High Performance Computing as part of the European digital strategy. https://digital-strategy.ec.europa.eu/en/news/european-declaration-high-performance-computing.

DTGEO (2023). Digital twin for geophysical extremes. https://dtgeo.eu/.

Environment Modules (2023). Environment modules open source project webpage. https://modules.sourceforge.net/.

EuroHPC JU (2018). EuroHPC JU webpage. https://eurohpc-ju.europa.eu/index_en.

EuroHPC JU (2023). EuroHPC JU supercomputers access policy. https://eurohpc-ju.europa.eu/access-our-supercomputers/access-policy-and-faq_en.

Eviden (2020). Bullsequana XH2000 product webpage. https://atos.net/wp-content/uploads/2020/07/BullSequanaXH2000_Features_Atos_supercomputers.pdf.

Falcucci, G., Amati, G., and Fanelli, P. e. a. (2021). Extreme flow simulations reveal skeletal adaptations of deep-sea sponges. *Nature*, 595:537–541.

Gamblin, T., LeGendre, M., Collette, M. R., Lee, G. L., Moody, A., de Supinski, B. R., and Futral, S. (2015). The Spack Package Manager: Bringing Order to HPC Software Chaos. In *Proceedings of the International Conference Supercomputing 2015 (SC15)*, Austin, Texas, USA. LLNL-CONF-669890.

GEOIN (2023). Geosphere Infrastructures for Questions into Integrated Research. https://www.geo-inquire.eu/.

Intel (2019). Xeon Platinum 8358. https://www.intel.it/content/www/it/it/products/sku/212282/intel-xeon-platinum-8358-processor-48m-cache-2-60-ghz/specifications.html.

Intel (2023). Xeon Platinum 8480+. https://ark.intel.com/content/www/us/en/ark/products/231746/intel-xeon-platinum-8480-processor-105m-cache-2-00-ghz.html.

IO500 (2023). IO500 benchmark repository. https://github.com/IO500/io500.

ISCRA (2023). Italian supercomputing resource allocation policy. https://www.hpc.cineca.it/services/iscra.

Kim, J., Dally, W. J., Scott, S., and Abts, D. (2008). Technology-driven, highly-scalable dragonfly topology. *2008 International Symposium on Computer Architecture, Beijing, China*, pages 77–78.

Krashinsky, R., Giroux, O., Jones, S., Stam, N., and Ramaswamy, S. (2020). Ampere Architecture In-Depth. https://developer.nvidia.com/blog/nvidia-ampere-architecture-in-depth/.

Lustre (2023). Lustre official webpage. https://www.lustre.org/.

NVIDIA (2017). Volta V100 Tensor Core GPU datasheet. https://images.nvidia.com/content/technologies/volta/pdf/tesla-volta-v100-datasheet-letter-fnl-web.pdf.

NVIDIA (2020). Ampere A100 Tensor Core GPU datasheet. https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet-us-nvidia-1758950-r4-web.pdf.

NVIDIA (2020a). ConnectX-6 host channel adapter product webpage. https://www.nvidia.com/content/dam/en-zz/Solutions/networking/ethernet-adapters/connectX-6-dx-datasheet.pdf.

NVIDIA (2020b). GA100 whitepaper. https://images.nvidia.com/aem-dam/en-zz/Solutions/data-center/nvidia-ampere-architecture-whitepaper.pdf.

NVIDIA (2020c). High Data Rate Infiniband technologies. https://network.nvidia.com/files/doc-2020/wp-introducing-200g-hdr-infiniband-solutions.pdf.

NVIDIA (2020d). QM8700 switch product webpage. https://network.nvidia.com/files/doc-2020/pb-qm8700.pdf.

NVIDIA (2021). Skyway product specifications. https://nvdam.widen.net/s/xcrjcmcxtj/infiniband-datasheet-rebranding-nvidia-skyway-1870668-r5.

Shpiner, A., Haramaty, Z., Eliad, S., Zdornov, V., Gafni, B., and Zahavi, E. (2017). Dragonfly+: Low cost topology for scaling datacenters. *2017 IEEE 3rd International Workshop on High-Performance Interconnection Networks in the Exascale and Big-Data Era (HiPINEB), Austin, TX, USA*, pages 1–8.

SLURM (2023). Slurm workload manager documentation. https://slurm.schedmd.com/documentation.html.

Succi, S., Amati, G., Bernaschi, M., Falcucci, G., Lauricella, M., and Montessori, A. (2019). Towards Exascale Lattice Boltzmann computing. *Computers & Fluids*, 181:107–115.

Tecnopolo (2023). Tecnopolo Manifattura Data Valley Hub webpage. https://www.tecnopolomanifattura.it/en/home.

Top500 (2023). Top500 list June 2023. https://www.top500.org/lists/top500/2023/06/.

LEONARDO UserGuide (2023). LEONARDO User guide. https://wiki.u-gov.it/confluence/display/SCAIUS/UG3.4%3A+Leonardo+UserGuide.

A Benchmark results

The procurement process of LEONARDO relied on the evaluation of prominent and application-specific workloads. Here, results for the following benchmarks are sketched as a list of tables.

- Synthetic HPC benchmarks
  - High Performance Linpack (HPL)
  - High Performance Conjugate Gradients (HPCG)
  - IO500
- Application benchmarks
  - QuantumEspresso: electronic-structure calculations and materials modeling
  - SPECFEM3D Globe: global seismic wave propagation
  - PLUTO: astrophysical gas dynamics
  - MILC: lattice QCD calculations
  - Lattice Boltzmann Method (LBM): computational fluid dynamics

A.1 HPL, HPCG and Green500

Table 4 summarize the metrics of LEONARDO Booster as presented at TOP500 in June 2023 (Top500, 2023). The most relevant is a 238.7 petaFLOPS measured HPL performance out of 304.5 petaFLOPS of theoretical peak. Such a result was achieved by using 3300 compute nodes. Total power consumption was 7.4 MegaWatts, thus giving an average performance per unit of power of 32.2 gigaFLOPS/W. These results earned LEONARDO the rank 4^{th} and 15^{th} in the TOP500 list and Green500 list respectively. In the same edition, LEONARDO was also ranked 4th in the HPGC category with a performance of 3.11 petaFLOPS.

Benchmark	Performance [petaFLOPS]	Rank
HPL	238.7	4
HPCG	3.11	4

Table 4: LEONARDO performance at TOP500 in June 2023.

A.2 IO500

At ISC 2023 among the production machines, LEONARDO was 1^{st} in the bandwidth category of IO500 list (IO500, 2023). Related performances figures are shown in Table 5. Standard `ior` benchmark results are 1533 GiB/s and 1883 GiB/s bandwidth for `ior-easy-write` and `ior-easy-read` respectively.

Benchmark	Score	BW (GiB/s)	MD (KIOP/s)	Rank
IO500	649	807	522	1

Table 5: LEONARDO IO500 performance at ISC2023.

A.3 Application Benchmarks

Table 6 shows the results of domain-specific application benchmark in terms of *Time-to-Solution* (TTS) in seconds and *Energy-to-Solution* (ETS) in kWh. The job size in terms of number of nodes spans from 12 to 32. In case of PLUTO the ETS has been estimated using CPU power consumption only, since the program does not use GPUs.

Application name	Domain	Nodes	TTS	ETS
QuantumEspresso	Quantum Chemistry	12	439	1.14
MILC	Quantum Chromodynamics	12	178	0.56
SPECFEM3D	Solid Earth	16	270	1.43
PLUTO	Astrophysics	32	2874	11.7

Table 6: LEONARDO application benchmarks performance.

Figure 5 presents the weak scaling of the Lattice Boltzmann Method (LBM) benchmark, a code that has been described in details by Falcucci et al. (2021) and Succi et al. (2019). Considering the scaling efficiency of the same code on *Marconi100*, a CINECA's GPU based cluster equipped with NVIDIA V100, a significant better performance has been measured. In terms of TTS, LEONARDO was about 2.5 times faster than *Marconi100* as documented by Amati et al. (2021). Table 7 presents the performance of LEONARDO in terms of Lattice Updates per Seconds (LUPS) and explicits the total number of GPUs in each point of the scaling.

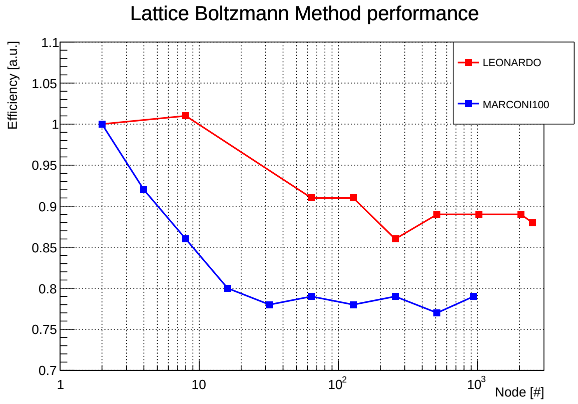

Figure 5: LBM Weak scaling efficiency comparison

Nodes [#]	#GPUs	Performance [LUPS $\times 10^{12}$]	Efficiency
2	8	0.0476	1.00
8	32	0.192	1.01
64	256	1.38	0.91
128	512	2.76	0.91
256	1024	5.24	0.86
512	2048	10.8	0.89
1024	4096	21.6	0.89
2048	8196	43.3	0.89
2475	9900	51.2	0.88

Table 7: LBM Weak scaling efficiency

B Hardware components

As of this writing, LEONARDO supercomputer consists in the following components.

Booster partition

- 3456 nodes (13824 GPUs)
- single node *Da Vinci* blade, based on the BullSequana X2135
  - 1x 32-core Intel Xeon Platinum 8358 CPU, 2.6 GHz (*Icelake*)
  - 8x 64 GB DDR4-3200 (512 GB)
  - 4x NVIDIA custom Ampere A100 GPU 64 GB HBM2
  - 2x dual-port HDR network interface (400 Gbps aggregated)

Data-Centric partition

- 1536 nodes (172032 CPU cores)
- three-node BullSequana X2140 blade, each node with
  - 2x 56-core Intel Xeon Platinum 8480+ CPU, 2.0 GHz (*SapphireRapids*)
  - 16x 32 GB DDR5-4800 (512 GB)
  - 1x SSD 3.84 TB M.2 NVMe
  - 1x single port HDR100 network interface (100 Gbps)

Storage

- Fast Tier, 5.7 PB full flash
  - 31x DDN appliance ES400NVX2 configured with
    - 24x SSD 7.68 TB NVMe with encryption support (184.3 TB)
    - 4x InfiniBand HDR ports (800 Gbps aggregated)
    - metadata resource included
- Capacity Tier, 137.6 PB
  - 31x DDN appliance ES7990X configured with
    - 1 Controller head (82 disks) + 2 expansion enclosures (SS9012, 164 disks)
    - 246x HDD 18 TB SAS 7200 rpm (4.4 PB)
    - 4x InfiniBand HDR100 ports (400 Gbps aggregated)
  - 4x DDN appliance SFA400NVX for metadata (322 TB), configured with
    - 21x SSD 3.84 TB NVMe with encryption support (80.8 TB)
    - 8x InfiniBand HDR100 ports (800 Gbps aggregated)

Frontend partition

- 32 Frontend nodes (16 login + 16 graphical)
  - 2x 32-core Intel Xeon Platinum 8358 CPU, 2.4 GHz (*Icelake*)
  - 16x 32 GB DDR4-3200 (512 GB)
  - 1x HDR100 network interface
  - 2x 50 GbE network interface.
  - *Login*, BullSequana X430-E6
    - 6 TB HDD in RAID1
  - *Graphical*, BullSequana X450-E6
    - 6.4 TB NVMe
    - 2x GPU PCIe NVidia Quadro RTX8000 48 GB

Service partition

- 11 Operational Management nodes (3 Master + 8 Worker)
  - 1x 64-core CPU AMD EPYC 7h12 (Rome), 2.6 GHz, TDP 280 W
  - 1x dual-port HDR100 network interface
  - 1x dual-port 50 GbE network interface
  - 1x dual-port 10 GbE network interface
  - *Master*
    - 8x 16 GB DDR4-3200 (128 GB)
    - 2x 960 GB NVMe with M.2 slots
    - 2x 3.84 TB 2.5 inches SATA3 SSD
  - *Worker*
    - 16x 32 GB DDR4-3200 (512 GB)
    - 2x 3.2 TB NVMe U.2
    - 4x 3.84 TB 2.5 inches SATA3 SSD
    - 8x 12 TB 3.5 inches SATA3 HDD