Case Study

High Performance Computing
4th Gen Intel® Xeon Scalable Processor
3rd Gen Intel® Xeon Scalable Processor

Cineca Drives Toward Exascale HPC with 250 PetaFLOPS Leonardo Supercomputer

Optimized for accelerated computing to enable scientific breakthroughs and industry innovation, Leonardo's hybrid architecture includes 3rd and 4th Gen Intel® Xeon® Scalable processors with built-in accelerators for significant workload performance.

Leonardo Supercomputer Highlights:

- Built by Atos on BullSequana XH2000 platform
- 250 petaFLOPS HPL (Rmax)/ 10 exaFLOPS FP16 AI performance
- 3,456 servers with Intel® Xeon® 8358 processor and NVIDIA Ampere GPUs
- 1,536 servers with 4th Gen Intel® Xeon® Scalable processors
- 5PB of high-performance storage/ 100PB of large-capacity storage

Executive Summary

Italy has a long history of innovation and support for High Performance Computing (HPC) for research and industry throughout Europe. At the center of Italy's commitment to HPC is Cineca, a private, nonprofit consortium made up of the Ministry of Education, the Ministry of University and Research, 69 Italian universities, 28 national public institutions, and 13 national research institutes. The organization provides HPC resources and high-level support to its members. All members collaborate in a wide range of research projects across Europe. Discoveries and insight are used to further scientific exploration and for commercial applications, making Cineca a technological bridge between academics and science domains and industry.

Cineca deploys new HPC systems periodically to continue to provide advanced technologies to its customers. Cineca recently deployed its most powerful supercomputer built on the latest generations of Intel® Xeon® Scalable processors and NVIDIA GPUs. The new HPC system, named Leonardo, ranks #4 on the November Top500.org list. Leonardo is designed to deliver 250 petaFLOPS HPL compute performance and 10 exaFLOPS of FP16 AI performance, giving Cineca a new achievement along its roadmap to be a leading supercomputing center for Europe.

Challenge

Italy's Cineca provides HPC services throughout Europe to enable discovery and innovation in science and industry. It supports advanced research in material

Leonardo, built from advanced technologies including 4th Gen Intel® Xeon® Scalable processors, will deliver amazing capabilities to Cineca and its customers.

	FAST TIER	CAPACITY TIER
Net Capacity	5.4 PB	100 PB
Disk Technology	Full FLASH (NVMe and SSD)	NVMe and HDD
Bandwidth	Aggregated: 1400 GB/s r/w; io500: 676 GB/s	Aggregated: 744 GB/s read; 620 GB/s write; io500: 197 GB/s

Table 1. Leonardo storage partition summary (courtesy of Cineca)

science, astrophysics, engineering, bioinformatics, weather and climate, and other fields. As research data expands dramatically along with HPC technologies and methodology advancements, workloads need ever-demanding computing resources. Adding Artificial Intelligence (AI), machine learning (ML), and deep learning (DL) into the workflows calls for advanced supercomputing architectures. Thus, Cineca's philosophy is to stay on the very competitive edge of HPC and follow an aggressive Exascale roadmap to maintain Cineca as a leader in world supercomputing.

To drive research forward, Cineca needed a new system with computing capabilities not yet available in Italy that complemented its many pre-Exascale systems, including:

- Several Marconi tiers, including Marconi-A3 built on Intel® Xeon® 8160 processors and Intel® MCU architecture.
- Marconi100, a 100 petaFLOPS system.
- Galileo100 and ADA Cloud designed around Intel® Xeon® 8260 processors

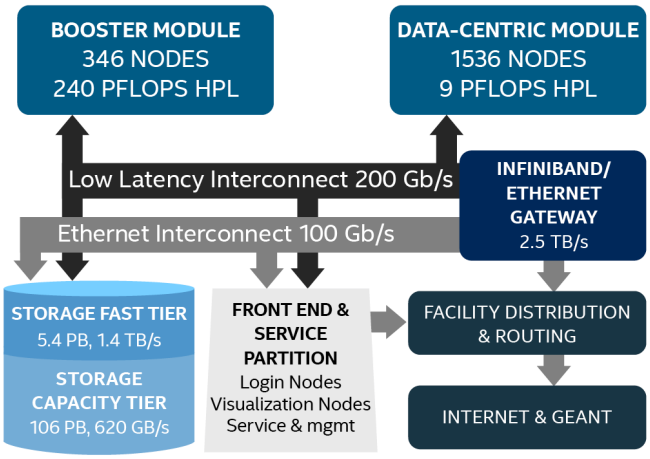


Figure 1. Leonardo system overview (courtesy of Cineca)

Supporting supercomputing in Europe, the EuroHPC Joint Undertaking (EuroHPC JU) project helps fund expansion of supercomputing resources in the continent. Together, Cineca's and EuroHPC JU's aggressive supercomputing plans enabled the building of a new HPC resource at Cineca, called Leonardo, with next-generation data center and supercomputing technologies for a wide variety of traditional workloads, visualization, and AI.

Solution

Leonardo is the first of many HPC systems being deployed across Europe under the EuroHPC JU. With funding from EuroHPC JU, Cineca and other European HPC centers, are on track to deliver Exascale supercomputing in the near future to meet the demands of the world's grand challenges.

Cineca's customers' workloads present a range of demands on computing resources, including memory bandwidth, data throughput, floating point and matrix computation, and others. Such workloads include ab initio materials science and molecular modeling, weather and climate modeling, plasma physics simulation, large-scale bioinformatics, AI and ML, and many other demanding applications. Thus, Leonardo needed to offer both high performance general purpose HPC and AI capabilities in a balanced manner to eliminate bottlenecks for the various workloads. For Leonardo, Cineca chose a hybrid architecture with over a million CPU and GPU cores designed for compute-intensive and data-intensive HPC workloads.

System Summary

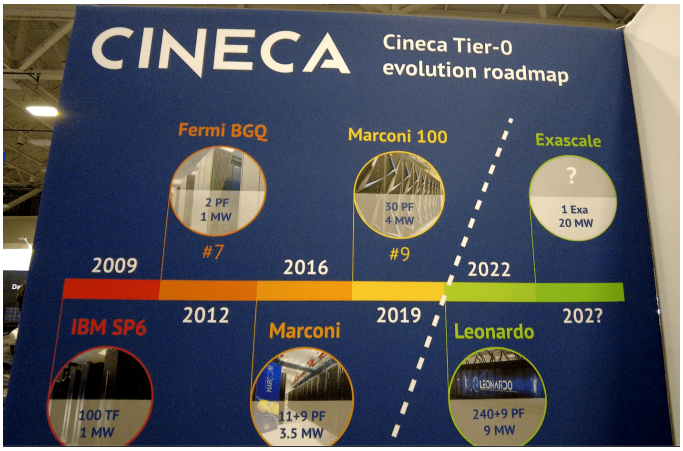
Leonardo was built by Atos on BULLSequana XH2000 supercomputer nodes. The system includes four partitions and more than 136 BULLSequana XH2000 Direct Liquid cooling racks. Leonardo's partitions include a front-end/service tier, storage tier, compute accelerator (booster) tier, and compute (data-centric) tier. The two compute and booster tiers deliver nearly 250 petaFLOPS HPL and 10 exaFLOPS AI 16-bit floating point operations per second.

Front-end/service partition: These provide the login, service, and visualization nodes.

Storage partition: Designed to support both high data throughput and capacity, the storage partition includes a 5-petabyte fast tier and 100-petabyte capacity tier (Table 1). This architecture enables the system to address demanding I/O use cases with extreme bandwidth and IOPS, while providing capacity for the large datasets seen in today's computational problems and AI.

Computing (data-centric) partition: With 1,536 BULLSequana X2610 3-node compute blades built on 4th Gen Intel® Xeon® Scalable processors (56 cores each), the compute (data-centric) partitions delivers nine petaFLOPS HPL, according to Cineca.

Compute accelerator (booster) partition: Each of the 3,456 BULLSequana X2135 DaVinci compute blades house an Intel® Xeon® 8358 processor (32 cores) and four custom-designed NVIDIA Ampere GPUs. This partition is designed to satisfy the computational demanding requirements of GPU-accelerated workloads.



Leonardo is the latest addition that complements many of the pre-exascale systems at Cineca.

The system is interconnected by a 200 Gbps InfiniBand Architecture network with 100 Gbps inter-node communications.

4th Gen Intel® Xeon® Scalable Processors

4th Gen Intel Xeon Scalable processors in the compute partition integrate built-in accelerators that are optimized for specific workloads. They deliver increased performance at improved efficiency for optimal total cost of ownership.[1] These accelerators include Intel® Advanced Matrix Extensions (Intel® AMX) with support for BFloat16 and int8 to accelerate neural network computations, Intel® QuickAssist Technology (Intel® QAT) to accelerate cryptography and data compression and Intel® Advanced Vector Extensions (AVX-512) to speed up AI processing, among others.

Power is a key metric in supercomputing centers today. The 4th Gen Intel Xeon Scalable processors are Intel's most sustainable data center processors, with many features for managing power and performance, making improved use of CPU resources to achieve key sustainability goals.

LISA Expansion

Additionally, Cineca has been approved for a major expansion called LISA. Cineca expects the expansion will increase Leonardo's already formidable computing power by around 100 petaFLOPS and broaden the use cases the system can support. LISA will add two new modules. The first is a module with conventional nodes that utilize high-bandwidth memory, aimed at improving the performance of tasks that require fast data transfers between the memory and the CPU. The second module for high-end acceleration will be powered by next-generation GPU server nodes with the goal of providing significant efficiency in terms of performance per watt.

Result

Leonardo will deliver advanced HPC capabilities to Cineca and its members, enabling new discoveries and innovation.

The pre-production phase of the Leonardo supercomputer has begun. The Leonardo Early Access Program (LEAP) aims to support projects with high scientific impact and can take advantage of Leonardo's many computational resources. Researchers across science and industry and public sectors can submit proposals, regardless of their nationality.

Solution Summary

Supercomputing technology continues to advance and computational methodologies are evolving rapidly. These enable new and advanced HPC systems to be deployed around the world. New supercomputers, like Cineca's Leonardo, allow the world's scientists to gain greater insights and achieve new discoveries around the grand challenges they work on.

With funding assistance from EuroHPC JU, Cineca deployed Leonardo, the world's 4th fastest supercomputer, according to Top500.org. Leonardo is built from advanced technologies, including the 4th Gen Intel Xeon Scalable processors. With over 250 pFLOPS HPL and 10 ExaFLOPS 16-bit AI performance, Leonardo will enable Europe's researchers to delve ever deeper into the problems facing the world and innovate new solutions for science and industry.

Find out more about Intel® Xeon® Scalable Processor family at https://www.intel.com/content/www/us/en/products/processors/xeon/scalable.html.

Find out more about Leonardo at https://leonardo-supercomputer.cineca.eu/hpc-system/.

Solution Ingredients

- Built by Atos on BULLSequana XH2000 platform
- 250 petaFLOPS HPL (Rmax) / 10 exaFLOPS FP16 AI performance
- 3,456 servers with Intel® Xeon® 8358 processor and NVIDIA Ampere GPUs
- 1,536 servers with 4th Gen Intel® Xeon® Scalable processors
- 5PB of high-performance storage / 100PB of large-capacity storage

[1] https://edc.intel.com/content/www/us/en/products/performance/benchmarks/4th-generation-intel-xeon-scalable-processors/

032023/RJMJ/DC/PDF ♻ Please Recycle 354707-001US