The EuroHPC JU Supercomputers

Analysis of the Petascale and Pre-exascale systems

September 2021

Table of Contents

Table of Figures

Preface

The EuroHPC Joint Undertaking (EuroHPC JU) is a joint initiative between the European Union, European countries, and private partners. The goal of the JU is to coordinate the efforts and pool resources of its members with the objective to deploy a world-class exascale-level supercomputing infrastructure in Europe and develop innovative supercomputing technologies and applications. The JU was established in 2018 and one of the first goals was the procurement of a number of world-class supercomputing systems that would significantly increase the overall computing capacity in the EU.

This report presents the technical details of EuroHPC JU supercomputers under deployment or currently operational and provides a snapshot of the systems' status as of September 2021. The document is intended to be updated periodically following the evolution of systems deployment and availability.

The document is split into two parts: the first part (Section 1) offers an overview of the systems, summarising their main characteristics and performance, in terms of peak and sustained Linpack performance, and key architecture details. In the second part (Sections 2 and 3), each system's architectures, technologies, and the target applications are presented in detail, along with the benchmarks used during the procurement process.

Technical details of each system have been provided by the respective Hosting Sites. A particular mention to the following people who contributed to the report:

- Mirko Cestari, CINECA
- Mihail Iliev, Sofiatech
- Branislav Jansik, IT4Innovations
- Pekka Manninen, CSC
- Rui Oliveira, MACC
- Valentin Plugaru, LuxProvide
- Dejan Valh, IZUM

1. Introduction

1.1. Overview

The development of an advanced High-Performance Computing Infrastructure is one of the key goals mandated to the EuroHPC Joint Undertaking under Regulation (EU) 2018/1844. The implementation of this strategic goal started in 2019 with the Call for Expression of Interest for Hosting Entities for Petascale and Pre-exascale systems. Eight EU countries responded successfully to the two calls and started working together with the EuroHPC JU on the design of the procurement processes and the technical specifications of the systems. The table below summarises the countries, the hosting sites, and locations where the systems will be installed.

Table 1 – The EuroHPC JU Supercomputers

System type	System name	Hosting site	Site location	Country
Pre-exascale	LUMI	CSC	Kajaani	Finland
	LEONARDO	CINECA	Bologna	Italy
	MareNostrum 5	BSC	Barcelona	Spain
Petascale	Deucalion	MACC	Guimaraes	Portugal
	Discoverer	Sofiatech/PSB	Sofia	Bulgaria
	Karolina	IT4Innovations	Ostrava	Czech Republic
	MeluXina	LuxProvide	Bissen	Luxembourg
	Vega	IZUM	Maribor	Slovenia

Figure 1 shows the geographic distribution of EuroHPC JU systems across the EU.

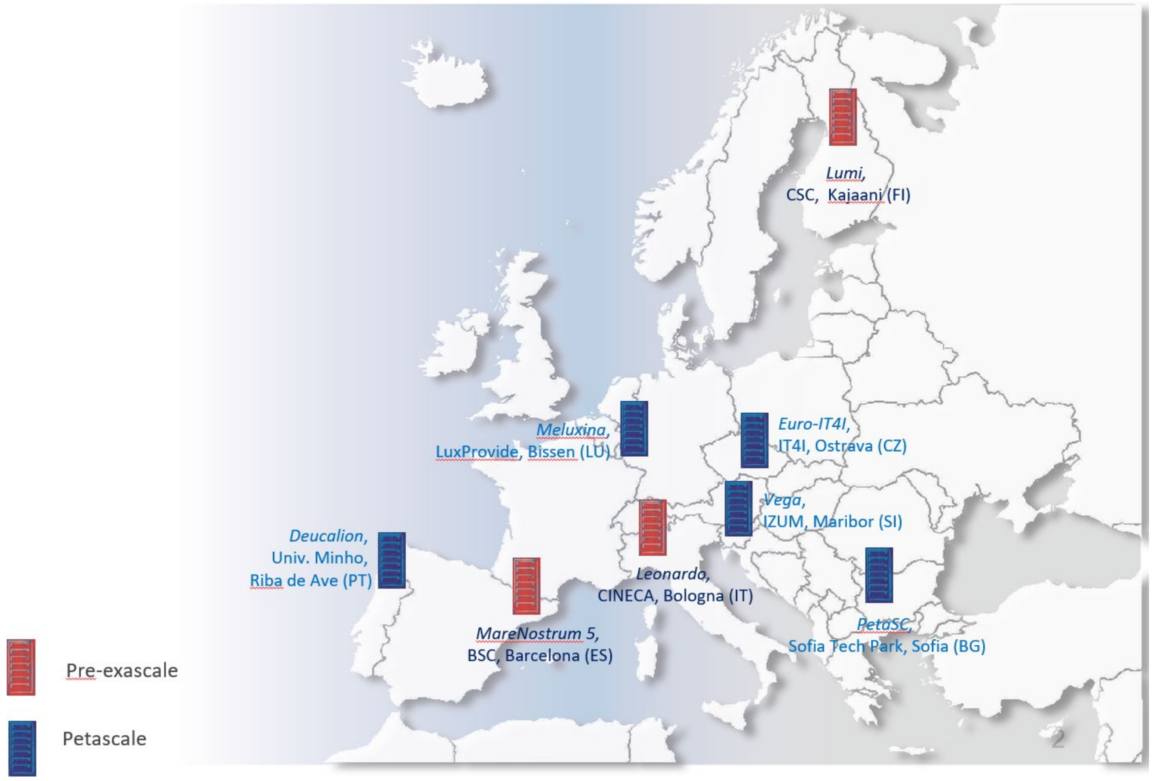


Figure 1 - Site locations of EuroHPC JU supercomputers

At time of writing, seven systems have been successfully tendered and are currently under deployment or have already been accepted and are operational. The only system still pending is MareNostrum 5, whose initial call for tenders was cancelled and is expected to be re-launched by the end of 2021.

1.2. Systems performance

The initial performance targets set for these systems were the following: at least 150 PFlops sustained (Linpack) performance for the pre-exascale and an order of magnitude lower PFlops for the Petascale systems, putting them in the range of 1-100 PFlops. The overall goal was to significantly increase the total raw computing power available to European HPC users.

Table 1 provides an overview of committed sustained performance for ongoing installations or confirmed submitted performance numbers for systems that are already operational. The latter results were submitted to Top500.org and were included in the June 2021 list. The “Aggregated Theoretical Performance” depicts the sum theoretical performance of all individual partitions of the given system. The “Aggregated Sustained Performance” is either the achieved Linpack performance (for Discoverer, Karolina, MeluXina, and Vega) or the one committed by the vendor (for LUMI, Leonardo and Deucalion) across all partitions. As indicated, most of the systems (with the exception of Discoverer) rely on two or more individual compute partitions which offer different configurations and performance characteristics. The “Largest partition Sustained performance” column indicates the confirmed Linpack performance for the partition that offers the largest share of PFlops within the respective system. Accordingly, the “Largest partition Top500.org ranking”, depicts the ranking of the abovementioned largest partition as it was listed in the June 2021 edition of the Top500.org rankings.

Table 2 - Performance characteristics

System name	Aggregated Theoretical Performance (PFlops)	Aggregated Sustained Performance (PFlops)	Largest partition Sustained performance	Largest partition Top500.org ranking (June ’21)
Discoverer	6.0	4.5	4.5	#91
Karolina	15.7	9.4	6.0	#69
MeluXina	18.0	12.8	10.5	#36
Vega	10.1	6.9	3.8	#106
LUMI	*552*	*375.0 (*)*	*N/A*	*N/A*
Leonardo	*322.6*	*249.5 (*)*	*N/A*	*N/A*
Deucalion	*10.0*	*7.2 (*)*	*N/A*	*N/A*

(*) Estimated

1.3. Architecture and technologies

EuroHPC JU systems have been designed with the aim to satisfy the computing requirements of a large and diverse range of applications, from traditional computationally intensive arithmetic simulations to data-intensive processing applications commonly incurred for example in the Artificial Intelligence domain (Machine Learning, Deep Learning etc). For these purposes the JU systems need to offer a broad variety of different architectural paradigms striking a balance between diversity and the choice of best-in-class technologies that can facilitate the quick uptake by modern scientific applications.

Almost all the EuroHPC JU systems (with the exception of Discoverer) rely on modular architectures, incorporating a large number of system partitions offering different configurations. Typically, all systems offer a CPU-based partition which does not include GPUs and an Accelerated partition incorporating GPUs or in the case of MeluXina, also FPGAs. In many systems within the CPU partitions there are variations within the nodes as in many cases the nodes incorporate different memory configurations in order to satisfy requirements of applications requiring large, shared memory (fat nodes). In addition, many systems offer partitions for cloud services and data centric applications.

The predominant technology for CPUs is the x86 instruction set, with the notable exception of Deucalion which will offer a large ARM based partition incorporating the Fujitsu A64FX processor. The x86 CPUs are provided by Intel and AMD. The majority of GPU partitions rely on NVIDIA solutions (mostly the A100 Tensor Core GPU) with the exception of LUMI which will offer a large, accelerated partition incorporating next generation AMD GPUs.

Regarding storage, all systems incorporate state of the art multi-tiered, high-performance storage partitions designed to satisfy capacity and performance requirements of modern HPC applications. The predominant file system used by the JU supercomputers is Lustre. It is worth noting that the systems that incorporate a cloud partition, they pair it with a Ceph storage solution to facilitate storage management for Virtual Machines.

Table 3 - Architecture characteristics

System	CPU partition	GPU partition	FPGA partition	Cloud partition	Data Centric partition	Storage
LUMI	x86_64	AMD Instinct	-	Yes	Yes	87 PB Lustre (multi-tiered) & 30 PB Ceph
Leonardo	x86_64	NVIDIA Ampere	-	-	Yes	111.4 PB Lustre (multi-tiered)
Deucalion	ARM & x86_64	NVIDIA A100	-	-	-	11.3 PB Lustre (multi-tiered)
Discoverer	x86_64	-	-	-	-	2 PB Lustre
Karolina	x86_64	NVIDIA A100	-	Yes	Yes	1.3 PB Lustre (all flash)
MeluXina	x86_64	NVIDIA A100	Intel Stratix 10MX	Yes	Yes	20 PB Lustre (multi-tiered) & 96 TB Ceph
Vega	x86_64	NVIDIA A100	-	Yes	-	1 PB Lustre & 19 PB Ceph

All systems rely on InfiniBand HDR (100 or 200) for the application fabric and are connected to GÉANT with at least 100 GB/s links.

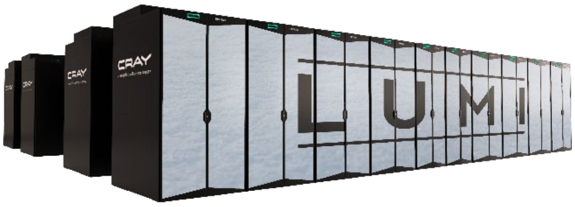
2. Pre-exascale systems

2.1. LUMI

2.1.1. System overview

LUMI will be an HPE Cray EX supercomputer consisting of several partitions targeted for different use cases. The peak performance of LUMI is an astonishing 552 petaflop/s meaning 5.52 *10^{17} double-precision floating point operations per second. This figure makes LUMI one of the world's fastest supercomputers.

- The number-crunching capability of LUMI is primarily due to its GPU partition (LUMI-G). It is based on the future-generation AMD Instinct GPUs. One LUMI-G node will have four AMD MI200 GPUs and one host CPU.
- The GPU partition is complemented by a x86 CPU partition (LUMI-C), featuring 64-core AMD EPYC "Milan" CPUs, 1536 dual-socket nodes i.e. 196 608 cores. 1376 of these nodes will have 256 GB memory, 128 nodes with 512 GB and 32 with 1 TB memory.
- LUMI's data analytics partition (LUMI-D) has 32 aggregated terabytes of memory and 64 visualisation GPUs. This partition is used for interactive use in e.g., visualisation, heavy data analysis, meshing, and pre/post-processing.
- LUMI's storage system will consist of three components. There will be a 7-petabyte partition LUMI-F of ultra-fast flash storage, combined with a more traditional 80-petabyte capacity storage (LUMI-P), both based on the Lustre parallel filesystem, as well as a data management service LUMI-O, based on object storage technology (Ceph), coming in at 30 petabytes in volume. The I/O bandwidths (read or write from/to LUMI-G) are nearly 2 TB/s for LUMI-F and 1 TB/s for LUMI-P.



- LUMI will also have a small OpenShift/Kubernetes container cloud platform for running microservices.
- LUMI's high-speed interconnect is based on HPE Cray's Slingshot technology. The injection bandwidth to a LUMI-G node is 800 Gbit/s and to a LUMI-C node 200 Gbit/s. It will be possible to combine LUMI-G and LUMI-C nodes for the same job.

The system will be installed in two phases. In the first phase in the autumn of 2021, everything except the LUMI-G partition will be installed. The second phase will be installed in the first quarter of 2022, and the system will be generally available after the acceptance of the second phase, during the first quarter of 2022.

The first phase also includes a small-scale porting platform for the AMD GPUs. This porting platform will include hardware that is similar to that which will be used in the GPU partition of the LUMI system, as well as a pre-release of the software stack that will be used for the GPUs.

This platform will be accessible to the end-users and they can carry out porting work on it; it is not intended for actual production workloads.

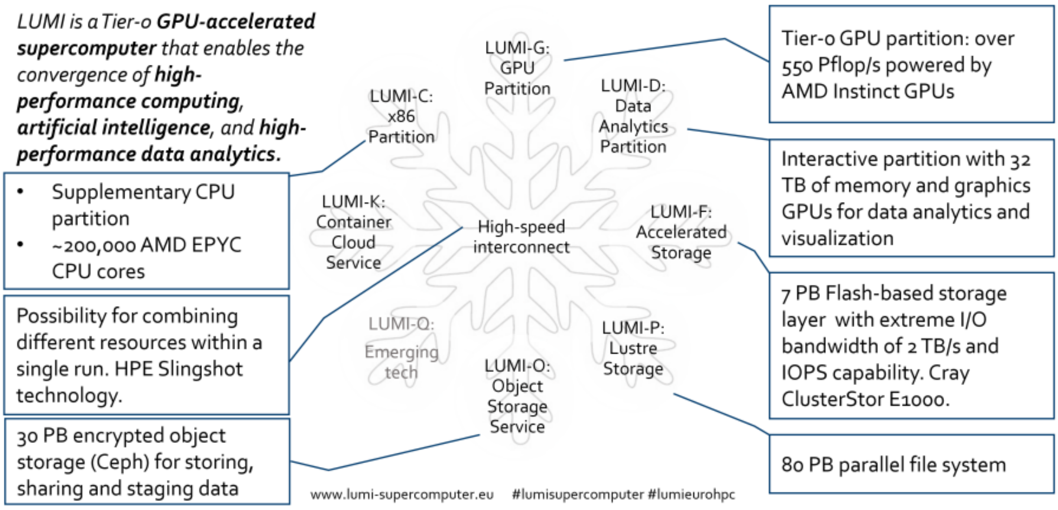

LUMI is a Tier-0 GPU-accelerated supercomputer that enables the convergence of high-performance computing, artificial intelligence, and high-performance data analytics.
• Supplementary CPU partition
• ~200,000 AMD EPYC CPU cores
Possibility for combining different resources within a single run. HPE Slingshot technology.
30 PB encrypted object storage (Ceph) for storing, sharing and staging data
LUMI-G: GPU Partition
LUMI-C: x86 Partition
LUMI-D: Data Analytics Partition
LUMI-K: Container Cloud Service
High-speed interconnect
LUMI-F: Accelerated Storage
LUMI-Q: Emerging tech
LUMI-O: Object Storage Service
LUMI-P: Lustre Storage
Tier-0 GPU partition: over 550 Pflop/s powered by AMD Instinct GPUs
Interactive partition with 32 TB of memory and graphics GPUs for data analytics and visualization
7 PB Flash-based storage layer with extreme I/O bandwidth of 2 TB/s and IOPS capability. Cray ClusterStor E1000.
80 PB parallel file system
www.lumi-supercomputer.eu #lumisupercomputer #lumieurohpc

Figure 2 - The LUMI system at a glance

2.1.2. Programming environment

The AMD programming environment comes under the ROCm (Radeon Open Compute) brand. As the name suggests, it consists mostly of open-source components, which are actively developed and maintained by AMD. The source code is hosted at RadeonOpenCompute/ROCm, and the documentation can be found on the ROCm documentation platform.

The ROCm software stack contains the usual set of accelerated scientific libraries, such as ROCBlas for BLAS functionality and ROCfft for FFT, etc. The AMD software stack also comes with the necessary compilers needed to get code compiled for the GPUs. The AMD GPU compiler will have support for offloading through OpenMP directives. In addition, the ROCm stack comes with HIP, which is AMD's replacement for CUDA, and the tools required to facilitate translating code from CUDA to HIP. The code translated to HIP still works on CUDA hardware.

The ROCm stack also includes the tools needed to debug code running on the GPUs in the form of ROCgdb, AMD's ROCm source-level debugger for Linux based on the GNU Debugger (GDB). For profiling, the ROCm stack also comes with rocProf, implemented on the top of rocProfiler and rocTracer APIs, allowing you to profile code running on the GPUs. AMD provides its MIOpen library, an open-source library for high performance machine learning primitives for machine learning. MIOpen provides optimised hand-tuned implementations such as convolutions, batch normalisations, poolings, activation functions, and recurrent neural networks (RNNs).

In addition to the AMD software stack, LUMI will also come with the full Cray Programming Environment (CPE) stack. The Cray Programming Environment is equipped with the required compilers and tools that help users port, debug, and optimise for GPUs and conventional multi-core CPUs. It also includes fine-tuned scientific libraries that can use the CPU host and the GPU accelerator when executing kernels.

2.1.3. Application domains

LUMI is designed as a 'Swiss army knife' targeted to a wide spectrum of use cases and user communities. We expect to catalyse many different domains, such as

- enabling more precise climate models and the interconnection of different climate models;
- life science: calculation of protein function, structural protein-protein interactions, distributions of binding free energies, and simulations to understand electron transfer, for example;
- artificial intelligence (deep learning): analysing large data sets (simulated and measured) and reanalysing in atmospheric science, environmental science, climate modelling, material science, and linguistics;
- digital humanities and social sciences: large-scale data set analytics from social networks and the modelling of complex societal phenomena; training very large language models.

Furthermore, we expect the fast-track capability for urgent computing to be of importance in addressing in time- and mission-critical simulations e.g., related to national or EU public health or security or other major crises.

2.1.4. Benchmarks

The procurement process of the LUMI system relied on the following list of benchmarks for the evaluation of submitted tenders:

- LUMI-G
    - Synthetic benchmarks: High Performance Linpack (HPL) benchmark, High Performance Conjugate Gradient (HPCG) benchmark, OSU MPI benchmark
    - Application benchmarks: Gromacs, CP2K, ICON, GridTools, MLPerf (Image Classification, Object Detection, Translation)
- LUMI-C
    - Synthetic benchmarks: High Performance Linpack (HPL) benchmark, High Performance Conjugate Gradient (HPCG) benchmark
    - Application benchmarks: Gromacs, CP2K
- LUMI-D
    - Synthetic benchmarks: Graph500 (BFS) benchmark on the CPUs only
- LUMI-F/P
    - Synthetic benchmarks: IO500 bw from both LUMI-G and LUMI-C, IO500 md on both LUMI-P and LUMI-F

2.2. Leonardo

2.2.1. Overview

CINECA will host one of the three EuroHPC JU precursor of exascale systems. The Leonardo supercomputer will be installed on the newly built data centre located in the Bologna Big Data Technopole (henceforth denoted as Technopole). Emilia Romagna region and the Ministry of University and Research established a collaboration in order to promote the Technopole to national and international level. Thanks to this collaboration, the ECMWF decided to relocate its data centre to the Technopole. The combined presence of ECMWF and CINECA data centres will thereby make Bologna Technopole one of the main European hubs for computing and data processing.

CINECA new data centre will follow a two-stage evolution plan. In the first stage (2021-2025) the data centre will feature: 10 MW of IT load, 1240 m^2 of computing room floor space, 900 m^2 of ancillary space, a direct liquid cooling capacity of 8MW, a chilled water (18° -23°) cooling capacity of 6+2 MW and power capacity of 3+1 MW (no-break) and 9+3 MW (short-break). The second stage (2025-2030) will see an increase to 20 MW IT load available and an additional 2600 m2 computing room space floor available, with mechanical and electrical infrastructure able to comply with two different expansion strategies. Stage 2a Liquid Cooling Expansion (16 MW direct liquid cooled + 4 MW air Cooled) or stage 2b Air Cooling Expansion (8 MW direct liquid cooled + 12 MW air cooled). In designing the data centre, particular care was devoted to containing the PUE, estimated to be below 1.1. Simulation of the PUE was based on loads and losses calculated for all systems (IT, mechanical, and electrical) and on Bologna historical external conditions, following a strategy compliant with Level 3 Green Grid/ASHRAE.

CINECA overviewed the technical aspects of the procurement procedure and the design of the supercomputing system architecture, significantly promoting the competition between the candidates and resulting in final offers remarkably better than tender minimum requirements. The result of the procurement procedure is a system capable of nearly 250 PFlops providing 10 to 20 times better time to results with respect the current CINECA flagship system Marconi-100.

This document will briefly cover system technical specifications and performance alongside the targeted usage models better suited to exploit the most out of Leonardo system.

2.2.2. System Architecture

The overall architecture is composed of two main modules:

- a Booster Module, whose purpose is to maximise the computational capacity; it is designed to satisfy the most computational-demanding requirements in terms of TTS, while optimising the ETS. This result is achieved with 3456 nodes, each of them equipped with four NVidia Ampere[1] based GPUs and with two 32-cores Intel Ice Lake CPUs, for a full computational performance of 240.5 PFlops of sustained performance. Details are listed in subsection 2.1.1;
- A General Purpose/Data Centric Module (labelled GP-DC from now on) aims to satisfy a broader range of applications; its 1536 nodes are equipped with two 56-cores Intel Sapphire Rapids CPUs per node in order to reach 8.97 PFlops of sustained performance. Details are listed in section 2.1.2;

[1] Commercial name pending.

These modules are coupled with the following in order to build the whole infrastructure:

- A front-end interface composed of 16 login nodes with 2 Ice Lake CPUs (32 cores each), 512 GB RAM and 6 TB disks in RAID-1 configuration; in addition, 16 additional nodes are equipped with 6.4 TB NVMe disks and two NVidia Quadro RTX8000 48GB to be used as visualisation nodes;
- A fast storage area, with a read performance of 744 GB/s and write performance of 620 GB/s on the top of 106 PB of net capacity; details in section 2.3;
- 200 Gbs Infiniband Network; details in section 2.2;

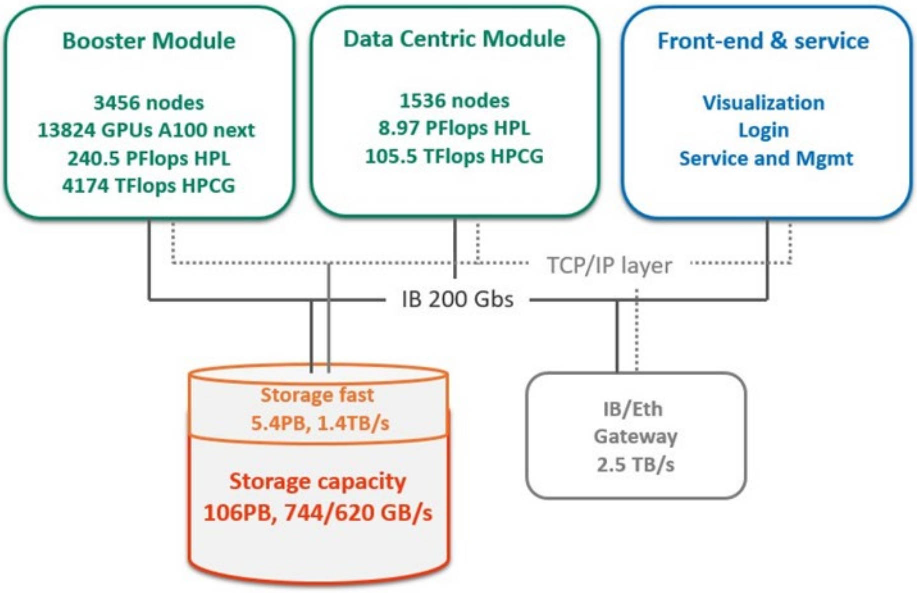


Figure 3 - Leonardo system architecture

Compute node design

Booster module

System specifications:

- Intel Xeon ICP06 (Ice Lake), single socket with 32 core, 2.5 GHz per core, supporting AVX-512 instruction set.
- 256 GB DDR4 RAM, 8 memory channels per socket.
- CX200 ad hoc board from ATOS.
- 360 W of thermal design point.
- 4 NVidia Ampere GPU (SXM).
- CPU-GPU connection via PCIe4 16x connection through HDR Connect6 HCA.
  - PCI passthrough.
  - 16 PCI links towards CPU, 16 links towards GPU.
  - Bandwidth: 64 GBs duplex.
- Full NVLink GPU-GPU connection.
  - 200 GB/s bi-directional.

- No PCI switch between host and external network.
    - Low latency.
- Out-of-band telemetry information.

Data centric module

- Intel Xeon Sapphire Rapids, dual socket supporting AVX-512 instruction set.
- 512 GB DDR5 RAM, 8 memory channels per socket.
- 350 W of thermal design point.

High speed interconnect

The low latency high bandwidth interconnect is based on a Mellanox HDR200 solution and designed as a Dragonfly+ topology. This is a relatively new topology for Infiniband based networks that allows to interconnect a very large number of nodes with a moderate number of switches, while also keeping the network diameter very small. In comparison to non-blocking fat tree topologies costs can be reduced and scaling-out to a larger number of nodes becomes feasible. In comparison to 2:1 blocking fat tree, close to 100% network throughput can be achieved for arbitrary traffic. Dragonfly+ topology features a fat-tree intra-group interconnection, with 2 layers of switches L1 and L2, and an all-to-all inter-group interconnection with a third layer of switches L3.

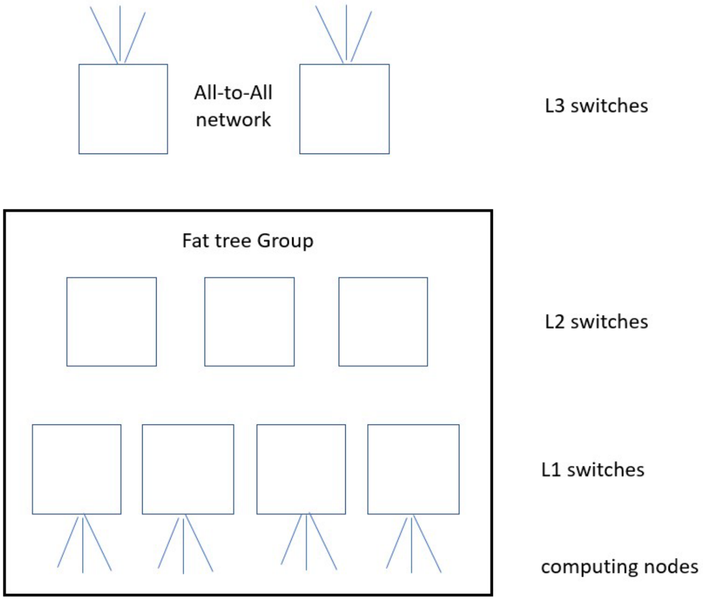


Figure 4 - Leonardo high speed interconnect

For Leonardo, Dragonfly+ is designed as following:

Module	Number of switches			Intra-group blocking	Inter-group blocking
	L1 (downlinks)	L2 (uplinks)	L3		
Booster	18 (20)	18 (22)	18	1.11:1	non-blocking
Data Centric	18 (16)	18 (22)	2	non-blocking	non-blocking

The solution comes with improved adaptive routing support, which is crucial for facilitating high bisection bandwidth through non-minimal routing. In fact, intra-group routing and inter-group routing need to be balanced to provide low hops count and high network throughput. This is obtained with routing decisions evaluated in every router on the packet's path and allows a minimum network throughput of ~ 50%.

I/O

The storage system features a capacity and fast tier. This architecture allows great flexibility and the ability to address even the most demanding IO use cases, in terms of bandwidth and IOPS. The storage architecture, in conjunction with the booster compute node design and its GPUDirect capability, allows in principle to improve IO bandwidth and reduce IO latency towards the GPUs, therefore improving the performance for a significant number of use cases.

Fast tier

It is based on DDN Exascaler and will act as a high-performance tier specifically designed to support high IOPS workloads. This storage tier is completely full flash and based on NVMe and SSD disks therefore providing high metadata performance especially critical for AI workloads and in general when the creation of a large number of files is required. A wide set of options are available in order to integrate fast and capacity tier in order to make them available to end users.

Leonardo will feature a fast tier with the following characteristics:

Storage Fast Tier	
Net capacity	5.4 PB
Disk technology	full flash (NVMe and SSD)
Bandwidth:	Aggregated: 1400 GB/s r/w io500[2] score: 676

Capacity Tier

The capacity storage will provide the parallel filesystem which will be based on Lustre. Some key security features that will be evaluated are: i) Lustre multi-tenancy, available from version 2.10, aiming to improve security and isolation so that only authenticated users can access a selected portion of the storage namespace: ii) Lustre encryption at rest, available from version 2.14, based on cryptofs, transparent to the parallel filesystem and able to handle file multiple access (multiple client).

Leonardo will feature a fast tier with the following characteristics:

Storage Capacity Tier	
Net capacity	106 PB
Disk technology	NVMe and HDD
Bandwidth:	Aggregated: read performance of 744GB/s and write performance of

[2] io500 rules can be found at https://www.vi4io.org/io500/rules/start

	620GB/s io500[3]: 197 GiB/s

2.2.3. Software ecosystem

All the nodes are equipped with Red Hat Enterprise Linux 8; the provided scheduler is Slurm. In addition, the vendor will provide the system with some preinstalled software:

- Several compiler suites are pre-installed: Intel Parallel Studio Cluster Edition in order to optimise the performances on CPU-based applications, and Nvidia HPC SDK in order to take advantage of all the GPU features on the Booster Module;
- Debugger and profilers are included. Particularly the Arm Forge Ultimate Suite brings Arm DDT for CPU and GPU debugging will be installed, as well as Arm MAP and Arm Performance Reports for profiling and the Intel Parallel Studio brings Intel Debugger too. As for the Nvidia tools, also the graphical tool Nsight System is provided for profiling.
- Some of the most important optimised numerical libraries are given through the Intel MKL library;
- Containerisation is supported through several different tools:
  - Atos Containers Framework allows management of containers via Singularity;
  - Nvidia Container Framework enhance containers creation and automated deployment;
  - Slurm integration is improved via Pyxis;
  - ParTec Parastation also supports the execution of containerized applications, improving the flexibility of a pure Singularity approach;
- Monitoring is granted via Atos SMC xScale suite, based on Prometheus, and using Grafana as frontend;
- Detection and tracking of issues is performed by Parastation HealthChecker;
- Additional software will be provided by CINECA staff in order to satisfy requirements from scientific and industrial communities running on the machine.

2.2.4. Energy Efficiency

The vendor will provide two different software tools which grants a dynamical adjustment of power consumption: Bull Energy Optimiser keeps track of energy and temperature profiles via IPMI and SNMP protocols. Such tools can interact with Slurm scheduler in order to tune some of its specific features, like a selection of the jobs based (also) on the expected power consumption or a dynamical capping of the CPUs frequencies based on the overall consumption.

This dynamical tuning procedure is enhanced by a second tool called Bull Dynamic Power Optimiser, which monitors consumptions core by core in order to cap frequencies to the value which grants optimal balance between energy saving and time degradation of the running application.

[3] io500 rules can be found at https://www.vi4io.org/io500/rules/start

About the GPU consumptions, NVIDIA Data Centre GPU Manager is provided, which throttles the GPUs clock when it overcomes a custom threshold. Atos also claims to obtain a CPU energy saving of ~17% with a related time degradation of ~2.8%.

2.2.5. System performance and targeted use models

Scalable and throughput computing

Leonardo is designed to be as a general-purpose architecture in order to serve the needs of all scientific communities and to satisfy the needs of R&D industrial customers. Scalable and high throughput computing typically refer to scientific use cases that require large amount of computational resources either through highly parallel jobs on large scale HPC architectures or by launching a large number of jobs to evaluate different parameters. Leonardo system is expected to support both models by providing a tremendous speed-up for workloads able to exploit accelerators. Leveraging the booster architecture described in section 2, early benchmarks figures report a 15-30x time-to-science improvement for applications already ported on NVIDIA GPUs (QuantumEspresso, Specfem3D_Globe, Milc QCD) compared to CINECA Tier-0 system Marconi-100, currently 11th in November Top500 list. Examples of applications used in production on Marconi-100 (NVIDIA V100 based) can be found here: https://www.hpc.cineca.it/hardware/marconi100. The number of applications able to run on GPUs is increasing day by day, thanks to the diffusion of the existent programming paradigms and to a more supporting ecosystem (EU Centres of Excellence, Hackathons, local support of the computing centres).

AI-based applications can leverage state-of-the-art GPUs providing large low precision peak performance, dedicated Tensor cores and a system architecture designed to support I/O bound workloads thanks also to GPUDirect RDMA feature and IO fast tier caching.

Non-traditional HPC usage models: interactive and urgent computing

CINECA will foster “non-traditional” HPC usage models on Leonardo. End-users requiring interactive computing sessions will benefit from the experience gained within the Fenix-ICEI project’s system architecture features including visualization and data centric partitions. Interactive computing aims to provide a tightly integrated HPC computing experience enabling: a) access to capable computing resources such as state-of-art servers; b) access to data produced on HPC computing resources; c) access to capable storage infrastructure (high IOPS, high bandwidth and reduced latency) for data intensive workloads; d) access to remote desktop for GUI applications; e) access to interactive frameworks (e.g. Jupyter notebook). Various factors motivate the need of interactive computing from the users’ perspective. Some of them are: i) the need for real-time interaction with a program runtime or running simulations, ii) the need to estimate the state of a program or its future tendency, iii) the need to access intermediate results. iv) the need to steer the computation by modifying input parameters or boundary conditions as more input data (live data) comes into play. We expect this functionality to be more and more relevant the more data are produced with the supercomputing system.

A high return value use case is the so-called urgent computing typically required in extreme cases or conditions. An example of such type was recently conducted in CINECA within the Excalate4cov project with the aim to process a huge ligand library containing 70 billion molecules against the 15 active sites of COVID-19 virus to find the most promising interactions. Even if CINECA has a proven track record in HPC operations, the ambition here is to be able to serve such use cases on-demand and with major time constrains. All the software ecosystem, from the application software stack, its deployment model, the productivity environment including the batch scheduler and system health tools play a major role and need to be sufficiently flexible and ready for the scope.

Quantum computing emulators

CINECA HPC system Leonardo will play a fundamental role in testing and benchmarking quantum algorithms and devices. Thanks to the high amount of RAM per node, it will be possible to exactly emulate quantum algorithms of about 35 qubits using a single node and up to 45 qubits if the entire Leonardo system were to be used. Moreover, exploiting new techniques such as Tensor Networks, a single Leonardo node will be able to approximate quantum algorithms on hundreds of qubits. Such emulations will be highly accelerated by the 4 GPUs available on each node.

2.2.6. Benchmarks

The procurement process of the Leonardo system relied on the following list of benchmarks for the evaluation of submitted tenders:

Synthetic benchmarks:

- HPL and HPCG

Application benchmarks:

- QuantumEspresso
- SFPECFEM3D_Globe
- PLUTO
- MILC

3. Petascale systems

3.1. Deucalion

3.1.1. System Architecture

The Deucalion supercomputer consists of two main general purpose compute partitions based on different processor architectures and one GPU-based compute partition. Users have access to the system through a set of front-end servers. All nodes are connected to a high-performance shared storage through an Infiniband high-speed interconnect and to an Ethernet network.

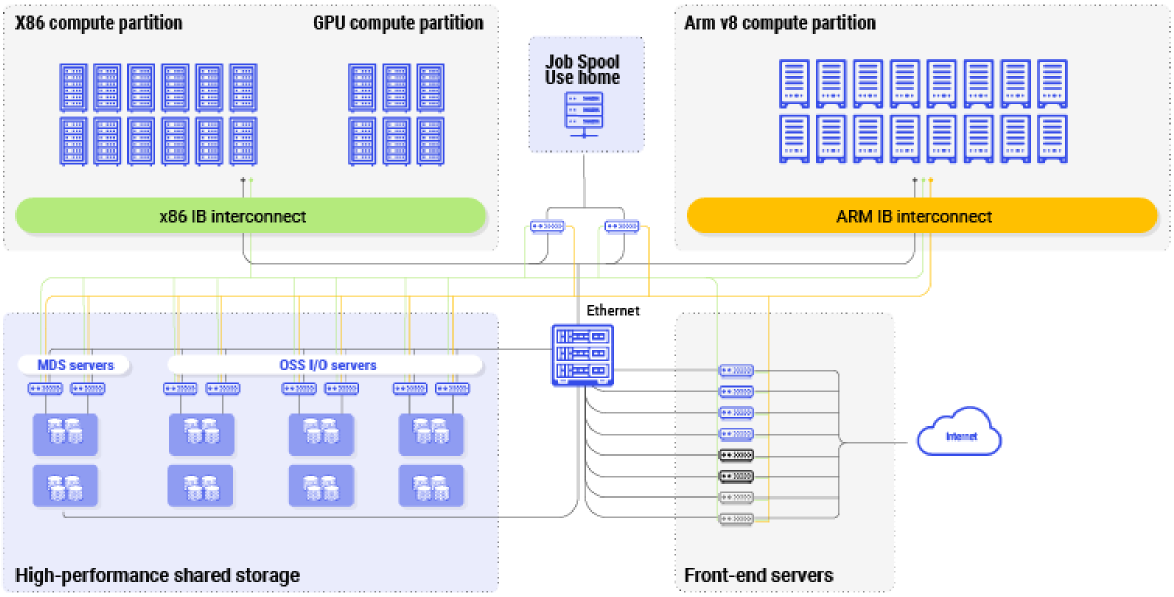


Figure 5: The Deucalion supercomputer architecture

Compute partitions

Deucalion's largest compute partition is based on the ARMv8-A architecture and consists of 1632 Fujitsu PRIMEHPC FX700 nodes equipped with Fujitsu A64FX 2.0GHz processors. Each node offers 16GB of in-processor High Bandwidth Memory. The partition peaks at 5PF of double precision performance.

The second compute partition is based on the x86 architecture and consists of 500 ATOS Bull Sequana X440 dual processor nodes equipped with AMD EPYC Rome 7742 (2.25 GHz) processors. Each node offers 256GB of RAM. The partition peaks at 2.3PF of double precision performance.

The GPU-based compute partition consists of 33 ATOS Bull Sequana E410nodes equipped with four NVIDIA A100 GPUs and two AMD EPYC Rome 7742 (2.25 GHz) processors. Each node offers 80 GB of High Bandwidth Memory per GPU and 512GB of RAM. The partition peaks at 2.7PF of double precision performance.

Shared storage

All nodes have access to a high-performance shared storage based on the DDN EXAScaler subsystem using ES400VNX controllers. The storage offers a 10.6PB parallel filesystem and a 430TB NVM multi-purpose high-speed layer. An additional NetAPP All Flash A220 network attached storage with 70TB is available for user homes and spooling.

Interconnect network

The interconnect network is built on Infiniband HDR technology based on Mellanox QM8790 switches.

3.1.2. Programming environment

All standard HPC programming models are supported. This includes C, C++ and Fortran compilers from various vendors, including AMD, Intel, PGI, GNU and LLVM. The multicore programming is available via OpenMP support, the AMD ROCm and Intel ecosystems including the IPP and TBB. Further various mathematical libraries with multicore support are provided. The multi-node programming model is supported via MPI, GASPI and similar message passing technology. The accelerator programming is supported primarily via the CUDA API, the OpenACC and the OpenMP offloading, with other technologies such as OpenCL also available. The ROCm HIP is also available. Optimized tools, libraries and compilers for the ARM ecosystem are provided by Fujitsu and made available for native and cross-compiled scenarios.

3.1.3. Application domains

Deucalion was designed to stand out in several high-performance computational workloads, both from academia and industrial workloads, mainly focused on high-performance, low-power architectures. The emphasis was given to integrating low-power processor (FUJITSU A64FX processor) partition as the leading resource alongside the traditional X86 CPU and GPU accelerated partitions. Overall the architecture provides high throughput for most scientific and industrial workloads on a generic programming and computing platform while providing a testbed for low-power architectures critical for exascale computing.

The Singularity framework provides security, isolation, and ease of use within the supercomputing infrastructure containerization, allowing users to deploy an application within a self-contained environment, even using legacy software stacks if needed. The option of deploying traditional scheduling and software stack provided by the OpenHPC project emphasizes the supercomputer's generic nature and provides excellent support for the most traditional HPC applications. Furthermore, the software stack provided focuses on open-science and open-source software such as computational fluid-dynamics (OpenFOAM), molecular dynamics (GROMACS), Genomics, Oil-and-gas (Reservoir Simulation), ML/AI/DL workloads (PyTorch and TensorFlow both on generic architecture and GPU accelerated) as well as scientific visualization workloads (Paraview/Visit).

3.1.4. Benchmarks

The procurement process of the Deucalion system relied on the following list of benchmarks for the evaluation of submitted tenders:

Synthetic benchmarks

- HPC Challenge Suite
- IO500

Application Benchmarks

- ALYA

- CP2K
- GROMACS
- NEMO
- SPECFEM3D
- ResNet-50
- OpenNMT

3.2. Discoverer

3.2.1. System overview

- Supercomputer model: ATOS BullSequana XH2000 ;
- Double precision peak performance (Rpeak): 6+ PetaFlops ;
- Linpack Benchmark performance (Rmax): 4.45 PetaFlops
- % DP TeraFlop/s peak vs Linpack: 0.74;
- Total number of compute nodes: 1128 ;
- Number of Fat Nodes: 18 (included in the above total number of compute nodes}
- Total number of processors: 2256 ;
- Total number of cores: 144384;
- Total main memory capacity: 302592 GB ;

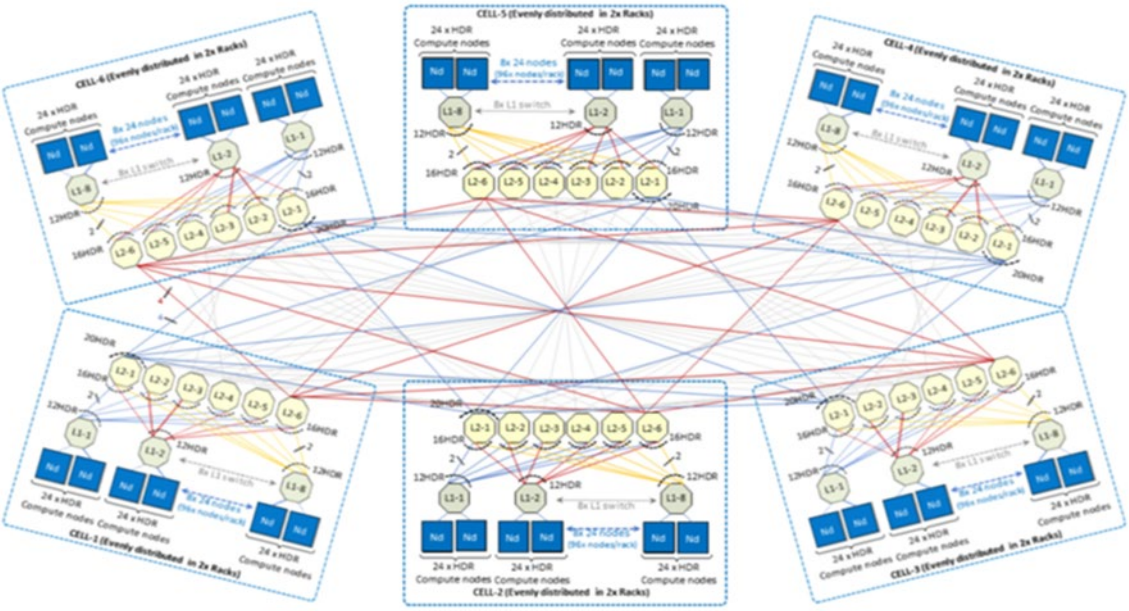
Compute node design

- CPU model: AMD EPYC 7H12, 64core, 2.6GHz, 280W; Next generation x86 “Zen2”
- CPU sockets per node: 2 ;
- CPU Cores per node: 128;
- Main memory per node : 256GB (Each of the 18x Fat nodes has 1024GB Memory)
- Memory type and frequency: 16GB DDR4 RDIMM 3200MT/s DR; (The fat nodes are equipped with 64GB DDR4 RDIMM 3200MT/s DR)
- Node DP TeraFlop/s peak: 5.325TFlops
- % DP TeraFlop/s peak vs Linpack: 74% ;
- TFlop/s sustained Linpack: 3.940TFlops;
- Linpack node power consumption: 665.1 W per 256 GB compute node; 747.0 W per Fat compute node (Cooling subsystem power consumption excluded);
- Number and bandwidth of network interfaces: 1x 200Gbps HDR;

High performance Network

- Interconnect family: IB HDR;
- Interconnect bandwidth per link: 200Gbps (IB HDR) ;
- Expected latency (worst case for a 1 kB message) : 520ns ;

- Interconnect topology: DragonFly+ ;
- Number of compute nodes per isle (2 Racks) : 192 ;
- Blocking factor within isle: 2:1;
- Number of links to I/O partition: 120;


CELL-6 (Evenly distributed in 2x Racks)
CELL-5 (Evenly distributed in 2x Racks)
CELL-4 (Evenly distributed in 2x Racks)
CELL-1 (Evenly distributed in 2x Racks)
CELL-2 (Evenly distributed in 2x Racks)
CELL-3 (Evenly distributed in 2x Racks)
24 x HDR Compute nodes
Nd
8x 24 nodes (96x nodes/rack)
8x L1 switch
L1-8
L1-2
L1-1
L2-6
L2-5
L2-4
L2-3
L2-2
L2-1
12HDR
16HDR
20HDR
2

Performance:

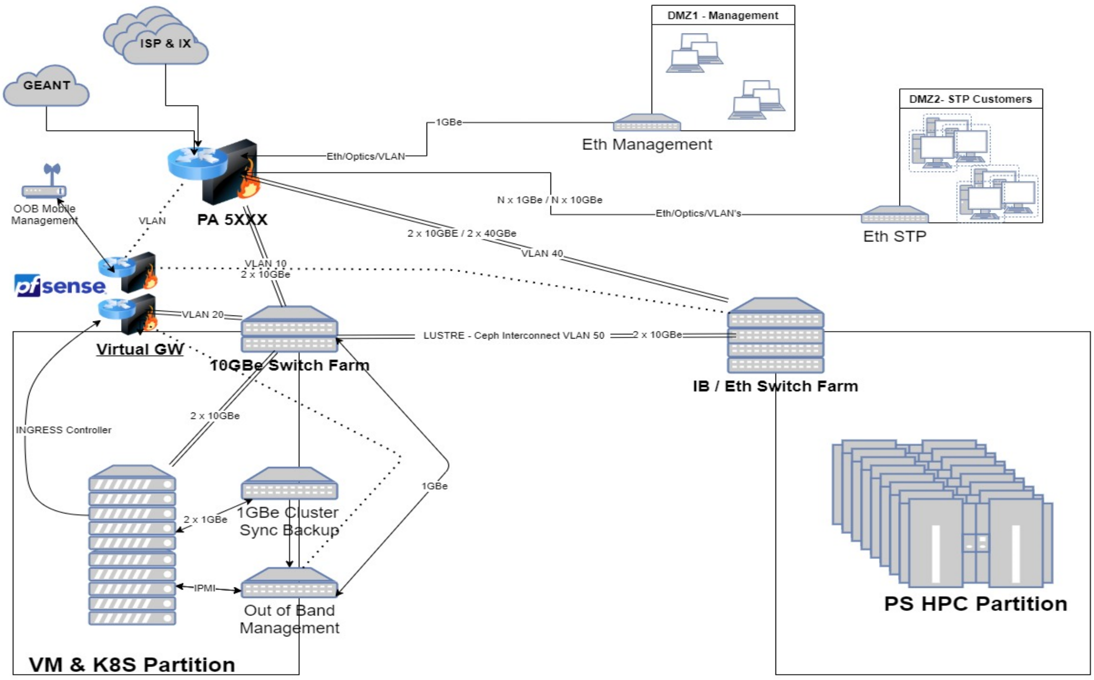
- 40 X HDR 200Gb/s ports in a 1U switch
- 80 X HDR100 100Gb/s ports in a 1U switch
- 16Tb/s aggregate switch throughput
- Up to 15.8 billion messages-per-second
- 130ns switch latency

Management Network

- Network family: Ethernet ;
- Network bandwidth: 10GbE/1GbE ;

Public Access

- Number of login nodes: 2;
- Links and bandwidth per login to external network: 2x 10GbE ;



3.2.2. Software Environment

System Software

- Operating system: Red Hat Enterprise Linux 8;
- Compiler Suite(s): AMD Optimizing C/C++ Compiler ; AMD Open64 SDK; x86 Open64 Compiler System;

Numerical libraries

- AMD Optimizing CPU Libraries (AOCL);

Debugging/ profiler tools

- AMD μProf; IO Instrumentation (IOI) ;Lightweight Profiler (LWP) ;Modular End-of-Job Report (MEJR);

Resource and workload manager

- SLURM; Singularity;

Monitoring

- Bull Energy Optimizer (BEO); Dynamic Power Optimizer (BDPO); Interconnect Management Suite (IMS); Prometheus

3.2.3. Storage system

I/O Partition Summary

- Total net capacity of data: 2 031.89 TB;
- Total addressable capacity (no file system);
- Total net capacity for metadata storage + home/apps - 15.25 TB; Total addressable capacity (no file system), this is for parallel file system metadata only. User home folders and application binaries will be in Data; The useable capacity of the filesystem will be 1 to 2% below these numbers area;
- Aggregated performance (TB/s): 20 GB/s;
- Number of data modules: 164 HDD ;
- Number of metadata modules: 11 SSD ;

Data Module

- Net capacity provided (PB): 2.03 PB ;
- Performance provided (GB/s): 20 GB/s;
- Number of storage elements: 164 + 11;
- Type of storage element: HDD and SSD;
- Size (TB) per storage element: 6TB + 1.92TB ;
- CPU Cores per server: 10 ;
- Main memory per server: 150 GB ;
- Memory type and frequency: DDR4 2666 MT/s ;
- Number and bandwidth interfaces to control data network: 4 x GigE RJ45 for OS access and hardware management;
- Number and bandwidth interfaces to bulk data network (RDMA): 4x HDR100 IB / 100GbE ports (same ports as for metadata);

Metadata Module

- Net capacity provided (PB): 0.01PB
- Performance provided (GB/s) : 50K file create/s ;
- Number of storage elements: 11
- Type of storage element: SSD
- Size (TB) per storage element: 1.92;
- Number of storage servers: 2;
- CPU Cores per server: 10;
- Main memory per server: 150 GB;
- Memory type and frequency: - DDR4 2666 MT/s;

- Number and bandwidth interfaces to control data network: 4 x GigE RJ45 -> for OS access and hardware management;
- Number and bandwidth interfaces to bulk data network (RDMA): 4x HDR100 IB;

3.2.4. Target Applications:

The consortium, managing the supercomputer have experience in the following fields:

- Bioinformatics and Genomics
- Computational Chemistry
- Molecular Dynamics, Molecular Mechanics and Molecular interactions
- Quantum Chemistry
- In silico Drug Discovery
- 3D Protein Structure Prediction
- Seismic Wave Impact Simulation
- Computational Fluid Dynamics
- Finite Element Computer Simulation
- Monte Carlo Simulations

3.2.5. Benchmarks

The procurement process if the Discoverer system relied on the following list of benchmarks for the evaluation of submitted tenders:

Synthetic benchmarks:

- MPI benchmark: latency
- MPI benchmark: bandwidth
- Memory benchmark
- Multi-flow IP benchmark
- Graph500 BFS
- Graph500 BFS (incl. NVM storage)
- High Performance Conjugate Gradient (HPCG)
- High Performance Linpack (HPL)

Application benchmarks:

- GROMACS
- NAMD
- OpenFOAM

- Quantum Espresso
- ALYA

3.3. Karolina

3.3.1. System Overview

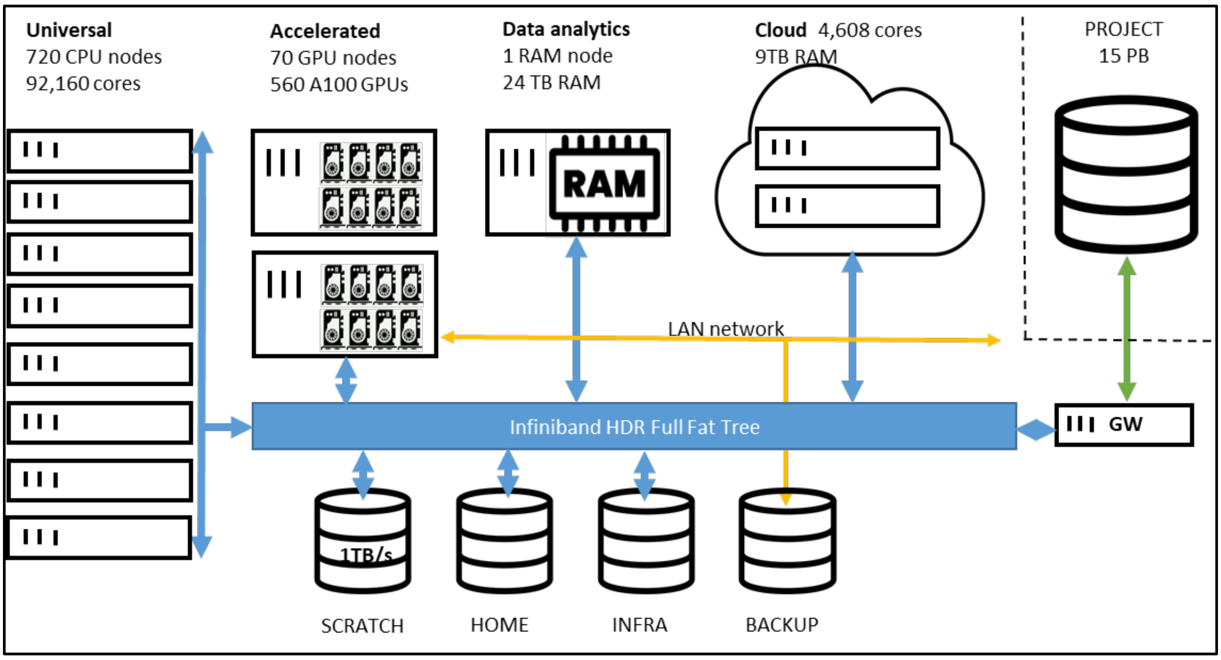


Figure 6 - Karolina supercomputer architecture

The Karolina supercomputer consists of four major compute partitions. The Universal partition, the Accelerated partition, the Data analytics partition and the Cloud partition. The partitions are connected to the high-performance SCRATCH storage and the auxiliary storages HOME, INFRA and BACKUP via the Infiniband HDR Compute network. All components are also interconnected by Ethernet LAN Management network. The system is connected to external PROJECT storage via dedicated data gateways.

Universal partition

The Universal partition consists of 720 nodes. Every node features 2 AMD EPYC processors, 128 cores and 256GB of memory per node. The nodes are connected to the Infiniband HDR network at 100Gb/s rate. Via the network, nodes can access the SCRATCH, HOME and PROJECT storage. The partition provides 2.8PF of double precision performance.

- Nodes and performance: 720 nodes, 2.8PF
- CPU 2x AMD EPYC 7H12, 2x64 cores
- RAM 256GB RAM DDR4
- Interconnect 100Gb/s Infiniband HDR100

Accelerated partition

The Accelerated partition consists of 72 nodes fitted with 576 Nvidia A100 GPU accelerators. Every node contains 8 Nvidia A100 GPUs with 40GB of HBM2 memory, attached via Gen4 PCIe bus. The 8 accelerators are interconnected by an NVLINK2 fabric featuring NVSwitch technology. This enables 320GB of HBM2 memory addressable across the accelerators in unified virtual address space. The nodes are connected to the Infiniband HDR network with 4x200Gb/s links to achieve very high throughput to the network and the SCRATCH storage, necessary for distributed machine learning applications. The partition provides 6.6PF of LINPACK performance.

- Nodes and performance: 72 nodes, 6.6PF
- CPU: 2x AMD EPYC 7763, 2x64 cores
- RAM: 1TB RAM DDR4
- Interconnect: 4x200Gb/s Infiniband HDR
- Accelerator: 8xNvidia A100, 40GB HBM2 memory

Data analytics partition

The Data analytics partition consists of single shared memory node. The HPE Superdome Flex node is fitted with 32 Intel Xeon high end processors providing 768 cores and 24 576 GB of shared DDR4 RAM memory. The processors are interconnected by an internal NUMA network in all-to-all topology, mitigating the NUMA effect. The node is connected to the Infiniband HDR network with 2x200Gb/s links to achieve high throughput to the network and the storage, necessary data analytics workloads. The partition provides 41TF of LINPACK performance.

- Nodes and performance: 1 node, 41TF
- CPU 32x Intel Xeon 8268, 768 cores
- RAM 24 576 GB RAM DDR4
- Interconnect 2x200Gb/s Infiniband HDR

Cloud partition

The cloud partition consists of 36 nodes providing 4608 cores and 9TB of RAM. The nodes are further equipped with 960GB of local NVMe storage. The nodes are connected to the Infiniband HDR network at 100Gb/s rate. Via the network, nodes can access the SCRATCH, HOME and PROJECT storage. The cloud partition nodes are also connected to a dedicated, independent Ethernet LAN network to support on demand networking in cloud environment. The partition provides 142TF of LINPACK performance.

- Nodes and performance: 36 nodes, 142TF
- CPU 2x AMD EPYC 7H12, 2x64 cores
- RAM 256 GB RAM DDR4
- Interconnect 100Gb/s Infiniband HDR100 + 2x10Gb/s Ethernet
- Local storage 960 GB NVMe

SCRATCH storage

The high-performance SCRATCH storage features very high throughput of 1200GB/s at 1330TB of storage capacity. To achieve those performance figures, all-Flash storage array ClusterStor E1000 is deployed. The ClusterStor E1000 is configured from three storage building blocks. An MDU which contains 2 active Lustre MDSes, SSUs each running 2 active OSSes and single SMU system management unit that manage the entire storage solution in an active/passive mode. The system is installed and configured with the Lustre parallel distributed filesystem.

Compute network

The Compute network is built on Infiniband HDR technology configured in non-blocking Fat Tree topology with 40 Leaf HDR switches and 20 Spine HDR switches. The network interconnects all partitions, storages and other components of the supercomputer at 100Gb/s or at multiples of 200Gb/s link speed.

Other components

Besides the main components listed above, the supercomputer contains Login nodes, Visualization nodes, Data management nodes, Data gateways, HOME, INFRA and BACKUP storages, as well as Infrastructure and management nodes and Management LAN network. These components are necessary for operation of the supercomputer and provisioning of computing services and data access.

3.3.2. Programming environment

All standard HPC programming models are supported. This includes C, C++ and Fortran compilers from various vendors, including AMD, Intel, PGI, GNU and LLVM. The multicore programming is available via OpenMP support, the AMD ROCm ecosystem and Intel ecosystem including the IPP and TBB. Further various mathematical libraries with multicore support are provided. The multi-node programming model is supported via MPI, GASPI and similar message passing technology. The accelerator programming is supported primarily via the CUDA API, the OpenACC and the OpenMP offloading, with other technologies such as OpenCL also available. The ROCm HIP is also available, providing accelerated code interoperability with the EuroHPC LUMI supercomputer.

3.3.3. Application domains

The Karolina supercomputer is designed to stand out in complex scientific and industrial workloads in particular the machine learning and artificial intelligence computations. This is reflected in the hardware setup, where besides the standard CPU based partition emphasis was given on integration of heavily GPU accelerated partition, high speed all-flash storage and high memory Data analytics partition supported by a closely integrated Cloud partition. This setup provides strong and balanced performance for most workloads encountered in scientific and industrial computing, with special advantage for multi-GPU accelerated applications.

To provide simple and intuitive access to the supercomputing infrastructure an in-house application framework called HEAppE (High-End Application Execution) Middleware has been developed. The HEAppE is IT4Innovations implementation of the HPC-as-a-Service concept, which provides an abstraction layer that enables the users to run pre-prepared workflows without the difficulty of command-line access. The in-house HyperLoom/HyperQueue applications allow the users to efficiently schedule large complex workflows with dependencies on the supercomputer. The IT4I research labs also develop in-house highly parallel ESPRESO framework which enables KAROLINA users to solve real world multiphysics engineering problems consisting of hundred billions of unknowns that enables to tackle recent computational challenges in both science and industry. The ESPRESO leverages development of automatized workflows for simulation of industrial digital twins and other real world problems. It is distributed under the free and open-source license which allows to create new Solver-as-a-Service solutions and thus make HPC technologies more affordable for industrial users especially SMEs.

3.3.4. Benchmarks

The procurement process if the Karolina system relied on the following list of benchmarks for the evaluation of submitted tenders:

Universal Partition:

- Top500 (HPL)
- HPCG
- Graph500
- Green500
- IO500

Accelerated Partition:

- Top500 (HPL)
- HPCG
- Graph500
- Green500
- IO500
- MLPerf

Data Analytics Partition:

- Top500 (HPL)
- HPCG
- Graph500

3.4. MeluXina

3.4.1. System architecture

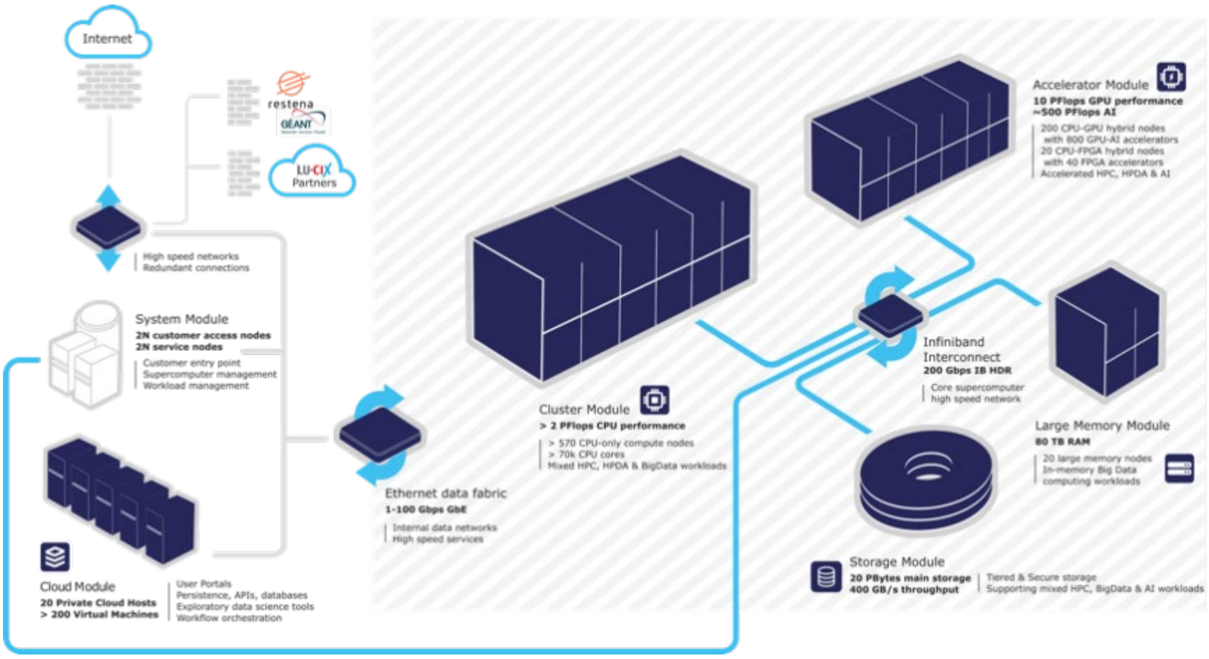


Figure 7 - MeluXina supercomputer architecture

Hardware specifications – Compute environment

- **Cluster Module**: 573 CPU nodes, each with:
  - CPU: 2x AMD EPYC Rome 7H12 (64cores @ 2.6 GHz, 128 physical cores total)
  - RAM: 512 GB DDR4-3200
  - Interconnect: 1x HDR (200 Gbps InfiniBand)
  - No local storage
- **Accelerator Module (GPU):** 200 CPU-GPU hybrid nodes, each with:
  - CPU: 2x AMD EPYC Rome 7452 (2x 32cores @ 2.35 GHz, 64 physical cores total)
  - RAM: 512 GB DDR4-3200
  - Accelerator: 4x Nvidia Ampere A100-40 (40GB HBM, NVlink)
  - Interconnect: 2x HDR (200 Gbps InfiniBand, 400Gbps in dual-rail)
  - Local storage: 1.92TB SSD
- **Accelerator Module (FPGA):** 20 CPU-FPGA hybrid nodes, each with:
  - CPU: 2x AMD EPYC Rome 7452 (32-cores @ 2.35 GHz, 64 physical cores total)
  - RAM: 512 GB RAM DDR4-3200
  - Accelerator: 2x Intel Stratix 10MX (16GB HBM)
  - Interconnect: 1x HDR (200Gbps InfiniBand)

  - Local storage: 1.92TB SSD
- **Large Memory Module:** 20 CPU nodes with extended memory capacity, each with:
  - CPU: 2x AMD EPYC Rome 7H12 (64-cores @ 2.6GHz, 128 physical cores total)
  - RAM: 4 TB RAM DDR4-3200
  - Interconnect: 1x HDR (200GBps InfiniBand)
  - Local storage: 1.92TB NVMe

Hardware specifications - Data storage environment

- Tier-0 storage (“Scratch”):
  - Total size, performance: 0.5 PetaBytes, 400 GB/s read-write throughput
  - Filesystem: Lustre, backed by SED (Self Encrypting Drives) NVMe storage
  - Connectivity: InfiniBand HDR
  - Intended usage: “scratch” directories; highly intensive, short term IO workloads
- Tier-1 storage (“Project”):
  - Total size, performance: 12 PB, 190 GB/s read-write throughput
  - Filesystem: Lustre, backed by SED HDDs for data, SED NVMe storage for metadata
  - Connectivity: InfiniBand HDR
  - Intended usage: user home and project directories; intensive, project-length IO workloads
- Tier-2 storage (“Backup”):
  - Total size, performance: 7.5 PB, 30 GB/s
  - File system: Lustre, backed by SED HDDs for data, SED NVMe storage for metadata
  - Connectivity: InfiniBand HDR
  - Intended usage: project backups for requesting projects
- Tier-3 storage (“Archive”), hosted off-site:
  - Total size: 5PB Tape Library
  - Filesystem: LTFS
  - Intended usage: long-term storage of data for requesting projects
  - Protocol export capabilities:
  - 20 service servers able to re-export the Tier-1,2,3 filesystems for Cloud and object storage access (S3 gateway)
  - Connectivity: InfiniBand HDR and 100Gbps

Hardware specifications – Cloud environment

- **Cloud Module (RedHat OpenStack)**: 20 virtualization hosts, each with:

  - CPU: 2 x AMD EPYC Rome 7H12 (64-cores @ 2.6 GHz, 128 physical cores total)
  - RAM: 512 GB DDR4-3200
  - Capability to host multiple tenants, with tens to hundreds of VMs for APIs, User Portals, Workflow and Orchestration tools.
- **Cloud Module CEPH storage**:
  - Total size: 96TB with replication factor 3 (288TB raw)
  - Capability to provide both block-level and object storage to the VMs

Network connectivity – internal fabric & external networks

Infiniband HDR 200Gbps interconnect:

- Topology: DragonFly+ with 7 groups (cells) of nodes
  - each cell is a non-blocking 2-layer fat tree
  - all-to-all connections between the cells
- Blocking factor:
  - Cluster Module: 2:1
  - Accelerator Module: 1.33:1
- Bisection bandwidth: 76.8Tbps (bidirectional)

External networks:

- GEANT connectivity:
  - 2x 100Gbps connections through the Restena NREN
- Internet connectivity:
  - High bandwidth redundant connections to the national Internet Exchange Point LU-CIX
  - High bandwidth redundant ISP connections

3.4.2. Software environment

Initial composition of the user software stacks on MeluXina's Compute Environment:

- **Compute node OS**: RHEL8/CentOS8 compatible
- **Comprehensive software stack**: based on EasyBuild scientific software delivery toolkit
- Possibility to deploy user tools also using Spack
- **Compilers and SDKs**: Intel Compilers (OneAPI), GCC, AOCC, NVIDIA HPC SDK (including PGI compilers), Intel OpenCL SDK
- **MPI suites**: OpenMPI, IntelMPI, ParaStationMPI
- **Programming languages**: Python, R, Julia, Go, Fortran, C/C++, Scala
- **Numerical, data and parallel/accelerator libraries**: BLAS, LAPACK/ScaLAPACK, MKL, BLIS, FFTW, HDF5, netCDF, Eigen, ARPACK, CUDA, cuDNN, TensorRT, KOKKOS, NCCL, Intel TBB
- **Performance tools**: ARM Forge, Intel ITAC/VTune/Advisor/Inspector, GDB, Valgrind, NVIDIA NSight

- **Frameworks**: PyTorch, TensorFlow, Horovod, Keras, Spark
- **Visualisation**: VMD, ParaView, VisIT; XCS Portal for accelerated remote visualisation
- **Container system**: Singularity

3.4.3. Application domains

LuxProvide has been established as Luxembourg's national HPC centre with missions to provide high performance computing capabilities, high-speed connectivity and advanced applications on a national, European and international scale, serving public and private sector actors.

MeluXina, the first national supercomputer of Luxembourg, has been designed as a modular system able to accommodate a large variety of workloads, from HPC to HDPA and AI.

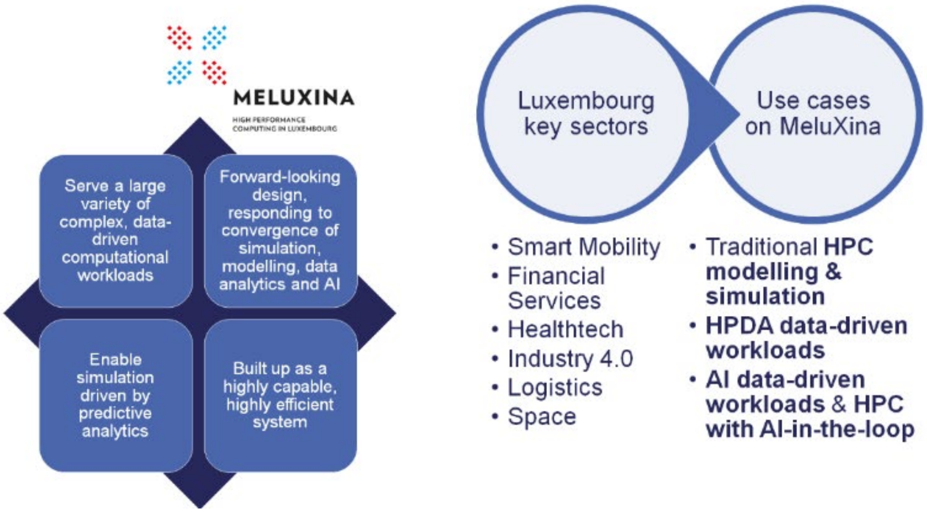

MELUXINA
HIGH PERFORMANCE COMPUTING IN LUXEMBOURG
Serve a large variety of complex, data-driven computational workloads
Forward-looking design, responding to convergence of simulation, modelling, data analytics and AI
Enable simulation driven by predictive analytics
Built up as a highly capable, highly efficient system
Luxembourg key sectors
Use cases on MeluXina
• Smart Mobility
• Financial Services
• Healthtech
• Industry 4.0
• Logistics
• Space
• Traditional HPC modelling & simulation
• HPDA data-driven workloads
• AI data-driven workloads & HPC with AI-in-the-loop

Figure 8 - MeluXina application domains

MeluXina will target emerging data science problems as well as traditional simulation and modelling problems. It can run both large-scale simulations and data-intensive workloads which can involve sensitive data. Distinct computing modules with different performance features are joined into a single system, able to execute heterogeneous applications concurrently on matching compute and data resources while sustaining reliably and securely hundreds of concurrent projects and their particular workflows.

Computational workloads coming from the following application domains are foreseen to run efficiently on MeluXina:

- Simulation and Modelling: in particular Materials Science through Computational Physics and Chemistry, Computer Aided Design and Engineering, Computational Fluid Dynamics;
- High Performance Data Analytics and AI: Machine Learning, Deep Learning and Big Data analytics, applied in particular to Life Sciences (Systems Biology, Bioinformatics, Genomics, Metabolomics, Proteomics), Computer Vision, Finance, Logistics, Smart Mobility and Space.

In particular, the Cluster Module will enable the execution of the widest range of applications and is optimised to address non-accelerated scalable science codes, coupling highly capable CPUs with sufficient main memory to fit expanding datasets and optimal bandwidth to allow the CPUs' full utilisation.

The Accelerator Module is optimised to address those applications that can exploit GPU-based parallelism (the main deliverable of this module), in particular data science AI applications that use Deep Neural Networks (DNNs). DNNs have been implemented in software frameworks successfully shown to exploit multiple GPUs per node for high speed-up and able to scale across many distributed nodes for both training and inferencing tasks. The high application performance will depend on the applications being able to fit into the fast memory of the GPUs, thus both a large High Bandwidth Memory (HBM) capacity per GPU and high bandwidth path to the node main memory are provided. The same Module hosts a set of reconfigurable computing nodes with FPGAs for highly specialised workloads.

The storage architecture of MeluXina is optimised to allow a concurrent workload mix with both HPC and HPDA/AI application usage patterns. High performance local storage is provided on several Modules at the node level for data staging, complemented by the larger scratch storage for application checkpoints and main storage for projects sufficient to host large data sets.

3.4.4. Benchmarks

The procurement process if the MeluXina system relied on the following list of benchmarks for the evaluation of submitted tenders:

Synthetic benchmarks:

- MPI benchmark (latency and bandwidth),
- Memory bandwidth benchmark,
- Graph500 benchmark,
- High Performance Conjugate Gradient (HPCG) benchmark,
- High Performance Linpack (HPL) benchmark,
- MLPerf benchmark suite (training and inference),
- IOR benchmark,
- IO500 benchmark.

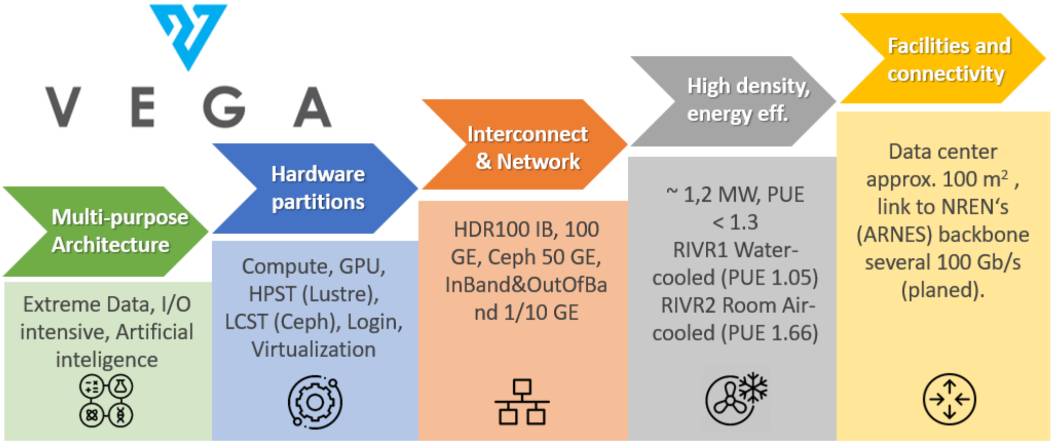
Application benchmarks:

- GROMACS,
- Quantum ESPRESSO,
- CP2K,
- FDS,
- NLP application (BERT/GLUE ran with containerized TensorFlow+Horovod),
- LAMMPS.

3.5. Vega

Commissioned as the primary supercomputer system of the Slovenian national research infrastructures upgrade project "HPC RIVR" and delivered as the first of EuroHPC Joint Undertaking systems, Vega is hosted at the Institute of Information Science - IZUM in Maribor. It increases the computing capacity in Slovenia and the European Union as a whole and helps researchers, as well as other users in the public and private sector.

HPC Vega is a driving force in innovation, helping Europe to compete globally in strategic areas, such as Artificial Intelligence, Advanced Data Analytics (HPDA), Personalized Medicine, Bioengineering, fighting climate change and the development of medicine and new materials. Main characteristics of HPC Vega are described below:



VEGA

Multi-purpose Architecture

Extreme Data, I/O intensive, Artificial inteligence

Hardware partitions

Compute, GPU, HPST (Lustre), LCST (Ceph), Login, Virtualization

Interconnect & Network

HDR100 IB, 100 GE, Ceph 50 GE, InBand&OutOfBand 1/10 GE

High density, energy eff.

~ 1,2 MW, PUE < 1.3 RIVR1 Water-cooled (PUE 1.05) RIVR2 Room Air-cooled (PUE 1.66)

Facilities and connectivity

Data center approx. 100 m^2 , link to NREN's (ARNES) backbone several 100 Gb/s (planed).

Figure 9 - The Vega system in a glance

With 960 CPU nodes (overall 1920 CPUs AMD Epyc 7H12 - 122000 cores) and 60 GPU nodes (overall 240 GPUs NVidia A100) sustained performance of HPC Vega is 6,9 petaflops (peak 10.1 petaflops).

Power supply and Cooling infrastructure for HPC Vega is divided into several areas. One 1.6 MVA 10k/400V transformer is installed in a separate facility room and the whole facility is backed up by 820 kVA Diesel Generator.

Atos supplied and installed 12 Liquid-cooled BullSequana Racks for Compute nodes in RIVR1 system room. Compute part in RIVR1 is backed-up with UPS autonomy of few minutes to enable clean shutdown procedure. Adiabatic cooling system is installed for warm-water regime with inlet 35°C and outlet 50°C enabling free cooling for 90 % of time, and is efficient up to the outdoor temperature of 33°C. This system is placed on the roof of IZUM building, and chiller for RIVR1 is placed in cellar. This system can provide cooling up to 880kW of heat loses in RIVR1.

RIVR2 room is equipped with classical in-row air conditioning. 7 standard racks are installed for 61 storage Ceph nodes, 10 DDN storage nodes, 30 virtualization/service nodes and 8 login nodes. One wider rack is for communication equipment. Routers are placed in a separate telecommunication system room. Power consumption of system area RIVR2 is limited to 110 kW. In the case of power outages, only RIVR2 room is supplied with power from UPS and diesel generator. In RIVR2, there is a redundant cooling system with standard cold-water regime.

3.5.1. System Overview

Network topology is shown below:

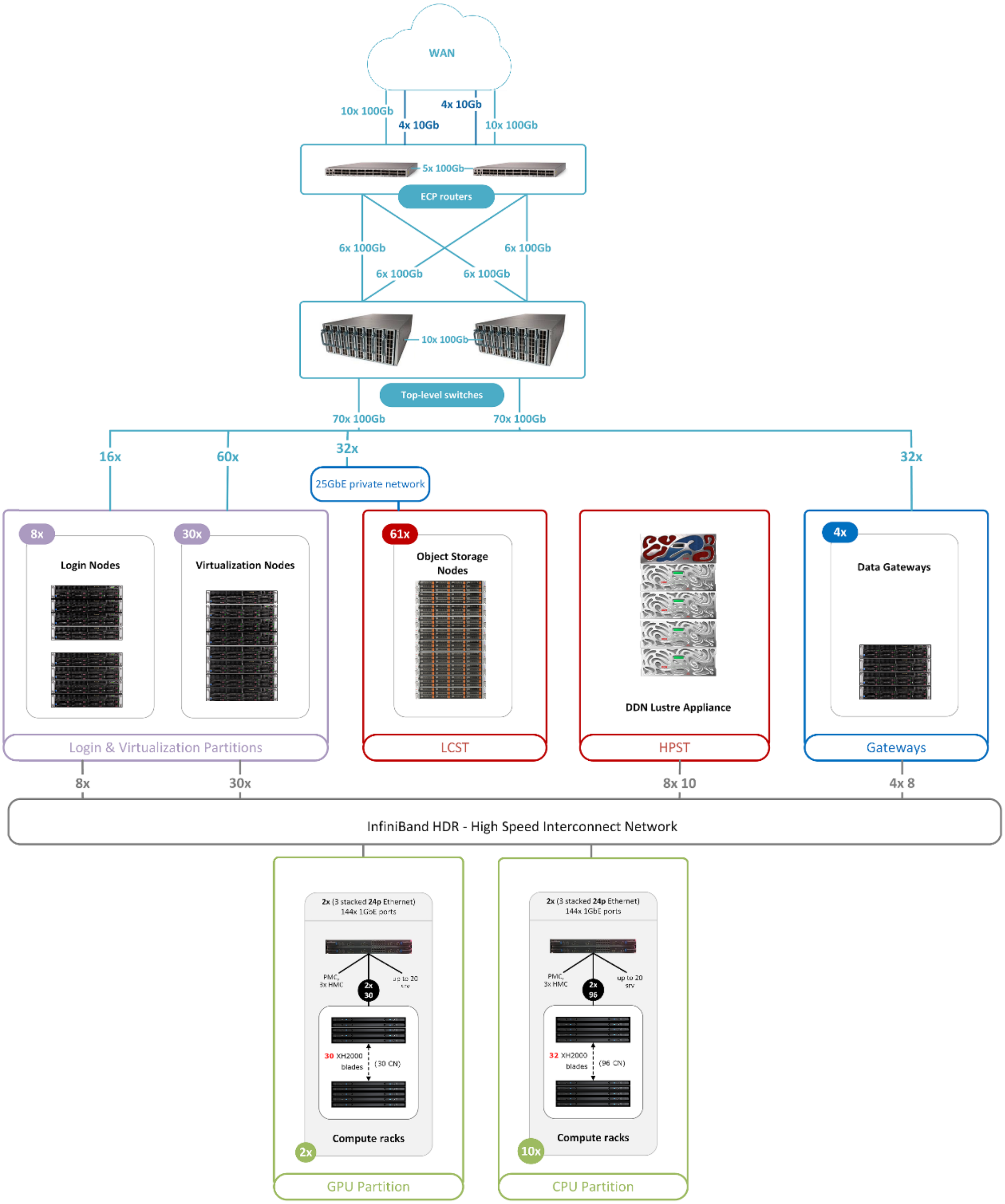


Figure 10 - Vega network topology

Connectivity and Network infrastructure

HPC Vega WAN-connectivity to ARNES backbone is redundant and provided by CPE routers. Main goals are to provide sufficient bandwidth capabilities to other data centres and supercomputers within SLING network, EU and worldwide for large data transfers and quality, reliable and secure connectivity for all user communities.

Top-level Ethernet network consists of two CPE routers Cisco Nexus N3K – C3636C-R, used for redundant connection of 5x 100GbE to WAN to backbone (to be provided by the end of 2021) and two top-level Ethernet switches Cisco Nexus N3K – C3408-S (192 ports 100GE activated) used to provide redundant connection for nodes in Login and Virtualisation and Service partitions and for storage subsystems High-Performance Storage Tier (HPST) and Large-Capacity Storage Tier (LCST).

- Top Management Network consist of two Mellanox 2410 switches (per switch 48x 10GbE ports). In/Out of Band Management Network consist of several Mellanox 4610 switches (per switch 48x1GbE + 2x 10GbE ports) and Rack Management Network consist of BullSequana integrated switches WELB (sWitch Ethernet Leaf Board) with three 24-port Ethernet switch instances and one Ethernet Management Controller (EMC).

Interconnect and Data gateways

Interconnect Network consist of 68x 40-port Mellanox HDR switch with Dragonfly+ topology:

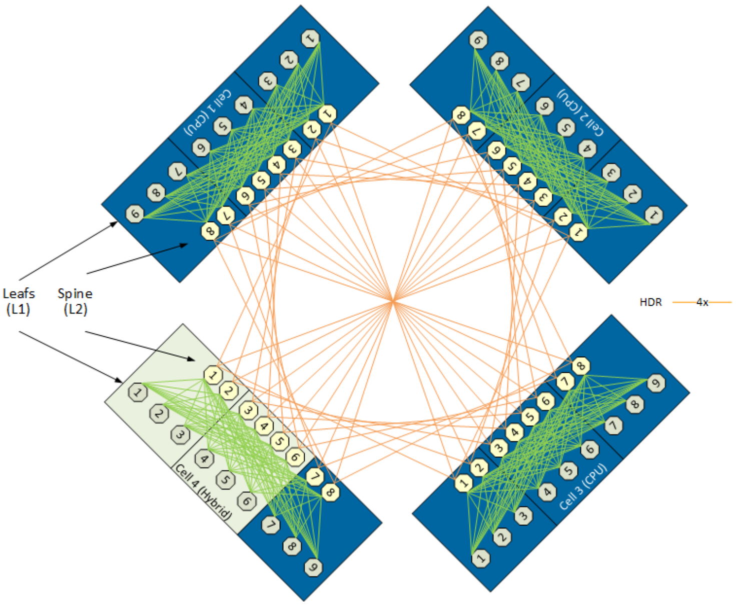
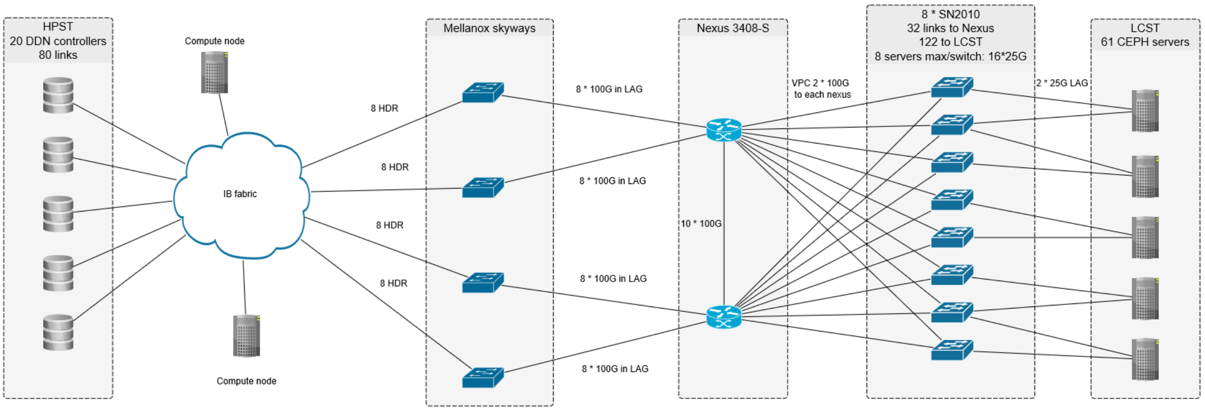


Figure 11 - Vega Dragonfly+ layout

All 960 Compute, 60 GPU, 8 Login, 30 Virtualization, 10 HPST storage nodes and 8 Skyway Gateways are connected with Mellanox ConnectX-6 (single or dual port).

Four NVIDIA (Mellanox) Skyway Gateway GA100 IP/Data Gateways with 8 connections on each side are used to route IP traffic between Infiniband on Compute cluster and LCST, based on Ceph. With this approach Compute nodes have access to the internet, so interaction between other data sources through secure networks such as LCHONE can be done.



Computing partitions

CPU partition consist of 10 BullSequana XH2000 DLC racks, with:

- 768 standard compute nodes (within 256 blades), each node with:
    - 2 CPUs AMD EPYC Rome 7H12 (64c, 2.6GHz, 280W), 256GB of RAM DDR4-3200, 1x HDR100 single port mezzanine, 1x local 1.92TB M.2 SSD
- 192 large memory compute nodes (within 64 blades), each node with:
    - 2 CPUs AMD EPYC Rome (64c, 2.6GHz, 280W), 1TB of RAM DDR4-3200, 1x HDR100 single port mezzanine 1x 1.92TB M.2 SSD

GPU partition consist of 2 BullSequana XH2000 DLC racks, with:

- 60 GPU nodes (60 blades), each node with:
    - 2 CPUs AMD EPYC Rome (64c, 2.6GHz, 280W), 512 GB of RAM DDR4-3200, local 1.92 TB M.2 SSD
    - 4x NVIDIA Ampere A100 PCIe GPU (3456 FP64 CUDA cores, 432 Tensor cores, Peak FP64 9.7 TFLOPS, FP64 Tensor Core 19.5 TFLOPS), each with 40 GB HBM2 and max. TDP 400W
- HPC Vega consist of 1020 compute nodes with at least 256 GB of RAM, all together 130560 CPU cores. Sustained performance on all CPUs is 3.8 PFLOPS. 240 GPU accelerators with all together 829440 FP64 CUDA cores and 103680 Tensor cores perform 3.1 PFLOPS.

High-performance Storage Tier (HPST) – Lustre

- HPST contains 10 DDN ES400NVX Building Blocks, each consists of:
- The ES400NVX base enclosure with 1+1 redundant storage controllers and fully redundant IO-paths to the storage devices, redundant Fans and Power Supplies
- 23x 6.4TB NVMe devices (DWPD=3), formatted in 2 x 10/0 flash pools
- 1 x NVMe device as HotSpare media, 8 x InfiniBand HDR100 frontend ports
- 4 embedded virtual machines (VM) to run the **Lustre** vOSS and vMDS with 1 OST and 1 MDT per VM.
- The building block provides 111 TB of formattable capacity. After applying filesystem overhead, the **usable Lustre capacity is around 1 PByte** for data and a maximum of 5 billion inodes in the system. Flash-based performance (MAX read and write) is more than 400 GB/s. HPST based on Lustre represents disk capacities for scratch space for I/O intensive applications and large scale jobs.

Large-Capacity Storage Tier (LCST) – Ceph

- LCST contains 61 Supermicro SuperStorage 6029P-E1CR24L servers, each consists of:
- 2x CPU Intel Xeon Silver 4214R (12c, 2.4GHz, 100W), RAM 256 GB DDR4 2933
- 2x 6.4 TB NVMe system disk and 24x 16 TB 3.5" SATA3 7.2K RPM data disk
- 2x 25GbE SFP28 to private internal network.
- 2x NVMe SSDs per server provide a 12.8TB raw NVMe capacity in total per OSD node. Each server contains 384 TB hard disk drive capacity and all together at least **19 PB**

of usable Ceph storage (with 16+3 erasure coded pools). Internal Ceph Network contains 8 Mellanox SN2010 switches, each 18x 25GbE + 4x 100GbE ports.

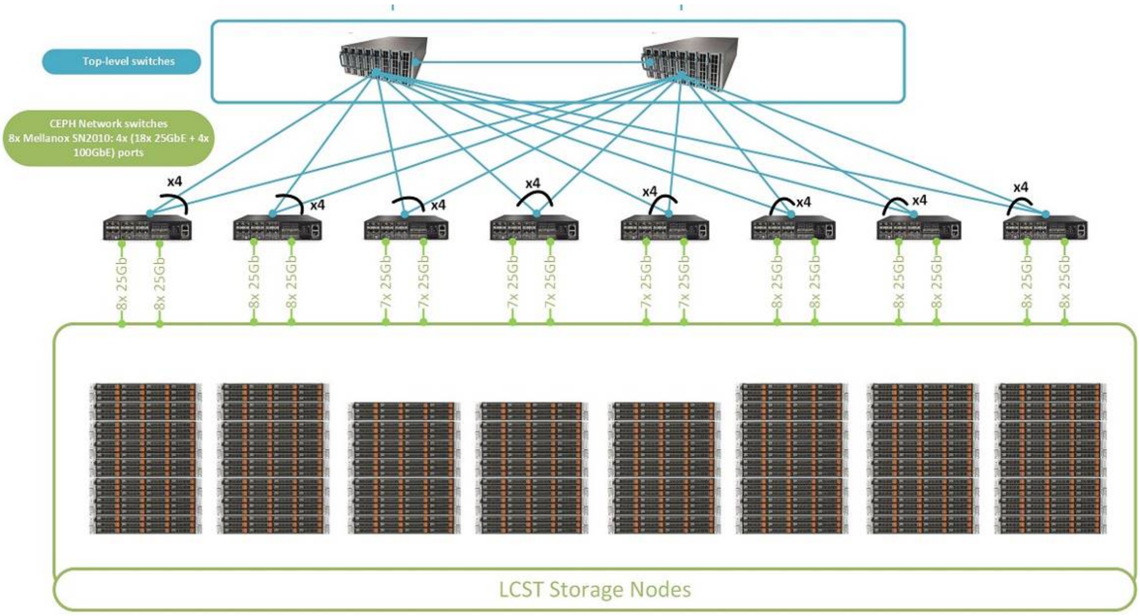


Figure 12 - Vega storage system layout

- Ceph storage performance is over 200GB/s. LCST based on Ceph represents disk capacities for:
- long-term services (system services and cloud virtualization)
- user home directories (100 GB defaults per user account)
- project group shared space and long-term scientific data repositories

Login nodes

- 4 Atos BullSequana X430-A5 **CPU Login nodes** consists of:
- 2x AMD EPYC 7H12 (same as on CPU partition),
- 256GB of RAM DDR4 3200, 2x local 7.6TB U.2 SSD
- 1x 100GbE DP ConnectX5, 1x 100Gb IB HDR ConnectX-6 SP.
- 4 Atos BullSequana X430-A5 **GPU Login nodes** consists of:
- 1x PCIe GPU NVIDIA Ampere A100 with 40GB (same model as on GPU partition)
- 2x AMD EPYC 7452 (32c, 2.35GHz, 155W)
- 256GB of RAM DDR4 3200, 2x local 7.6TB U.2 SSD
- 1x 100GbE DP ConnectX5 and 1x 100Gb IB HDR ConnectX-6 SP.

On Login nodes users can test their applications on the same CPUs and GPUs as they are used on CPU and GPU partitions.

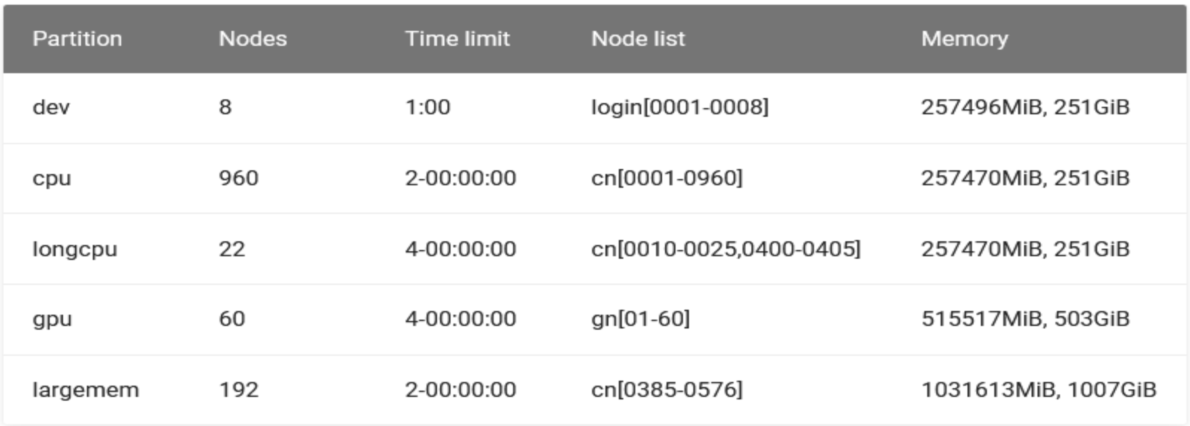
Virtualization and Service partition

- 30 Atos BullSequana X430-A5 **Virtualization nodes** consists of:
- 2x AMD EPYC 7502 (32c, 2.5GHZ, 180W)
- 512GB of RAM DDR4 3200

- 2x 7.6TB U.2 SSD
- 1x 100GbE DP ConnectX5 and 1x 100Gb IB HDR ConnectX-6 SP
- Virtualization and service partition is used for general purpose long term services, administrative tasks and tasks that do not correspond to the compute, GPU and login partitions. They are mainly used for system services and Cloud infrastructure.

3.5.2. Software environment

- **Atos System software:** Smart Management Centre, Bull Energy Optimizer, Performance Toolkit, IO Instrumentation
- **Monitoring, alerting:** DDN Insight Monitoring Software & ExaScaler Software Monitoring, Clary & Inventory software, Prometheus, Grafana, Icinga, NVIDIA NSight
- **Workload manager:** Bull Slurm
- **Orchestration:** BlueBanquise, Ansible
- **External Resource Access:** NorduGrid ARC, dCache, CVMS
- **Compute nodes Operating system**: Red Hat Enterprise Linux 8
- **Cloud/Service Virtualization:** Proxmox, RHEV, oVirt
- **Software stack**: EasyBuild toolkit, Modules (Lmod)
- **Compilers and SDKs**: Intel Compilers (licensed Parallel Studio XE), TotalView (licensed), GCC, AMD Optimizing C/C++ Compiler, NVIDIA SDK
- **MPI suites**: OpenMPI, IntelMPI
- **Programming languages**: Python, Fortran, C/C++, R, Julia, Go, Scala
- **Numerical, data and parallel/accelerator libraries**: BLAS, LAPACK/ScaLAPACK, MKL, BLIS, FFTW, HDF5, netCDF, Eigen, ARPACK, CUDA, cuDNN, cuFFT, cuRAND, cuSOLVER, cuSPARSE, cuTENSOR, TensorRT, KOKKOS, Jarvis, DeepStream SDK, DALI, AmgX, NCCL, Intel TBB, nvGRAPH, Thrust, nvJPEG, NVIDIA Performance Primitives, Video Codec SDK, Optical Flow SDK.
- **Frameworks**: PyTorch, TensorFlow
- **Containers**: Singularity Pro, Docker
- **Other:** Remote GUI X11-forwarding, User licensed Matlab
- ***Slurm partitions:*** The default partition is named *cpu*. Resource limits, node list and memory information are presented in the following table:



Partition	Nodes	Time limit	Node list	Memory
dev	8	1:00	login[0001-0008]	257496MiB, 251GiB
cpu	960	2-00:00:00	cn[0001-0960]	257470MiB, 251GiB
longcpu	22	4-00:00:00	cn[0010-0025,0400-0405]	257470MiB, 251GiB
gpu	60	4-00:00:00	gn[01-60]	515517MiB, 503GiB
largemem	192	2-00:00:00	cn[0385-0576]	1031613MiB, 1007GiB

3.5.3. Application domains

There are several domains and applications where Slovenian experts and expert researchers have developed advanced approaches or are part of the developement . In these domains, targeted high-level support or even co-development for substantial use of the infrastructure from user communities and projects in these spaces are foreseen. These domains and applications include:

- **Machine learning**, including container-based and optimised vector/GPU based application deployment for domain-specific systems as well as general frameworks (PyTorch, Theano, Caffe, Tensorflow) with optimizations for accelerator sharing, interconnect and RDMA architectural challenges and new deployments.
- **Biomolecular and materials simulations** based on force-field molecular dynamics, with support for NAMD, Gromacs, Amber, and LAMMPS. Modelling of enzymatic reaction with techniques based on the empirical valence bond approach.
- **Quantum chemistry and materials modelling** and simulations with support for Gaussian, NWchem, ORCA, and VASP and other popular packages.
- **Materials research** with Large-scale Monte-Carlo simulations of polymers, analysis of liquid crystal defects with Quantum ESPRESSO. Efforts in integration between machine learning, ab-initio models and Monte Carlo in progressive systems in high-energy physics, quantum chemistry, materials science, and complex matter research.
- **Medical physics,** fast, possibly on-line, GPU based analysis of medical tissues using RTE solution of layered media.
- **Non-equilibrium quantum and statistical physics**, dynamics and statistical properties of many-body system.
- **Fluid dynamics simulations and analysis in reactor physics and technology**, research and simulation in particle transport theory (neutrons, protons) with Singularity and Docker environments for Monte Carlo simulations for reactor physics of power reactors, research reactor physics, nuclear fusion, nuclear data evaluation and plasma physics.
- **Astrophysics,** data mining in large astrophysical collaboration datasets (Pierre Auger in Argentina and CTA), modelling of radiation from astrophysical structures. CORSIKA-based simulation and analysis of high energy cosmic ray particle-initiated air showers.
- **High-energy physics,** extreme scale data processing, detector simulation, data reconstruction and distributed analysis, ATLAS at CERN, Belle2 experiment at KEK, Japan.

- **Bioinformatics (Megamerge, Metamos),** co-development and data-flow optimizations for human genomics (GATK, Picard) with many applications in human, animal and plant-based research, also in medicine and medical diagnostics.
- **Fire-dynamics and simulations (FDS), fluid dynamics with OpenFOAM**.
- **Medical image processing** and research work with medical image analysis and modelling in medical physics.
- **Satellite images and astronomic image processing**, GIS processing using modern scalable machine-learning approaches, basic parallel and message-passing programming with new architectures and interconnects, support for basic mathematical and computational libraries, including Octave, Sage, Scilab.

Collaboration across disciplines, software stacks and system solutions has proven to be extremely effective and cross-development and interdisciplinary work, especially in combinations with ICT-related disciplines, machine learning, data processing, cross-pollination of detailed modelling, Monte-Carlo simulation and signal analysis, optimised modelling and machine learning model, with particular attention to introspective machine learning, in the stack of approaches to high energy physics, computation chemistry from quantum levels to protein folding and complex matter analysis and modelling.

3.5.4. Benchmarks

The procurement process if the Vega system relied on the following list of benchmarks for the evaluation of submitted tenders:

Synthetic benchmarks:

- Linktest
- IMB
- HPL
- HPCG
- STREAM
- IOR
- Mdtest
- Iperf3
- Firestarter

Application benchmarks:

- GROMACS
- Quantum ESPRESSO
- HEPSpec06
- OpenFOAM
- TensorFlow Resnet-50

4. Appendix I: Abbreviations and Acronyms

HPC	High Performance Computing
IO	Input/Output
IME	Infinite Memory Engine
ETS	Energy to solution
TTS	Time to solution
GUI	Graphic User Interface
NVMe	Non-Volatile memory express
SSD	Solid state drive
IOPS	I/O operation per second
GB	Giga bytes
PB	Peta bytes
Flops	Floating point operation per second
r/w	read / write
GPU	Graphic Processor Unit
CPU	Central Processing Unit
PUE	Power Usage Effectiveness
TTS	Time to solution
ETS	Energy to solution
Flops	Floating point operations per second
PFlops	Peta-Flops (10^{15} Flops)