Atos Proposition to the Contracting Authorities for the Purchase of Supercomputer Vega (JN-01P/2019-OP)

Authors: Vasile Chercea, Mathieu Dubois, Albert Gallina, Stefan Kero, Benjamin Pajot, Jure Sirena

Version: 9
Document date: 10th July 2020
Doc. No: 20 00064 - 9000719787

Contents

A. Management Summary

This document is part of the Atos' proposition to IZUM and EuroHPC JU for the project VEGA.

It gives additional details to the answers filled by Atos in the "Appendix_B_Evaluation_of_Requirements_23-04-2020(final)" document, and addresses requirements made in the Annex_Technical_Specifications_23-04-2020(final).

To ease the reading, the numbering of his document corresponds to the one of the "Annex_Technical_Specifications_23-04-2020(final). It therefore starts with section 5 Technical Requirements and finishes with section 12 Value for Money

It is with great pleasure that Atos submits its proposal to the procurement and evaluation team for the deployment of the VEGA Petascale Supercomputer to support the European scientific community and EuroHPC-JU's journey towards a European eXascale supercomputer infrastructure.

Atos is proud to propose its solution, services and collaboration approach and European added value for the Vega Petascale Supercomputer:

- **A 6.8 PFlops, flexible, hybrid and multi-purpose system capable of meeting the scientific workloads from 2020 and forward;**
- **A robust production system equipped with the most suitable technologies available;**
- **Delivering maximum performance within the given budget;**
- **Meeting all the mandatory requirements;**
- **Integrated into an adapted data center;**
- **Minimizing risks by Atos design choices, proven delivery methodology and solid support organization and processes.**

Atos, a European established company, through the Bull heritage and the Big Data & Security division, has a wealth of experience in supporting international high-level research centres of expertise in Europe and worldwide, and a strong track record in supercomputing innovation.

With a pedigree of over 50 years developing and implementing complex technical computing platforms, IZUM and EuroHPC-JU can be confident of the quality of the service and support offered by the Bull/Atos team.

B. Atos, the true European global player

Developing the European Exascale Ecosystem

Atos is playing a leading role in the formation of world-class exascale and post-exascale HPC technologies through the development of low-power micro-processors and related middleware, and their integration into supercomputing systems. This is primarily through a co-design approach with major European research labs, including many of the EuroHPC hosting entities. Exascale systems will need to simultaneously meet challenges related to performance, system cost, and energy efficiency. To find the right balance, global system level optimization will be necessary to ensure the resultant systems meet the performance requirements within a reasonable budget and are not too power hungry. Through its leadership position in the European Processor Initiative (EPI) Atos is bringing its expertise in supercomputing architectural design to help implement a new family of low-power European processors for extreme scale computing, high-performance big-data analytics and a range of emerging applications, ensuring system designs remain well balanced. While the hardware developments for exascale are now in sight, to achieve the real benefits that exascale can deliver in terms of international competitiveness requires additional capabilities in software and skills enablement. Currently there are very few application domains that can effectively exploit exascale systems and even fewer application developers capable of effectively scaling applications to run on them.

The current peta-scale and pre-exascale systems will provide an important steppingstone in the development of these capabilities, as will close integration with the major European research activities related to application development.
With this long term capability enablement in mind, Atos is actively involved in many large scale application development programmed with major research labs and industry, including CEA, Juelich Research Center, CINECA, BSC and regional universities, helping Europe to remain at the leading edge in a highly competitive international context, producing world-class science and ensuring the fast and broad exploitation of European research and technology across the Union, particularly within industry. Some relevant examples include:

- **CompBioMed.** CompBioMed is a H2020 funded Centre of Excellence focused on the use and development of computational methods for biomedical applications. The project, which is currently leading a coordinated research activity in Covid-19 research, includes major pharmaceutical companies, European research labs and the University of Bologna. Atos is the main technology partner and is active in several work packages related to applications scalability and porting to new technologies.
- **ESiWACE.** ESiWACE is the Centre of Excellence in Simulation of Weather and Climate in Europe (ESiWACE) and enables global storm- and eddy resolving weather and climate simulations. Current models of weather and climate provide valuable forecasts, but still cannot fully resolve storm systems in the atmosphere, and eddies, their oceanic counterpart. The project will use the upcoming pre-exascale

supercomputers to increase the resolution of the models and resolve new physics. Atos is actively involved in scaling weather dwarfs and integrating AI into weather and climate application workflows.

- **CHEESE.** CHEESE is the Centre of Excellence in Solid Earth Modelling and is targeting the preparation of 10 Community flagship European codes for the EuroHPC pre-Exascale and Exascale supercomputers.
- **Mont Blanc.** The successive Mont-Blanc projects have in common their investigations into new processors for HPC, based on European technologies. The first three Mont-Blanc projects that ran from 2011 to 2018 assessed the potential of Arm-based clusters to address Exascale HPC needs and developed the corresponding software ecosystem. The currently running Mont-Blanc 2020 project will pave the way to the future low-power European processor for Exascale (European Processor Initiative).
- **European Processor Initiative.** The European Processor Initiative (EPI) is a major project within the European Commission, where a consortium led by Atos aims is to design and implement a roadmap for a new family of low-power European processors for extreme scale computing, high-performance Big-Data and a range of emerging applications.
- **LEXIS.** The LEXIS (Large-scale EXecution for Industry & Society) project will build an advanced engineering platform at the confluence of HPC, Cloud and Big Data which will leverage large-scale geographically distributed resources from existing HPC infrastructure, employ Big Data analytics solutions and augment them with Cloud services. Driven by the requirements of the pilots, the LEXIS platform will build on best of breed data management solutions (EUDAT) and advanced, distributed orchestration solutions (TOSCA), augmenting them with new, efficient hardware capabilities in the form of Data Nodes and federation, usage monitoring and accounting/billing supports to realize an innovative solution. CIMA, the Centro Internazionale in Monitoraggio Ambientale (International Centre on Environmental Modeling) are also members of this project.
- **NEASQC (NExt ApplicationS of Quantum Computing).** The NEASQC project, coordinated by Atos, is devoted to the emergence of practical applications of quantum computing in its NISQ era, and to the construction of a strong community. NEASQC is use-case centric and organized around 9 real NISQ-compatible use cases defined by the industrial members of the project. As part of NEASQC, a complete quantum programming environment (QPE), composed of a special build of Atos' myQLM will be made available for free to the QT Flagship community.

More than any other company, through these projects, Atos is genuinely investing to create a European HPC ecosystem.

Enabling leadership-class supercomputers

High Performance Computing is a critical element for the digitization of industry and the data economy of the 21st century. Through this proposal Atos will provide IZUM with a leadership-class supercomputer system based on technologies designed and manufactured in Europe. The system will use hardware and software components which have been co-designed with leading European research labs with the specific intention of enabling a European HPC ecosystem.

The design and operation of the systems proposed takes into consideration environmental sustainability. The BullSequana XH2000 systems represent the state of the art in energy efficient supercomputing using direct liquid cooling for all major components to extract all the heat generated by the system to water, which can then be reused to heat buildings. Furthermore, our development of smart power management solutions, building on research undertaken in H2020 projects, allows intelligent energy metering and applications to be optimized for energy to solution.

Through our development of HPC portals and new user interface tools the Atos supercomputer can be used by a wide range of users from the research and scientific community, as well as the industry including SMEs and the public sector, for new and emerging data and compute-intensive applications and services.

Atos is playing a leading role in the formation of world-class HPC technologies through the development of low-power micro-processors and related middleware, and their integration into supercomputing systems. This is primarily through a co-design approach with major European research labs, including many of the EuroHPC hosting entities. Future exascale systems will need to simultaneously meet challenges related to performance, system cost, and energy efficiency. To find the right balance global system level optimization will be necessary to ensure the resultant systems meet the performance requirements within a reasonable budget and are not too power hungry. Through its leadership position in the European Processor Initiative (EPI) Atos is bringing it expertise in supercomputing architectural design to help implement a new family of low-power European processors for extreme scale computing, high-performance big-data analytics and a range of emerging applications, ensuring system designs remain well balanced.

The system Atos proposes for the Vega project will use hardware and software components which have been co-designed with leading European research labs with the specific intention of enabling a European HPC ecosystem

Reinforcing the digital technology supply chain in the Union

As Europe's leading manufacturer of supercomputing systems, Atos not only employs critical design and development skilled resources within Europe but also supports a wide supply chain within the Union.

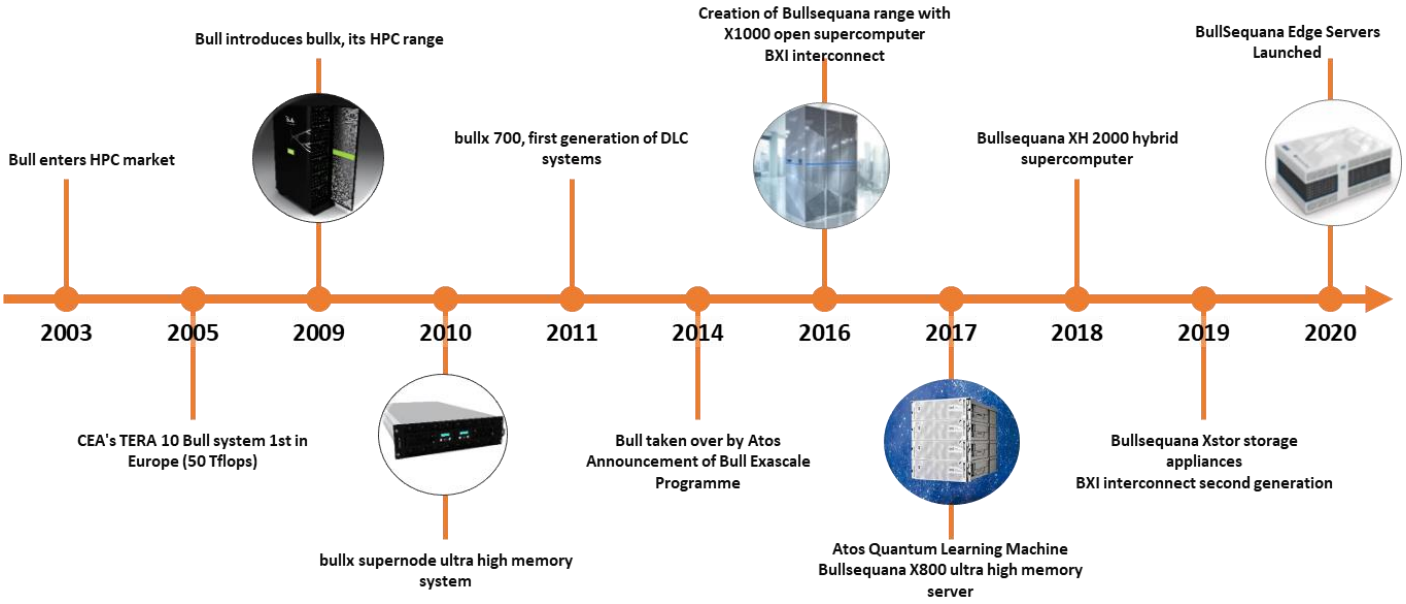
While we exist within a global economy and component parts are sourced from across the world, the key high value skills of design, integration and scalable software development are maintained within Europe. This allows Atos to offer its clients a full end-to-end solution capability, bridging the gap between research and development and the delivery of future exascale systems and software, reinforcing the digital technology supply chain within the Union while also integrating European technologies.

The Atos Scientific Community consists of 160 leading European scientists who drive the technological advancement of the company, setting its technical direction. It is supported by an Expert Community of over 2,000 people with domain specific expertise. Many of the people within these R&D communities have either previously worked, or are currently working, on collaborative R&D projects and co-design activities, building a competitive data and knowledge economy in Europe.

Ultimately, the manufacturing, co-design and R&D activities, undertaken in conjunction with Atos supply chain partners and academic collaborators, will lead to the development of a full European HPC ecosystem capable of developing new European technology with which furtur exascale supercomputers will be realized.

Integrating European technologies

The supercomputer solution proposed by Atos has European design credentials at its core and represents the company's largest ever R&D investment.



Specific examples of European technologies integrated into our systems are described below:

- **BullSequana XH2000 Supercomputer.** BullSequana XH2000 is the fifth generation of Atos supercomputer. Its heritage can be traced back to the early 2000s when Bull embarked on a co-design collaboration with CEA to develop a European sovereign capability in HPC and remove the dependency on foreign technologies. Progressively Bull, and subsequently Atos, has continued this approach, increasing the proportion of natively developed technology within each generation of system. The timeline above shows some significant milestones in that history
- **Direct Liquid Cooling Systems.** BullSequana XH2000 has a fan-less design and a unique Direct Liquid Cooling (DLC) technology that uses warm water up to 40°C to cool all critical components within the cabinet (compute blades, interconnect and management switches, and PSUs). The cooling system, designed by Atos R&D teams in France and representing the state of the art in energy efficient supercomputing, is composed of hydraulic chassis (HYC), primary and secondary manifolds, and an expansion tank. All critical components are direct liquid cooled and mount directly onto the secondary manifold via hydraulic non-spill quick disconnect couplings. The expansion tank prevents excessive pressure within the hydraulic circuit.
- **Blade Design.** Most blades within BullSequana systems are developed by Atos R&D teams in collaboration with key customers. A new generation of GPU blade

was developed in a co-design project with Julich and will be used to deliver Europe's most powerful supercomputer in Autumn 2020.

- **Memory Controllers and SMP servers.** Atos' shared memory supercomputer, BullSequana X800, uses a unique memory controller, now in its third generation, which allows Atos servers to scale to 32 sockets for large scale data intensive computing and AI applications. These servers, which are also sold through reseller agreements by Dell, Hitachi and Cisco, are also used in enterprise computing for large data base applications such as Sap/Hana. Furthermore, they are offered by both Google Cloud and Microsoft Azure as part of their global enterprise cloud offerings.
- **Edge Servers.** The requirement to undertake analytics at the edge is becoming increasingly important to avoid the costly movement of data. Analysts are predicting that by 2022 75% of data generated by IoT devices and distributed sensor networks will be processed at the edge. To address the challenge that this explosion of data presents, Atos Sequana Edge servers were introduced in 2019 to deliver powerful AI inference and streaming analytics capabilities at the edge while ensuring that all data remains safe and secure.
- **Interconnect.** A core component of any HPC system is the high-performance interconnect. Atos has many years' experience developing interconnect technologies and first introduced the Bull Exascale Interconnect (BXI) in the Tera1000 system at CEA in 2016. BXI networks scale up to 64k nodes and feature high-speed links (100 Gb/s), high message rates, and low latency across the network. Most importantly, to boost the communication performance of HPC applications, BXI provides full hardware offload of communications, enabling CPUs to be fully dedicated to computational tasks while communications are independently managed by BXI. The second generation of BXI will shortly be announced which introduces additional features and functionality. We have not proposed BXI for VEGA pending as its general availability is not compatible with the project timeframe.
- **Storage.** Atos will shortly introduce a range of natively developed storage products under the brand name BullSequana XStor. XStor is a new range of modular storage appliances designed for HPC environments, focussed on Performance, Modularity, Scalability and Cost Optimization. While currently most of our systems are sold with third party storage products, by introducing XStor we will be progressively increasing the amount of European developed technology sold as part of our solutions.
- **Software Products.** Over the last decade Atos has invested heavily in the development of software environments for HPC systems, ranging from systems management solutions to application development environments. The foundations of many of these tools were European FP7 and H2020 projects and they are now in widespread use across European academia and industry:
  - **Smart Management Centre (SMC)**: The Atos stable and robust software environment for HPC systems including: Optimised Operating

System, compilers and development environment, scheduling and monitoring tools.

- **Smart Data Management Suite.** This set of tools is used to monitor and store all IO related events giving a detailed visual analysis of IO patterns and bottlenecks. The tools provide customers with the ability to understand the low-level IO behaviour of the core applications, helping to inform their IO strategy.
- **Smart Power Management Suite.** These tools provide accurate information on energy and power usage of applications, to implement power capping policies and to optimize power consumption based on the individual phases of each application.
- **Cognitive Data Centre.** This uses Atos AI tools and instrumentation to monitor data centre usage, predicting potential failures before they occur, improving overall efficiency and service level compliance.
- **Bull Open MPI.** Based on an open source MPI stack Open MPI 2.x, Bull Open MPI is optimised for run time scalability and designed for exascale applications.
- **Atos Codex AI Suite for hybrid HPC/AI applications.** Designed for coupled HPC/AI workloads and including the main AI frameworks such as TensorFlow, Caffe etc., allowing the use of containers in a virtualised environment.
- **Atos Quantum Learning Machine.** Based on our shared memory technology and capable of emulating up to 41 qubits, the Atos Quantum Learning Machine is the most comprehensive quantum computing emulator appliance available today, allowing customers to prepare for the introduction of physical quantum computers in the future.
- **Atos Performance Toolkit**. This includes a set of tools to analyse application behaviour (Lightweight Profiler and HPC Toolkit) or system environment (PAPI and Binding Checker) in a synthetic way with very low overhead on performance.
- **HPCaaS Interface tools**. Atos Extreme factory computing studio (XCS) is a web portal designed for HPC cloud frontends. With XCS, organizations can easily and centrally manage their compute and visualization workloads. Atos's HPC-as-a-Service offer leverages XCS to power the Bull extreme factory on demand platform, enabling experts and non-experts alike to access the systems. Several European ISVs also use the XCS technology for their HPC Clouds, and many Atos HPC customers rely on XCS to access their private HPC resources on premises or hosted.

All the above hardware and software developments represent the results of decades of investment by Atos in a demand-oriented, user-driven and open collaborative approach to the design, development and integration of world-class supercomputing and data infrastructure, with which European scientists in industry and academia can maintain their international competitiveness and solve societal, environmental, economic and security challenges.

Furthermore, these technologies are now fundamental to Europe's defense and homeland security challenges where in several member states they are used to the undertake the modelling and simulation which underpins compliance with the international nuclear test ban treaty and the data analytics required to ensure the safety of European citizens from terrorist activity.

Atos is leading the vanguard for European supercomputing technology deployments in other parts of the world and now exports its supercomputing technologies to every inhabited continent. Our systems are internationally competitive and are widely respected, reflecting renewed confidence in European developments. We are the leading supercomputing supplier in South America and India and have an increasing footprint in Africa and Asia Pacific.

In the US, our quantum computing technology is deployed at the major DoE labs and our HPC systems are in production use at US based international conglomerates such as Procter and Gamble. Through recent acquisitions of US companies such a Xerox IT Services, Syntel, and Maven Wave and with our global strategic alliance with Google Cloud, Atos is extending its portfolio to further accelerate the transformation and growth of the North American market, reinforcing the company's commitment to extend and enrich its digital portfolio, including HPC, AI and Big Data systems, across the globe. This strategy is essential for the uptake of large-scale and emerging application fields such as personalized medicine, connected and automated driving or other lead markets that are underpinned by artificial intelligence, blockchain technologies, edge computing or more broadly by the digitalization of industry.

Contributing to the development of the European exascale ecosystem

Atos formally launched its programme to develop and market exascale systems in 2016 with the announcement at SC16 of the first generation of the BullSequana supercomputers and the Bull Exascale Interconnect programme. These announcement symbolized the significant ambition and long term vision of the company, which had only recently integrated Bull, to invest in the research and development needed to create a native European exascale capability and associated ecosystem using, where possible, technologies that have been designed and developed within the Union and building on Europe's collaborative R&D programmes.

Since those first announcements Atos has significantly increased its global HPC market footprint and has evolved the systems to the point where the BullSequana architecture is now capable of integrating the processor and interconnect components with which to achieve exascale.

While the hardware developments for exascale are now in sight, to achieve the real benefits that exascale can deliver in terms of international competitiveness requires additional capabilities in software and skills enablement. Currently there are very few application domains that can effectively exploit exascale systems and even fewer application developers capable of effectively scaling applications to run on them. The current peta-scale and pre-exascale systems will provide an important steppingstone in the development of these capabilities, as will close integration with the major European research activities related to application development.

Atos is actively involved in many large scale application development programmes with major research labs and industry, helping Europe to remain at the leading edge in a highly competitive international context, producing world-class science and ensuring the fast and broad exploitation of European research and technology across the Union, particularly within industry.

Contributing to Major European Projects

Atos is involved in many European projects regarding High Performance Computing. Much of the hardware and software which this response refers has been developed over the years through very close technical relationship and projects. The solution that will be delivered is based on Atos engineering and research done in partnership with major HPC centres, commercial partners, as also research centres.

The following projects are examples of Atos involvement which have bring new products on the market:

1) **Hybrid4HPC (ITEA2)** European project for the design of Exascale ready interconnect solution including software runtime environment. In this context the BXI solution has been designed and will be put on the market soon. Also, many improvements in the MPI stack and Slurm resource manager.
2) The **SAGE** (H2020) and **SAGE2** (H2020) European projects for the data management improvement on which we have based some development of our IO Instrumentation framework and tools as also our Smart Burst Buffer solution to speed-up data exchanges.
3) The **MontBlanc** (FP7), **MontBlanc2** (FP7) **MontBlanc3** (H2020) and **MontBlanc2020** (H2020) European projects in which we have developed our ARM-processor support as also energy efficiency of the overall solution. Our **Bull Energy Optimizer** (BEO) and **Bull Dynamic Power Optimizer** (BDPO) have been developed partially based on the research performed in those projects. Also, the

ARM support for most of Atos products (hardware and software) have been done inline to those projects.

4) The **TANGO** (H2020) project in which the resource and job management has been improved to manage heterogeneous resources or MPMD. This work has been performed in close collaboration with Slurm community and major research centers to extend Slurm features which are now integrated in the official Slurm version.

5) The **COLOC** (ITEA2) project in which the MPI stack and runtime environment has been enhanced to better support hybrid executions and optimize applications through tools such as the **Lightweight Profiler** (LWP) we have developed.

6) The **LEXIS** (H2020) project is for hybrid computing support to extend the HPC scope to new use cases (not only HPC-centric application) such as AI and Big Data. The scheduling and application workflow management are big parts of this project. The **SMC** project features (based on K8S) are fully in line with the targets of this project.

7) The **ELCI** project (French) to enhance the end-user environment in which the next-generation features for Slurm, MPI and OpenMP have been challenged. Some of the concept raised in ELCI has been integrated in Slurm and OpenMPI.

8) The **ParMA** (ITEA2) for improvement on multi-cores application, especially for OpenMPI intra-nodes enhancements.

9) The **European Processor Initiative** (EPI-SGA1) which aims to design and develop a European-made processor for multi-purpose usage. In this context the software environment developed by Atos follow the EPI project to ensure that it will support this news hardware environment.

10) The **Exascale project** (French) which is a multi-year technical cooperation with CEA to design and develop Exascale-class supercomputers. In this project we co-develop new features regarding MPI, management stack, interconnect, new hardware support, energy optimization, security. In this very large project, we have enhanced the reliability of our hardware and software (including high level security), the application performance 'out-of-the-box' as also many features in Slurm.

11) The **Mousquetaire-Group** which is an industry/research cooperation group to develop new solution for Exascale-class systems. In this context multiple thematic are addressed from modular design of the solution to IO patterns management.

12) Some others key H2020 projects regarding application performance such as **ASPIDE, VECMA, CHEESE, CompBioMed, CompBioMed2, ESCAPE, ESCAPE2, EsiWACE, EsiWACE2** in which our Application & Performance group is deeply involved close to customers to enhance real-case applications.

13) The HPC Quantum project to develop Quantum-computing solution and especially on the software side (development environment, etc.). In this project we are deeply involved **AQTION** (H2020) **and PASQUANS** (H2020).

14) **NEASQC (NExt ApplicationS of Quantum Computing).** The NEASQC project, coordinated by Atos, is devoted to the emergence of practical applications of quantum computing in its NISQ era, and to the construction of a strong community.

15) The **DEEP** series of projects (**DEEP, DEEP-ER, DEEP-EST, DEEP-SEA, IO-SEA, RED-SEA**) are developing the Modular Supercomputing Architecture for exascale systems in Europe from a software perspective, aiming to improve the power efficiency of HPC systems by an order of magnitude.

Embedding Energy Efficiency and Sustainability Considerations

Atos has long-term experience and a culture built around safeguarding the environment in the delivery of IT solutions and services. This is a core value of our business and has been embedded into our DNA since the launch of our environmental programme in 2008. With the recent appointment of our new CEO Elie Girard, we have launched the Spring programme, which focusses the resources of the company on achieving "*secure, decarbonized digital services".* We believe that because of this, we can deliver the most sustainable service possible for the EuroHPC supercomputing systems.

Since our programme started, from an environmental perspective, we have achieved a 60% reduction in carbon in intensity, set science-based emissions reduction targets, which are aligned with climate science and endorsed by the Science Based Target initiative (SBTi). In the process of doing this, we have consistently achieved the highest possible ratings from the main voluntary disclosure mechanisms, specifically, the GRI, CDP, EcoVadis, Carbon4, MCSI and the Dow Jones Sustainability Index, with the latter citing Atos as the world leader within the IT sector. Atos is also ISO14001:2015 certified, demonstrating that our operational processes are embedded across the business and adhere to maintaining strong environmental standards.

Due to our long-term global programme to drive down our environmental impacts, Atos has been carbon neutral since 2018.

With the ambition to achieve net zero by 2035, Atos has committed to follow a SBTi formally endorsed science-based emissions reduction target, driving a 1.5°C emissions reduction pathway while moving towards CO_2 sequestration of all residual emissions, following the latest recommendations from the IPCC.

Decarbonizing the EuroHPC Supercomputers

For VEGA, we intend to provide a supercomputer that has net zero emissions in its design creation and delivery and minimizes emissions during the run phase. Specifically, we will:

- Ensure that all emissions related to the bid process including travel, benchmarking and design activities are monitored and offset through sequestration.

- Ensure that emissions related to the manufacture of the supercomputers follows a 1.5°C pathway, over the course of the agreement, through our emissions reduction approach, described shortly.
- Ensure the supercomputers are designed to run using the most energy efficient direct liquid cooling technologies available today.
- Offset all residual emissions through CO_2 sequestration. These will cover all emissions sources, including embodied emissions and ongoing travel related to the service

Atos has a strict and standard approach following best practice to minimize our impacts, which reflects the approach that we will take towards the delivery of the VEGA systems. That is:

- Minimize energy consumption
- Source emissions-free, ideally renewable energy in place of fossil fueled energy in our own data centres, offices and manufacturing facilities
- Offset all residual emissions

Minimizing energy consumption

As is the case for all products and services, varying amounts of energy will be consumed across their full lifespan. In many cases, energy consumption leads to greenhouse gas emissions, which in turn drive climate change. Therefore, it is important to gain an understanding of where this energy is consumed, to perform a quantitative assessment and drive an emissions reductions programme to minimize the impacts from this energy. In the case of the Leonardo systems there are two value chains that will result in GHG emissions:

- For the supercomputers and related IT hardware, there is upstream energy consumption from the sourcing of raw materials, manufacturing of components, shipping, assembly and further shipping to the operational location. In the use phase, energy in the form of electricity is used to power the supercomputers and then finally energy is used for disposal in shipping, dismantling and recycling.
- There is also energy consumption within the manufacture, operation and disposal of purpose-built facilities such as data centres and associated cooling and power resilience plant.

Addressing IT Hardware Energy Consumption

The Atos BullSequana supercomputers are designed and manufactured in a state-of-the-art facility in Angers France. The facility, which opened in 2019, is equipped with an energy-efficient cooling system which uses low-GWP (Global Warming Potential)

refrigerant fluid and 'free-cooling' to chill the water which can result in energy savings of up to 75%. It has an energy recovery system which reuses the energy generated by the operation of the manufacturing plant to heat or cool the offices, operating at an energy-efficient Coefficient of Performance (COP) of 6. It is also equipped with a 'green' roof, electric vehicle charging terminals, and photovoltaic roof panels.

Once manufactured, the systems themselves will be shipped to the Maribor hosting location by road, avoiding high Carbon air freight travel.

Within the use phase, there will be significant energy consumption in the form of electricity to power the super computers. However, we can limit this consumption through the deployment of our highly efficient and innovative direct liquid cooling technology.

BullSequana XH2000 supercomputers utilize the fourth generation of Atos' patented warm water-cooling technology to directly cool all major components, including the power supplies, of the system. By using inlet water up to 40°C, Atos' Direct Liquid Cooling can extract up to 100% of heat generated to water, meaning that no mechanical cooling is required for the systems. This lowers the system PUE, providing exceptional energy efficiency and minimizing the Carbon footprint of the systems.

Offsetting of all residual emissions and net-zero.

Atos undertakes to minimize atmospheric greenhouse gas emissions wherever possible, however, with the world economy and current technologies inevitably there will be residual emissions that cannot be eradicated. Therefore, in line with best practice and as the final step of our emissions reduction programme, Atos offsets all residual Scope 1, 2 and operational Scope 3 emissions. Atos commenced its offsetting programme in 2012 by offsetting all strategic global data centres. In 2018, this programme was extended to cover all of Atos. The chosen method for offsetting at that time was emissions avoidance, through funding wind farm projects in India so that emissions are not generated by local fossil fuel power plants. Through carbon trading mechanisms, Atos has ensured that the CO_2 emissions avoided in India matches the CO_2 footprint of the company, thereby achieving carbon neutrality.

In recent times, climate science has favored a net-zero approach, although there is no globally recognized standard definition of exactly what this means. Atos recognizes and understands the value in following this direction and has chosen to adopt the definition of net-zero as defined by the Science Based Targets institute and endorsed by the Carbon Trust. The Carbon Trust website describes net zero as "achieving a state in which the activities within the value chain of a company result in no net impact on the climate from greenhouse gas emissions. This is achieved by reducing value chain greenhouse gas emissions, in line with 1.5°C pathways, and by balancing the impact of any remaining greenhouse gas emissions with an appropriate amount of carbon removals". In order to adhere to this definition, Atos is taking 3 significant steps within its environmental programme:

- Switching to a 1.5°C Science Based Target trajectory from the current 2oC target. Note that Atos was an early adopter of Science Based Targets and received endorsement for the current target in 2016, when Climate Science advised that a 2oC target was required.
- Modifying the offsetting programme, by moving from 100% emissions avoidance to 100% sequestration. Already in 2019 Atos has switched to an even split between avoidance and sequestration through:
- 50% avoidance, by continuing to support Indian Wind Farms
- 50% sequestration, through conservation of Peruvian rainforest.
- Developing methodologies to enable more accurate Scope 3 emissions (value chain) assessments so that these may be incorporated into the emissions reduction targets and ultimately offset in a meaningful way.

In Atos we believe that we can achieve the most practical sustainable solution for the Vega Supercomputer through our approach, experience and understanding of sustainable IT services.

We have experience with existing customers of successful transition to low carbon hosting.

We also take pride in our ability to work closely with customers on broader sustainability initiatives, to benefit the services and the wider community and environment.

5 Technical Requirements

5.1 Supply Infrastructure Requirements

5.1.1 Supply Infrastructure General Requirements

5.1.1.1 SUP1 [INHW] Power Consumption Limitations

Total VEGA HPC power consumption is designed to be below 1200 kW (at 1161 kW), as shown in the table.

ID	Type of HW	Description	Environment	Full power load [kW]	Operational power load [kW]
RIVR 1.11	Bull Sequana XH2000	CPU rack	Full load	52	47
RIVR 1.12	Bull Sequana XH2000	CPU rack	Full load	52	47
RIVR 1.13	Bull Sequana XH2000	CPU rack	Full load	52	47
RIVR 1.14	Bull Sequana XH2000	CPU rack	Full load	52	47
RIVR 1.15	Bull Sequana XH2000	CPU rack	Full load	52	47
RIVR 1.16	Bull Sequana XH2000	CPU rack	Full load	52	47
RIVR 1.21	Bull Sequana XH2000	CPU rack	Full load	52	47
RIVR 1.22	Bull Sequana XH2000	CPU rack	Full load	52	47
RIVR 1.23	Bull Sequana XH2000	CPU rack	Full load	52	47
RIVR 1.24	Bull Sequana XH2000	CPU rack	Full load	52	47
RIVR 1.25	Bull Sequana XH2000	GPU rack	Full load	57	51
RIVR 1.26	Bull Sequana XH2000	GPU rack	Full load	57	51
RIVR 2.1	Existing service rack	Virtu/login	Full load	17,5	15
RIVR 2.2	Existing service rack	Virtu/login	Full load	17,5	15

RIVR 2.3	Existing service rack	HCST	Full load	22	19
RIVR 2.4	Existing service rack	SERVICE/storage	Full load	11	10
RIVR 2.5	Existing service rack	Storage	Full load	10	9
RIVR 2.6	Existing service rack	Storage	Full load	10	9
RIVR 2.7	Existing service rack	Storage	Full load	10	9
RIVR 2.8	Existing service rack	Storage	Full load	10	9
UPC-C	Existing UPS system		Only losses considered	8	6
UPS-D	Existing UPS system		Only losses considered	30	22
UPS-E	New UPS system		Only losses considered	20	16
DryCool1	New adiabatic dry cooler		Partial load	15	7
DryCool2	New adiabatic dry cooler		Partial load	15	7
HA-C	Existing HA		Partial load	90	50
HA-D	Existing HA		Partial load Redundant operation	141	40
HO D1	Existing CRACK		Full load	6	4
HO D2.1	Existing CRACK		Partial load	3	2,5
HO D2.2	Existing CRACK		Partial load	3	2,5
HO D2.3	Existing CRACK		Partial load	3	2,5
HO D2.4	Existing CRACK		Partial load	3	2,5
HO C2.1	Existing CRACK		Partial load	3	2,5
HO C2.2	Existing CRACK		Partial load	3	2,5
HO C2.3	Existing CRACK		Partial load	3	2,5
HO C2.4	Existing CRACK		Partial load	3	2,5
OČ D1	Existing pump		Partial load	5	4
OČ D2	Existing pump		Partial load	4	3
OČ D3	Existing pump		Partial load	4	3
OČ D4	New pump		Partial load	5	4
OČ D5	New pump		Partial load	4	3
OČ D8	New pump		Partial load	4	3
OČ D6	New pump		Partial load	4	3
OČ D7	New pump		Partial load	4	3
OČ C1	Existing pump		Partial load	5	4

OČ C2	Existing pump		Partial load	4	3
OČ C3	Existing pump		Partial load		3
	Rest of consumers		Partial load	20	10
		Total Compute infrastructure		**742 kW**	**667 kW**
		Total Cooling infrastructure		**412 kW**	**218kW**
		TOTAL		**1154 kW**	**885 kW**

5.1.1.2 SUP2 [INHW] Electrical Power Design of Components

High level electrical design

Power distribution system is unchanged original system to which we add an additional UPS system – power line E. System is described on next picture. System is integrating a complete protection against electrical disturbances.

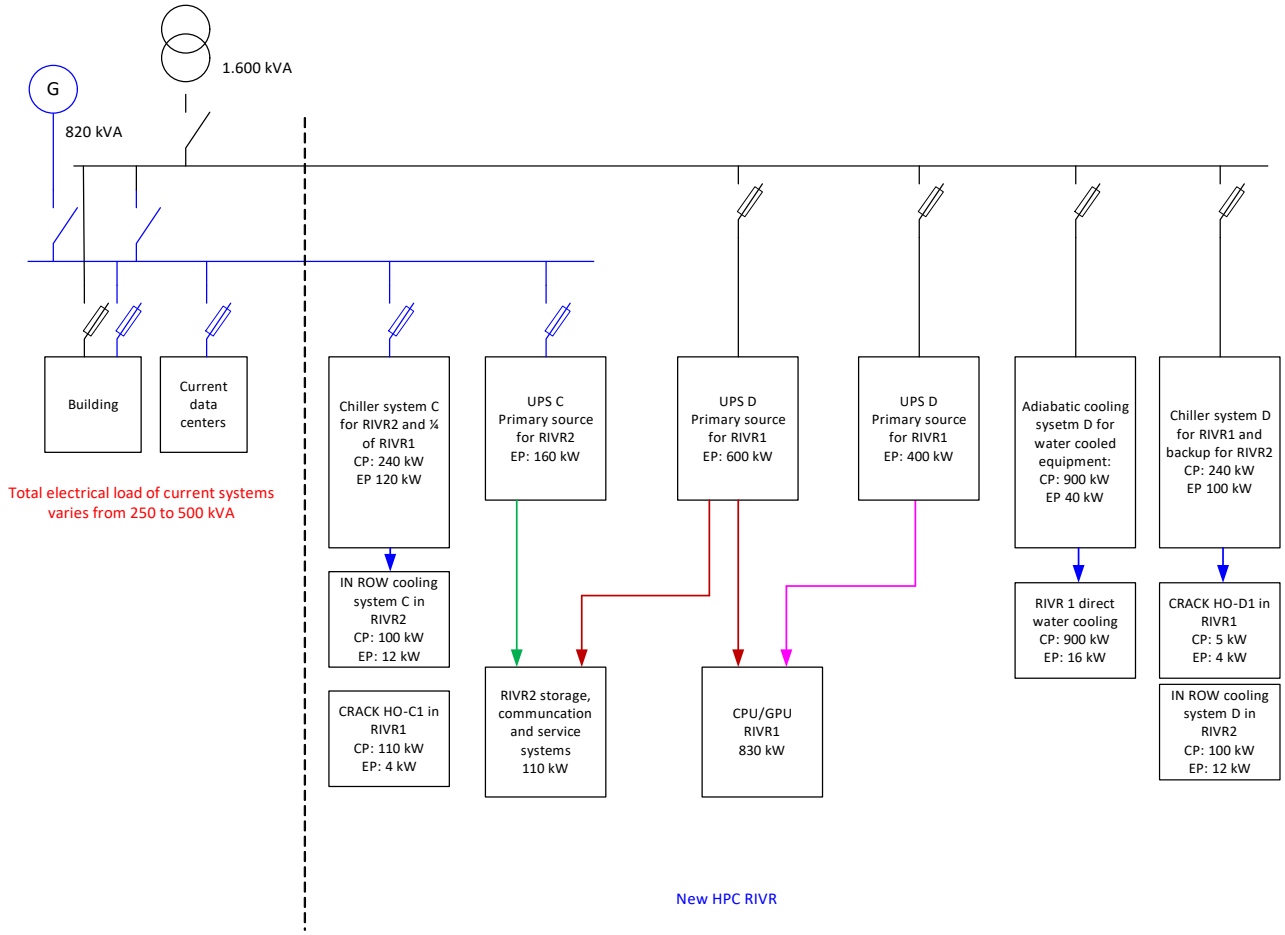


Figure 1 Power block diagram

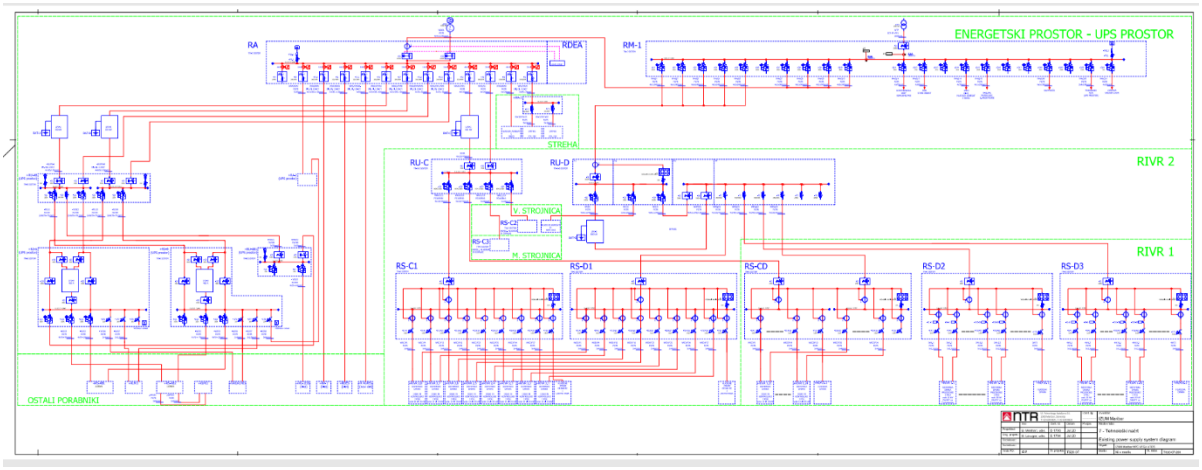


Figure 2 Existing Power supply diagram
(See Annex 10.2.4-Existing power supply system diagram.pdf)

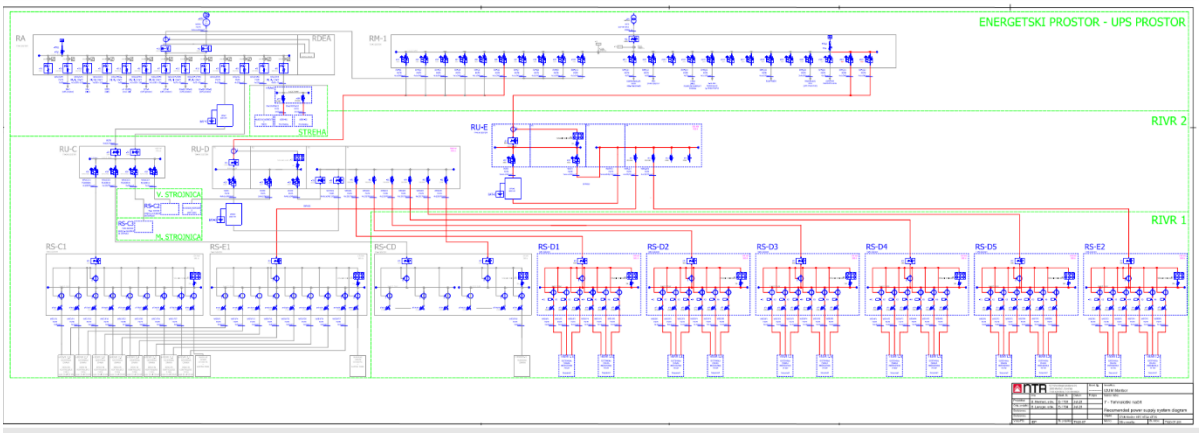


Figure 3 Recommended power supply diagram
(See Annex 10.2.5-Recomended power supply system design.pdf)

Key starting points for electrical power supply system are as follows:

Power:	230/400 V AC
Power Frequency:	50 Hz
Phases:	3
Power connectors:	IEC 60309 compliance
Electrical safety system:	as per attached plan
Dimensions of power cables:	as per attached plan

All other information on electrical power system are stated below.

Vega HPC Power Consumption

Table 1 HPC Power consumption

ID	Rack	Description	Cooling	Native Power	Weight
RIVR 1.11	Bull Sequana XH2000	CPU	DLC	69 kW	2033 kg
RIVR 1.12	Bull Sequana XH2000	CPU	DLC	69 kW	2033 kg
RIVR 1.13	Bull Sequana XH2000	CPU	DLC	69 kW	2033 kg
RIVR 1.14	Bull Sequana XH2000	CPU	DLC	69 kW	2033 kg
RIVR 1.15	Bull Sequana XH2000	CPU	DLC	69 kW	2033 kg
RIVR 1.16	Bull Sequana XH2000	GPU	DLC	70 kW	2033 kg
RIVR 1.21	Bull Sequana XH2000	CPU	DLC	69 kW	2033 kg
RIVR 1.22	Bull Sequana XH2000	CPU	DLC	69 kW	2033 kg
RIVR 1.23	Bull Sequana XH2000	CPU	DLC	69 kW	2033 kg
RIVR 1.24	Bull Sequana XH2000	CPU	DLC	69 kW	2033 kg
RIVR 1.25	Bull Sequana XH2000	CPU	DLC	69 kW	2033 kg
RIVR 1.26	Bull Sequana XH2000	GPU	DLC	70 kW	2033 kg
RIVR 2.1	Existing	Virtualisation /login	Air	17,5 kW	500 kg
RIVR 2.2	Existing	Virtualisation /login	Air	17,5 kW	500 kg
RIVR 2.3	Existing	HCST	Air	22 kW	500 kg
RIVR 2.4	Existing	Service /storage	Air	11 kW	800 kg
RIVR 2.5	Existing	Storage	Air	10 kW	800 kg
RIVR 2.6	Existing	Storage	Air	10 kW	800 kg
RIVR 2.7	Existing	Storage	Air	10 kW	800 kg
RIVR 2.8	Existing	Storage	Air	10 kW	800 kg
			Total	**938 kW**	

VEGA HPC Cooling Power Consumption

Table 2 List of main Datacenter infrastructure

ID	Type	Native Power	Purpose
UPS-E	New UPS system, P=400kW	P_{el}=400kW	Uninterrupted power supply, ca. 4% loss of output power, partial load
DryCool 1	New Adiabatic dry cooler	P_{hl}=630kW	Partial load at all types of operations
DryCool 2	New Adiabatic dry cooler	P_{hl}=630kW	Partial load at all types of operations
HO D1	Existing CRACK	P_{hl}=200kW	Partial load at all types of operations due to redundancy and over-estimation
OČ D4	New Pump		Partial load due to estimation
OČ D5	New Pump		Partial load due to estimation
OČ D8	New Pump		Partial load due to estimation
OČ D6	New Pump		Partial load due to estimation
OČ D7	New Pump		Partial load due to estimation
RU-E	Main Switch block		Main distribution for E UPS supply line
RS-D1	Switch block		Switch block
RS-D2	Switch block		Switch block
RS-D3	Switch block		Switch block
RS-D4	Switch block		Switch block
RS-D5	Switch block		Switch block
RS-E2	Switch block		Switch block

Table 3 Energy consumption of HPC support infrastructure

ID	Type	Native Power	Purpose	Peak power consumption [kW]	Full workload consumption [kW]
UPC-C	Existing UPS System	P_{el}=160kW	Uninterrupted power supply, ca. 4% loss of output power, partial load	8	6
UPS-D	Existing UPS System, P=600kW	P_{el}=600kW	Uninterrupted power supply, ca. 4% loss of output power, partial load	30	22

UPS-E	New UPS System, P=400kW	P_{el}=400kW	Uninterrupted power supply, ca. 4% loss of output power, partial load	20	16
DryCool 1	New Adiabatic dry cooler	P_{hl}=630kW	Partial load at all types of operations	15	7
DryCool 2	New Adiabatic dry cooler	P_{hl}=630kW	Partial load at all types of operations	15	7
HA-C	Existing HA	P_{hl}=230kW	Partial load at all types of operations due to redundancy and over-estimation	90	50
HA-D	Existing HA	P_{hl}=500kW	Partial load at all types of operations due to redundancy and over-estimation	141	40
HO D1	Existing CRACK	P_{hl}=200kW	Partial load at all types of operations due to redundancy and over-estimation	6	4
HO D2.1	Existing CRACK	P_{hl}=33kW	Partial load due to redundancy	3	2,5
HO D2.2	Existing CRACK	P_{hl}=33kW	Partial load due to redundancy	3	2,5
HO D2.3	Existing CRACK	P_{hl}=33kW	Partial load due to redundancy	3	2,5
HO D2.4	Existing CRACK	P_{hl}=33kW	Partial load due to redundancy	3	2,5
HO C2.1	Existing CRACK	P_{hl}=33kW	Partial load due to redundancy	3	2,5
HO C2.2	Existing CRACK	P_{hl}=33kW	Partial load due to redundancy	3	2,5
HO C2.3	Existing CRACK	P_{hl}=33kW	Partial load due to redundancy	3	2,5
HO C2.4	Existing CRACK	P_{hl}=33kW	Partial load due to redundancy	3	2,5
OČ D1	Existing Pump		Partial load due to estimation	5	4
OČ D2	Existing Pump		Partial load due to estimation	4	3

OČ D3	Existing Pump		Partial load due to estimation	4	3
OČ D4	New Pump		Partial load due to estimation	5	4
OČ D5	New Pump		Partial load due to estimation	4	3
OČ D8	New Pump		Partial load due to estimation	4	3
OČ D6	New Pump		Partial load due to estimation	4	3
OČ D7	New Pump		Partial load due to estimation	4	3
OČ C1	Existing Pump		Partial load due to estimation	5	4
OČ C2	Existing Pump		Partial load due to estimation	4	3
OČ C3	Existing Pump		Partial load due to estimation	4	3

Switch blocks

The existing switch blocks need to be upgraded or replaced. The list of blocks which needs an upgrade or replacement is as follows:

- RM-1 – switch block of network supply, upgrade with 2 new splits (for RU-E),
- RU-E – switch block of line E of UPS supply, new delivery and integration,
- RS-D1 – switch block of line D of UPS supply, new delivery and integration,
- RS-D2 – switch block of line D of UPS supply, new delivery and integration,
- RS-D3 – switch block of line D of UPS supply, new delivery and integration,
- RS-D4 – C
- RS-D5 – switch block of line D of UPS supply, new delivery and integration,
- RS-E1 – switch block of line D RS-D1 is renamed and source of supply is changed (from RU-E), switch block stays same,
- RS-E2 – switch block of line E of UPS supply, new delivery and integration,

Switch Blok RM-1

This is a switch block which is a part of NN field of transformation station that provides distribution of network supply to existing and new consumers. It containes source field with air supply switch and outline field with own cable space. New split lines wil lbe added to this block: - Q21 and -Q22.

Basic switch block info	
Basic switch block info	RM-1
Supplier	Legrand
Electrical equipment	Legrand
Level of internal assigment	I
Instalation system	TN-C
Operational Native Power	U= 400VAC/230VAC, 50 Hz
Native Voltage of service currents	U=24V DC
Install power	Pi= 1730 kVA,
Sync factor	1
Peak Power	Pk= 1730 kVA,
Power Factor	Cos fi=0.9
Native current	In= 2500A
Short current resistance	Icu= 50 kA
Mechanical protection	IP21
Electrical measurment	On supply line with thermical protection of transformator
Cable supply line	From floor

Switch block RU-E

Switch block is intended for central supply of UPS line E between switch blocks at single system rooms.

Basic switch block info	
Block name	RU-E
Supplier	EATON
Electrical equipment	EATON
Level of internal assigment	I
Instalation system	TN-C-S
Operational Native Power	U= 400VAC/230VAC, 50 Hz
Native Voltage of service currents	U=24V DC
Install power	Pi= 300 kVA,

Sync factor	1
Peak Power	Pk= 300 kW
Power Factor	Cos fi=1
Native current	In= 433 A
Short current resistance	Ik = 433 A
Mechanical protection	Icu= 50 kA
Electrical measurment	IP21
Cable supply line	From floor

Switch block RS-E1 (prej RS-D1)

Switch block for supply of UPS power line E.

Basic switch block info	
Block name	RS-E1
Supplier	Rittal
Electrical equipment	Eaton
Level of internal assigment	I
Instalation system	TN-S
Operational Native Power	U= 400VAC/230VAC, 50 Hz
Native Voltage of service currents	U=24V DC
Install power	Pi= 125 kVA,
Sync factor	1
Peak Power	Pk= 125 kVA,
Power Factor	Cos fi=1
Native current	In= 181A
Short current resistance	Icu= 10 kA
Mechanical protection	IP21
Electrical measurment	On supply line and all outputs
Cable supply line	From floor

Switch block RS-D1

Switch block for supply of UPS power line D.

Basic switch block info	
Block name	RS-D1
Supplier	Rittal
Electrical equipment	Eaton
Level of internal assigment	I
Instalation system	TN-S

Operational Native Power	U= 400VAC/230VAC, 50 Hz
Native Voltage of service currents	U=24V DC
Install power	Pi= 140 kVA,
Sync factor	1
Peak Power	Pk= 140 kVA,
Power Factor	Cos fi=1
Native current	In= 202A
Final current	Ik = 202 A
Short current resistance	Icu= 10 kA
Mechanical protection	IP21
Electrical measurment	On supply line and all outputs
Cable supply line	From floor

Switch block RS-D2

Switch block for supply of UPS power line D.

Basic switch block info	
Block name	RS-D2
Supplier	Rittal
Electrical equipment	Eaton
Level of internal assigment	I
Instalation system	TN-S
Operational Native Power	U= 400VAC/230VAC, 50 Hz
Native Voltage of service currents	U=24V DC
Install power	Pi= 140 kVA,
Sync factor	1
Peak Power	Pk= 140 kVA,
Power Factor	Cos fi=1
Native current	In= 202A
Final current	Ik = 202 A
Short current resistance	Icu= 10 kA
Mechanical protection	IP21
Electrical measurment	On supply line and all outputs
Cable supply line	From floor

Switch block RS-D3

Switch block for supply of UPS power line D.

Basic switch block info	
Block name	RS-D3
Supplier	Rittal
Electrical equipment	Eaton
Level of internal assigment	I
Instalation system	TN-S
Operational Native Power	U= 400VAC/230VAC, 50 Hz
Native Voltage of service currents	U=24V DC
Install power	Pi= 140 kVA,
Sync factor	1
Peak Power	Pk= 140 kVA,
Power Factor	Cos fi=1
Native current	In= 202A
Final current	Ik = 202 A
Short current resistance	Icu= 10 kA
Mechanical protection	IP21
Electrical measurment	On supply line and all outputs
Cable supply line	From floor

Switch Block RS-D4

Switch block for supply of UPS power line D.

Basic switch block info	
Block name	RS-D4
Supplier	Rittal
Electrical equipment	Eaton
Level of internal assigment	I
Instalation system	TN-S
Operational Native Power	U= 400VAC/230VAC, 50 Hz
Native Voltage of service currents	U=24V DC
Install power	Pi= 140 kVA,
Sync factor	1
Peak Power	Pk= 140 kVA,
Power Factor	Cos fi=1
Native current	In= 202A
Final current	Ik = 202 A

Short current resistance	Icu= 10 kA
Mechanical protection	IP21
Electrical measurment	On supply line and all outputs
Cable supply line	From floor

Switch block RS-D5

Switch block for supply of UPS power line D.

Basic switch block info	
Block name	RS-D5
Supplier	Rittal
Electrical equipment	Eaton
Level of internal assigment	I
Instalation system	TN-S
Operational Native Power	U= 400VAC/230VAC, 50 Hz
Native Voltage of service currents	U=24V DC
Install power	Pi= 140 kVA,
Sync factor	1
Peak Power	Pk= 140 kVA,
Power Factor	Cos fi=1
Native current	In= 202A
Final current	Ik = 202 A
Short current resistance	Icu= 10 kA
Mechanical protection	IP21
Electrical measurment	On supply line and all outputs
Cable supply line	From floor

Switch block RS-E2

Switch block for supply of UPS power line E.

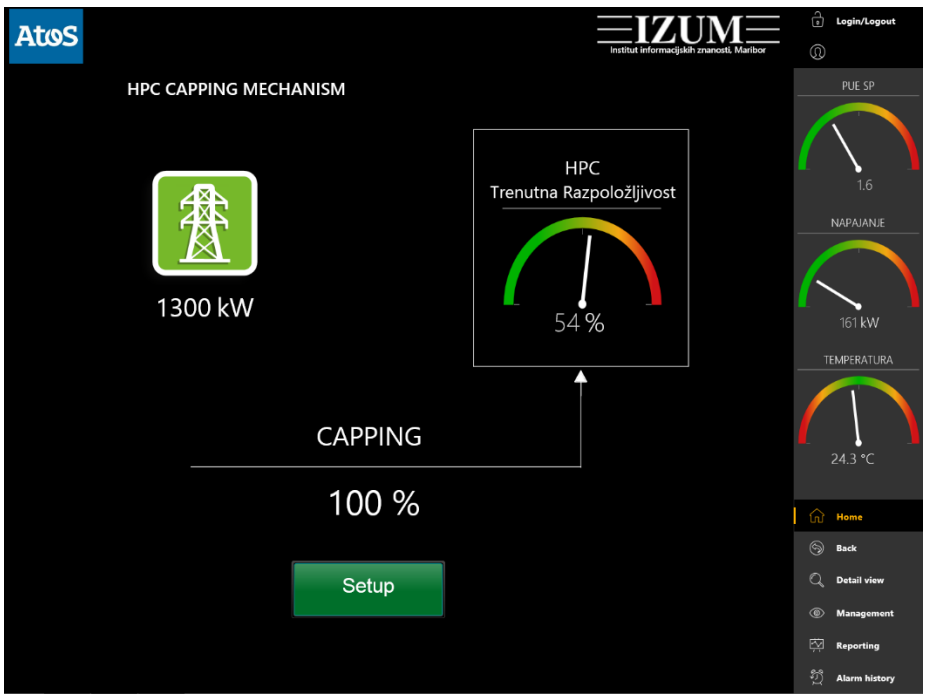
Basic switch block info	
Block name	RS-E2
Supplier	Rittal
Electrical equipment	Eaton
Level of internal assigment	I
Instalation system	TN-S
Operational Native Power	U= 400VAC/230VAC, 50 Hz
Native Voltage of service currents	U=24V DC
Install power	Pi= 140 kVA,

Sync factor	1
Peak Power	Pk= 140 kVA,
Power Factor	Cos fi=1
Native current	In= 202A
Final current	Ik = 202 A
Short current resistance	Icu= 10 kA
Mechanical protection	IP21
Electrical measurment	On supply line and all outputs
Cable supply line	From floor

5.1.1.3 SUP3 [INHW] Automatic Power Consumption Capping Mechanism

Atos offers an Automatic Power Consumption Capping Mechanism that will integrate a graphical monitoring system and ability to detect automatically Power peaks and cap them according to rules set by Administrator. The system will monitor and manage Datacenter and HPC Infrastructure.

Automatic power consumption capping mechanism as part of HPC DCIM management system will be implemented. The system will provide appropriate power consumption capping regarding working conditions. In case of any error, high temperature, malfunction or exceed of power capacities on power supply equipment, UPS, transformer station, power management starts to limit processor power at procured HPC system and decrease power demands



5.1.1.4 SUP4 [INHW] Failure Detection and Reaction

Atos solution will provide system-internal mechanism to enable real-time infrastructure related environment changes (e.g., leakage, pressure drops or temperature changes) and react in (near-) real time, in such a way that damage of the procured system is prevented.

The system will also include alerting system with a history record enabling full reporting

The existing DCIM Monitoring system will be upgraded with HPC Vega Management and Alert system, which will enable additional monitor features for the new HPC System:
In first phase it is intended to manage Datacenter Infrastructure, but it also features an additional feature like monitoring, alerting, power management, and status reporting.
Upgrade of existing DCIM system will ensure a central management and alerting system for Datacenter and HPC Solution.

One of the key functionalities will be detection and reaction to failures that is integrated in Alarm management system. DCIM Control system will with the help of sensors and plug-ins detect the failure and respond accordingly. The system also manages an alert system (to users or Administrators) and adjustment of electrical and thermical settings of the whole infrastructure with ability to stop, close, decrease, increase of critical systems if needed. This is valid for existing and new equipment and important in case of major failures (exp. Outage or overheating of main transformator, outage of cooling system etc.). There are several levels of failure detection and ability to react, The HPC Solution (BullSequana XH2000) integrates own HW and SW monitoring system, that controls crucial levels of operation (inlet temp, inlet pressure, node temperature, Water flow etc.) while BullSequana Rack CDU provides also HW 2+1 redundant protection. The system can detect and protect HW from node level to a complete rack level in case of overheat or outage of cooling elements detected on customer inlet loop. System is also able to adapt Power consumption (on CPU performance level) in case of detection of failures or thresholds close to critical values.

The supply infrastructure with DCIM management system will system-internal mechanism to detect, in real time, infrastructure related environment changes (e.g., leakage, pressure drops or temperature changes) and react, in (near-) real time, in such a way that damage of the procured system is prevented

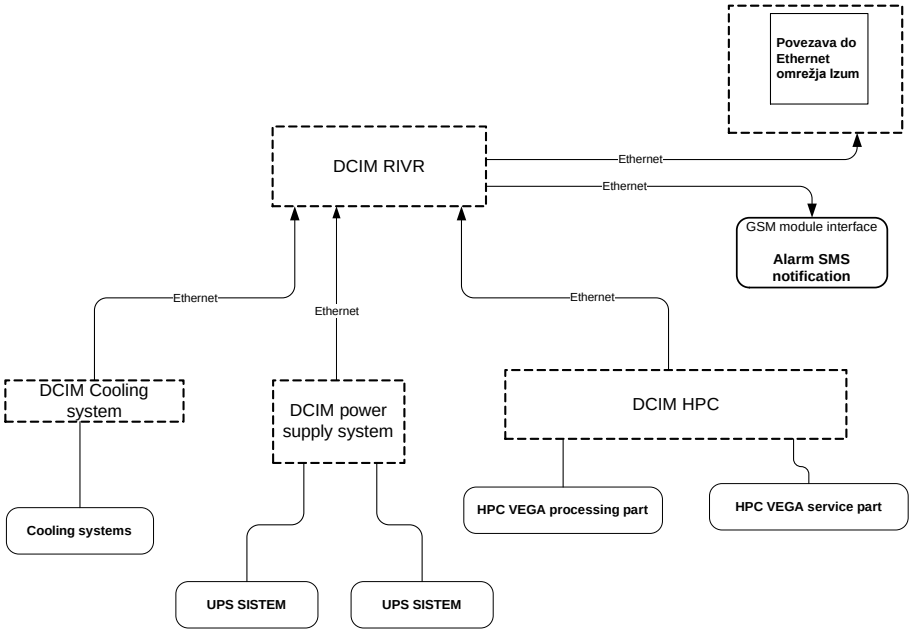


Figure 4 Block diagram of DCIM management system

5.1.1.5 SUP5 [INHW] Communication with Facility

The whole HPC Vega solution is made of three main parts: HPC compute partitions, HPC Service partition and Storage, and Datacenter Infrastructure. Each part integrates intelligent elements that enables to communicate with the central management system. The existing central management system DCIM can integrate these elements to ensure a central management point for the whole solution. Integration can then be done based on protocols like SNMP, Modbus TCP or customized API interfaces.

The elements of the HPC Vega solution will provide communication protocols to communicate with the existing DCIM monitoring system (exchange of data including events and alarms).

5.1.1.6 SUP6 [INHW] Work Execution and Supply Infrastructure Separation

All works on HPC Vega will be carried out in a way that the function of the exiting data center and equipment will not be compromised at any time. All support infrastructure will be separated and timing and execution plan prior to works will be prepared and coordinated with customer.

A key element for the supply of the Vega HPC Solution is the delivery of the cooling and power solution in the Datacenter based on clear and validated plans.

Atos and its subcontractor will ensure that all works, and deliverables will be executed in accordance with local laws and regulations.

All works and services are planned to be performed in a way that existing Computer infrastructure of IZUM will not require to be disconnected.

In order to assure such approach Atos will prepare a detailed plan of works and deliverables which includes:

- Plan of execution phases that includes protection of existing IZUM infrastructure
- Plan of protection of existing IZUM Computer room
- Timeline
- Plan of risk mitigation regarding execution of works on common infrastructure

5.1.1.7 SUP7 [INHW] Energy and Efficiency Measurement Separation

System Management Control system (DCIM) controls and manages recurrent power consumption on all active equipment and links. Current and historical power consumption data are available on every level of infrastructure. With this we can assure measurement of power consumption and energy efficiency of existing computer center as well as the new HPC Vega solution.

The DCIM monitoring system will provide power consumption measurement separate for supercomputer and separate for the rest of the facility. The system will provide measure energy efficiency of the procured system and appropriate reporting mechanism.

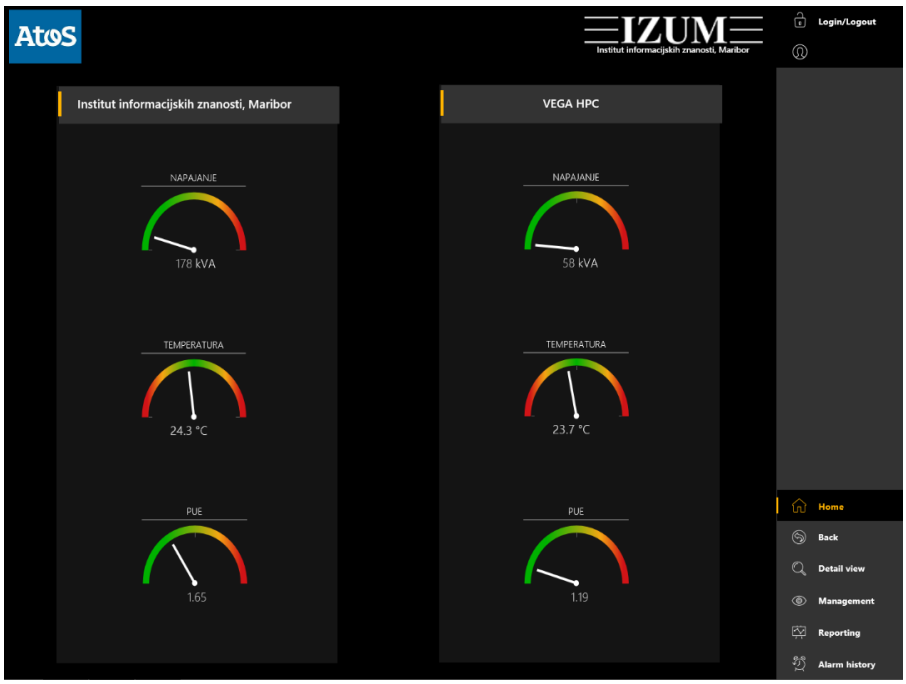


Figure 5 Power management dashboard

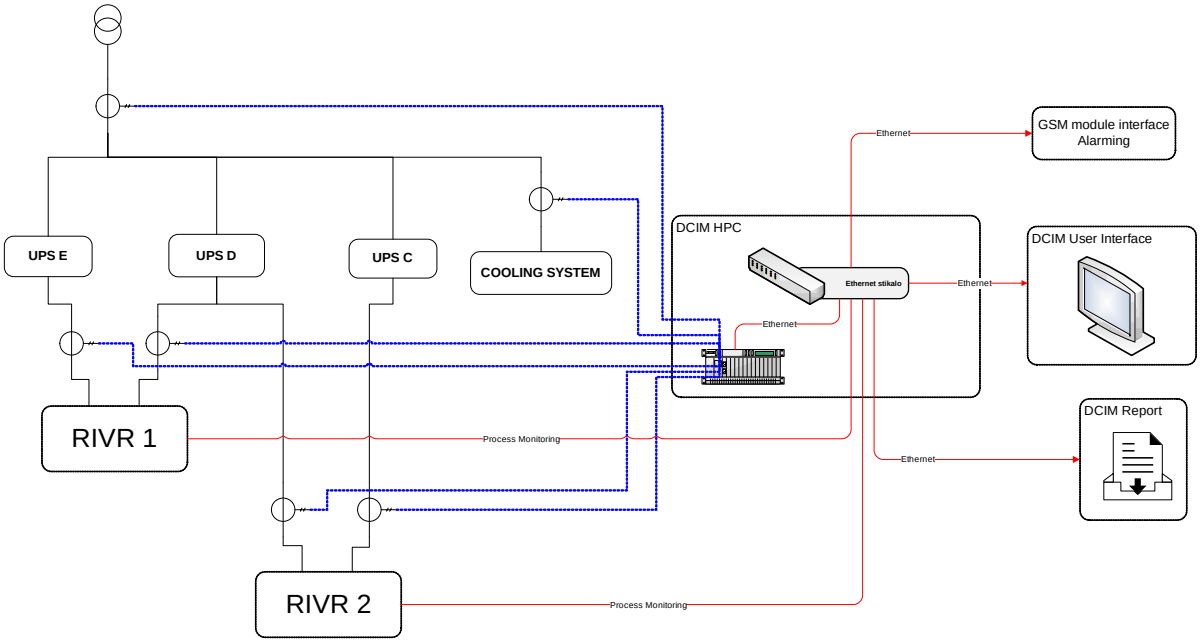


Figure 6 Power measurement system block diagram

Cene enot		
Obračunska moč	2,96561	EUR
Omrežnina VT	0,00723	EUR
Omrežnina MT	0,00557	EUR
Prispevek OVE+SPTE	4,10702	EUR
Prispevek za energetsko učinkovitost	0,0008	EUR
Prispevek za delovanje op.trga	0,00013	EUR
Energija VT	0,05899	EUR
Energija MT	0,0395	EUR
Trošarina - III. stopnja	0,00305	EUR

Razdelitev po specifikaciji k računu številka: 99001084-20050020-3

Datum od: 01.05.2020
Datum do: 31.05.2020

HPC				**Faktor razdelitve: 0,0439***
Tar. postavka	**Poraba**	**EM**	**Cena enote**	**Znesek v EUR**
Obračunska moč	6,41	kW	2,96561	19,01 €
Omrežnina VT	1.704,07	kWh	0,00723	12,32 €
Omrežnina MT	2.159,13	kWh	0,00557	12,03 €
Prispevek OVE+SPTE	6,41	kW	4,10702	26,32 €
Prispevek za energetsko učinkovitost	3.863,20	kWh	0,0008	3,09 €
Prispevek za delovanje op. trga	3.863,20	kWh	0,00013	0,50 €
Uporaba omrežja in prispevki za SODO d.o.o.:				**73,27 €**
Energija VT	1.704,07	kWh	0,05899	100,52 €
Energija MT	2.159,13	kWh	0,0395	85,29 €
Električna energija:				**185,81 €**
Trošarina - III. stopnja	3.863,20	kWh	0,00305	11,78 €
Trošarina in dodatni prispevek:				**11,78 €**
SKUPNI ZNESEK obračuna električne energije za obdobje 01.05.2020 - 31.05.2020.				**270,86 €**

OSTALO				**Faktor razdelitve: 0,9561***
Tar. postavka	**Poraba**	**EM**	**Cena enote**	**Znesek v EUR**
Obračunska moč	139,59	kW	2,96561	413,97 €
Omrežnina VT	37.112,93	kWh	0,00723	268,33 €
Omrežnina MT	47.023,87	kWh	0,00557	261,92 €
Prispevek OVE+SPTE	139,59	kW	4,10702	573,30 €
Prispevek za energetsko učinkovitost	84.136,80	kWh	0,0008	67,31 €
Prispevek za delovanje op. trga	84.136,80	kWh	0,00013	10,94 €
Uporaba omrežja in prispevki za SODO d.o.o.:				**1.595,77 €**
Energija VT	37.112,93	kWh	0,05899	2.189,29 €
Energija MT	47.023,87	kWh	0,0395	1.857,44 €
Električna energija:				**4.046,73 €**
Trošarina - III. stopnja	84.136,80	kWh	0,00305	256,62 €
Trošarina in dodatni prispevek:				**256,62 €**
SKUPNI ZNESEK obračuna električne energije za obdobje 01.05.2020 - 31.05.2020.				**5.899,12 €**

Figure 7 Example of power consumption report

5.1.2 Floor Space Arrangement Requirements

5.1.2.1 FLO1 [INHW] Maximum Area Load at RIVR1 room

The proposed floorplan is considering all technical aspects of the hardware and the floorspace available for the proposed solution.

Our proposed floorplan is as follows:

- RIVR 1 – main HPC room proposed for HPC computing part, with the following equipment:
  - 10 x HPC Direct Water-cooled racks Atos BullSequana XH2000 with CPU based nodes,
  - 2 x HPC Direct Water-cooled racks Atos BullSequana XH2000 with GPU based nodes,
  - Electrical distribution cabinets for electrical power supply in RIVR R1
  - Cooling Rack for Heat dissipated in Air
- RIVR 2 – HPC Service Room with HPC service partition and storage with the following equipment:
  - Existing service racks,
  - Existing cooling racks (In-Row coolers)
  - Existing Power Distribution Equipment
- New UPS room due to increase of Power of HPC system:
  - New additional 400kW UPS
  - Batteries for new UPS
  - Switch board for power distribution of new UPS
- Cooling racks for purposes of room cooling
  - External chillers space (on top of building roof)
  - System D external adiabatic units

Maximum Area load in RIVR1 will not exceed 3000 kg. Weight of Equipment, weight of raised floor, weight of load construction, weight of distribution systems of cooling and power is all considered in this calculation. Weight of personnel has also been considered.

Weight of one rack:	2100 kg
Area of one rack:	0,9 m^2
Weight of raised floor per m^2:	120 kg
Weight of distribution units-cooling per m^2:	250 kg
Weight of distribution units-power per m^2:	80 kg
Spare weight per m^2:	200 kg

5.1.2.2 FLO2 [INHW] Documentation of Floor Plan and Block Diagrams

RIVR 1 Layout

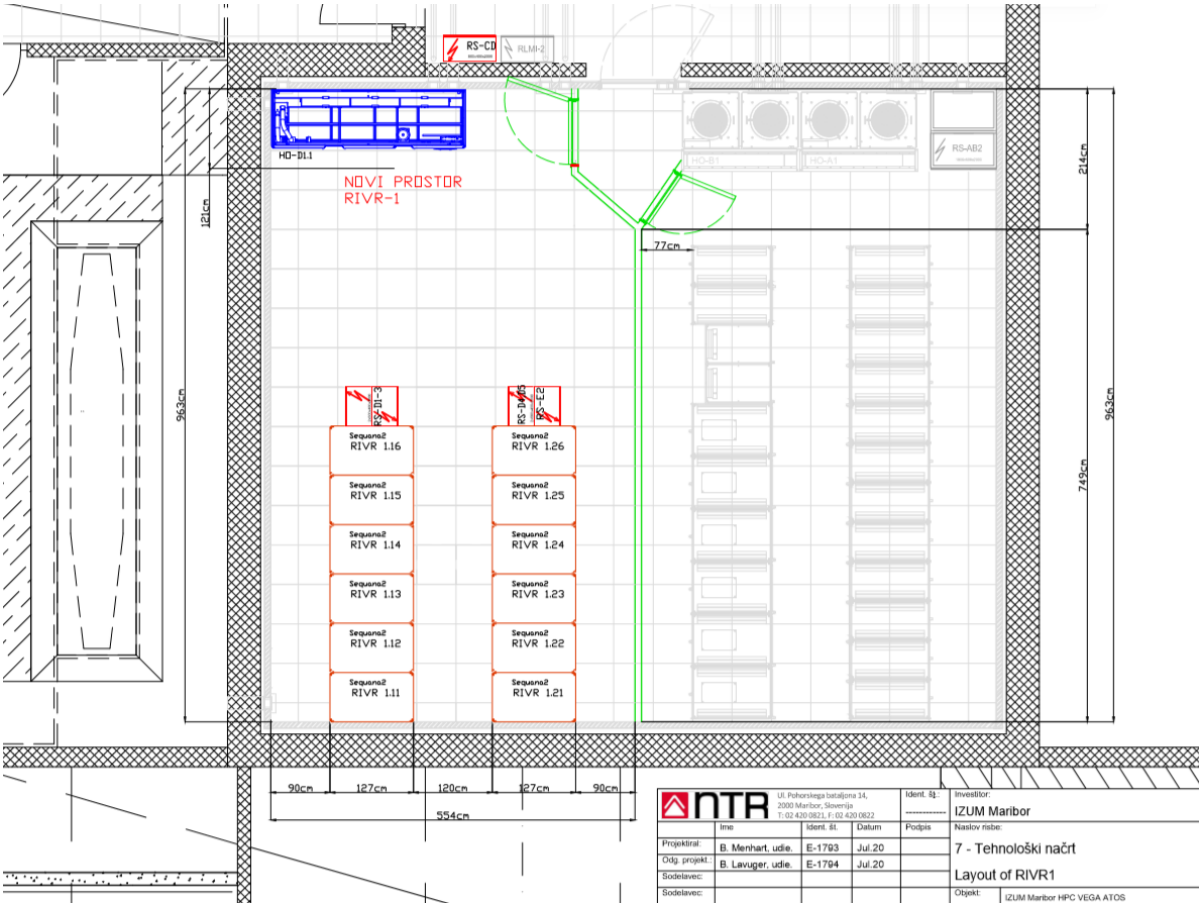


Figure 8 RIVR1 Floorplan

RIVR 2 Layout

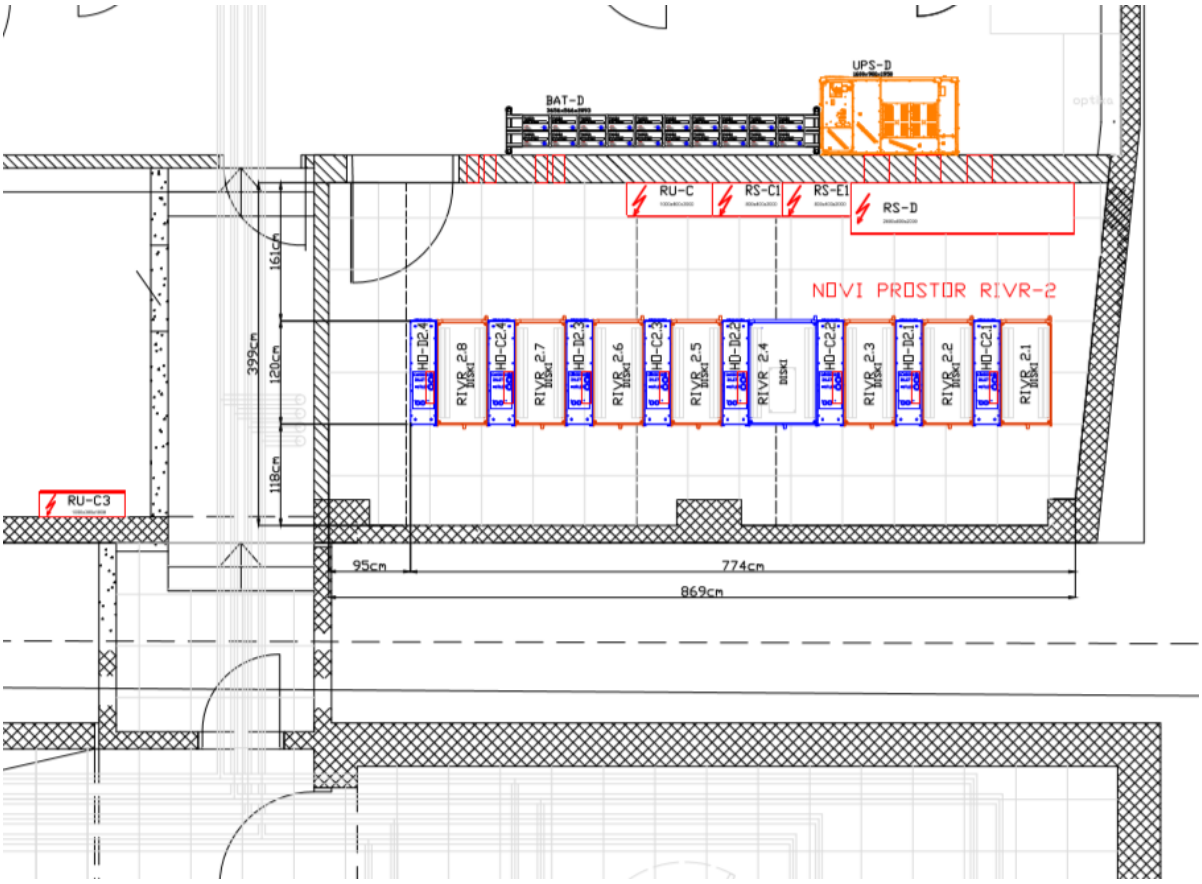


Figure 9 RIVR2 floorplan

Datacenter Infrastructure Layout

Basic floorplan of Datacenter infrastructure will remain the same. Due to the need for additional power consumption we plan to install two new dry coolers on the roof of the building, replacing existing ones. In the new room we plan to install power units of new power supply line E, as shown on next two pictures:

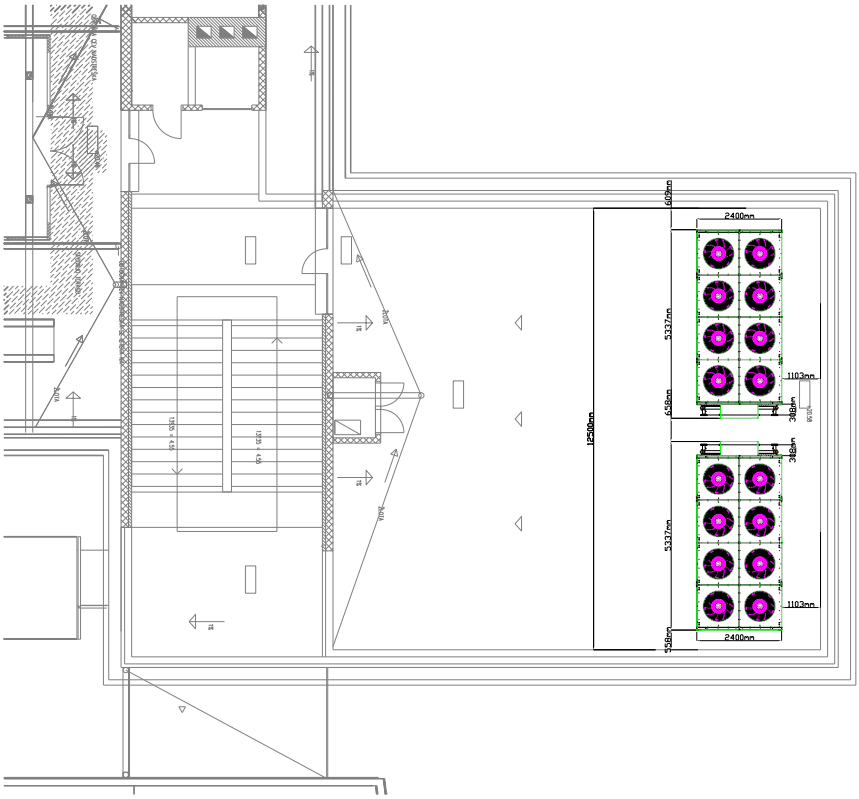


Figure 10 Floorplan of the roof with new units

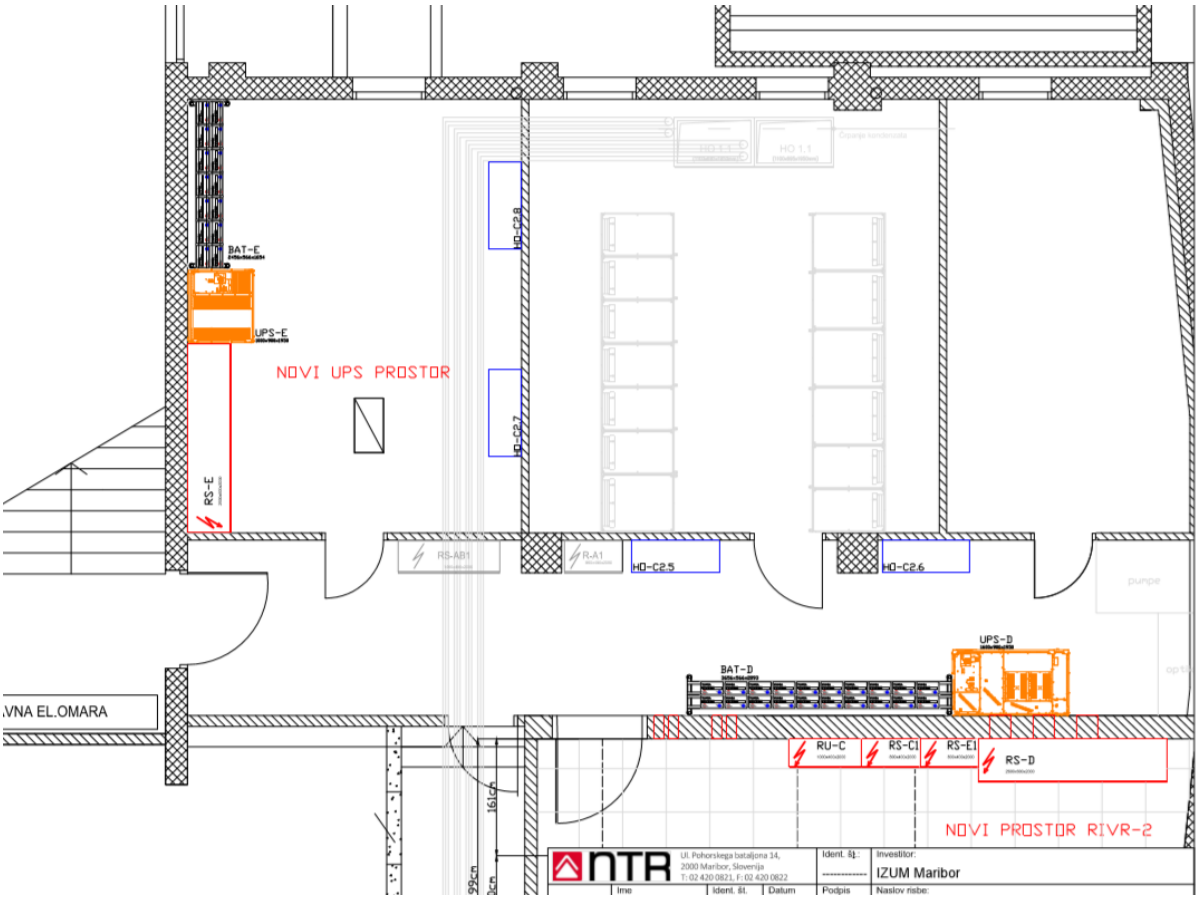


Figure 11 Floorplan of the new UPS room

5.1.3 Electrical Infrastructure Requirements

5.1.3.1 ELE1 |INHW] Compliance of Power Distribution for Connectors and Power Quality

Distribution of power will be installed with Siemens BD2 distribution system, as shown on next picture:

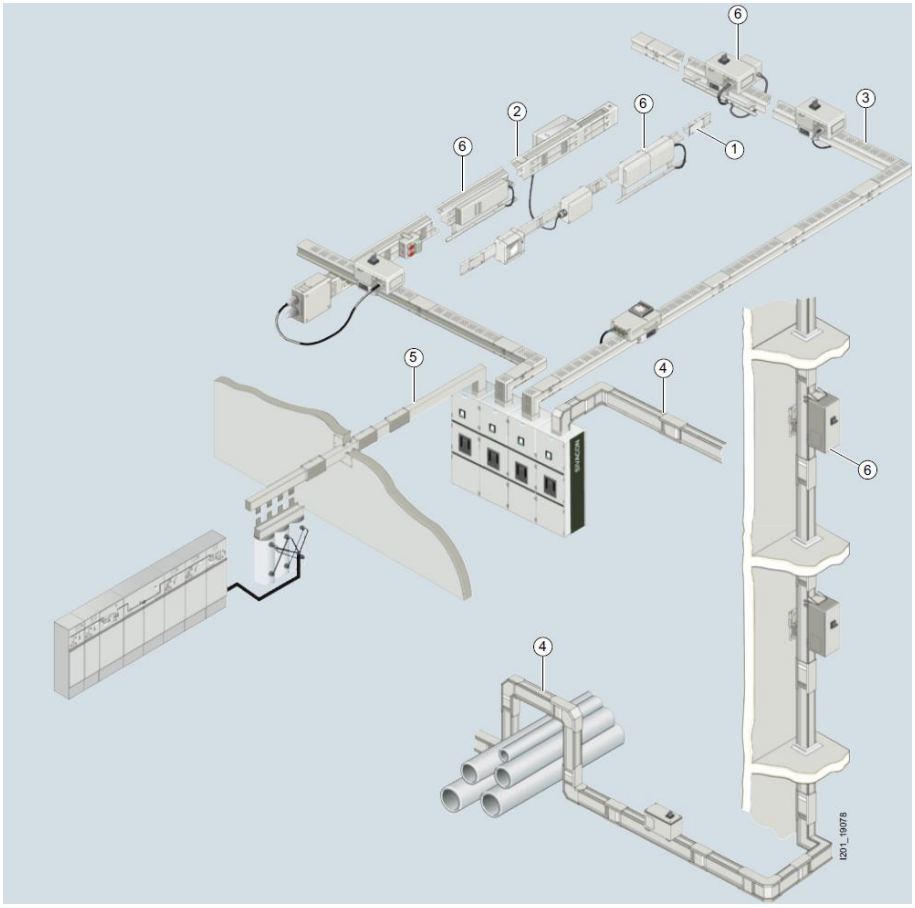


Figure 12 Power distribution system

The power supply of the system will comply with the DIN-VD and EN standards at the time of installation. Connectors will conform to IEC 60309/DIN EN 60309. All security and other standards are in technical description. Power connectors and power distribution units are marked with EU conformity declaration and certificate for EC marking. All units are conforming with the Low Voltage Directive 2014/35/EU. The power supply of the procured system will comply with power quality Level 2 and accordance SIST EN 61000-2-4.

5.1.3.2 ELE2 |INHW] Protection Against Main Disturbances

An additional UPS system is offered in order to protect against Power disturbances.

UPS power line E is a monolith UPS system, that can operate also in parallel configurations (for possible future upgrades). UPS system, native power of 400 kVA, ensures and provides power to electrical consumers. UPS-E line will be integrated together with batteries in new UPS room as per our block diagrams (called 'arhiv'). UPS is providing Level 2 quality of Power Distribution according to SIST EN 61000-2-4. As UPS system works under principal of double changeover it assures separation of high harmonic distortions and it complies with the level of requested power distribution. Offered UPS is Vertiv, model EXL S 400. With the following specification:

- Native power output 400 kVA,
- Monolit technology,
- Technology with no transformators,
- Native power 400V / 50Hz (3 phase) + N,
- Native frequency 50 Hz.
- power factor (cos FI) >0.99,
- native current min 588 A,
- nominal operating power 400 kVA,
- Peak power 400 kW,
- Power stability in accordance with IEC/EN 62040-3, Class 1,
- Short current min 2,2 x In for 200 ms,
- integrated IGBT power switch elements in inverter, converter, and battery charger,
- operation in VFI, VFD and VI regime according to IEC32040-3,
- support for parallel operation,
- allowed battery power voltage 396 to 700 V DC,
- VRLA, WET and NiCd battery support,
- Battery charge current with 240 cells >160A,
- VFI regime efficiency more than 96% at output load range 25% – 100%,
- Overload capacity:
  - 125% for 10 minutes
  - 150% for 1 minutes
  - 700% for 600 ms
  - 1000% for 100 ms
- dimensions 1950 x 750 x 900 mm,
- Cable line input from lower end,
- Integrated static bypass switch,
- Integrated service bypass switch,
- Integrated forced cooling system,
- Operating temperature up to 40C,
- Includes 2 min autonomy batteries at 100% load,
- Includes earthquake resistant battery stands, cables and switches and other safety elements

Liebert® EXL S1 Specifications

UPS RATING (KVA)	100	120	160	200	300	400	500	600	800	1000	1200
Output active power at 35 °C*(kW)	100	120	160	200	300	400	500	600	800	1000	1200
Output active power at 40 °C (kW)	90	108	144	180	270	360	450	540	720	900	1080
INPUT											
Nominal mains input voltage / voltage range* (V)	400 (250 to 460), 3Ph or 3Ph + N										
Nominal bypass input voltage / voltage range* (V)	400 (380/415 selectable), 3Ph or 3Ph + N										
Nominal frequency / frequency tolerance (Hz)	50±10%(60 selectable)										
Input Power Factor	≥ 0.99										
Input current distortion (THDi) (%)	≤3										
OUTPUT											
Nominal output voltage (V)	400 (380/415 selectable), 3Ph or 3Ph + N										
Nominal output frequency (Hz)	50 (60 selectable)										
Output voltage stability by load variation 0-100% (%)											
- static	±1										
- dynamic	Complies with IEC/EN 62040-3, Class 1										
Output frequency stability											
- synchronized with bypass mains (%)	±2 (2, 3, 4, 5 selectable)										
- synchronized with internal clock (%)	±0.1										
Inverter Overload Capacity*	110% continuous, 125% for 10mins, 150% for 1min										
Short circuit current for 200 ms (%)	2.2 In										
Load crest factor handled without derating the ups (Ipk/Irms)	3:1										
Compatibility with loads	Any power factor (leading or lagging) up to 1										
BATTERY											
Permissible battery voltage range (V)	396 to 700										
Float voltage for VRLA @ 20 °C (V/cell)	2,27										
End cell voltage for VRLA (V/cell)	1.65										
Float Voltage stability in steady state condition (%)	≤1										
DC ripple voltage without battery (%)	≤1										
GENERAL AND SYSTEM DATA											
Classification according to IEC/EN 62040-3	VFI-SS-111										
Operating Temperature (°C)	0-40										
Maximum relative humidity @ 20 °C (non condensing) (%)	up to 95										
Protection degree with open doors	IP 20										
Frame colour (RAL scale)	7021										
Noise @ 1 metre as per ISO 3746 (dBA ± 2dBA)	65	65	66	68	69	71	73	76	76	78	78
Noise @ 1 metre as per ISO 3746 (dBA ± 2dBA) (at partial load)	64	64	65	65	65	65	65	70	70	72	72
Parallel configuration	up to 8 units in parallel										
Access	Front and Top (no rear access required)										
AC/AC efficiency:											
- VFI according to IEC/EN 62040 definition (%)	up to 97%										
- VFD according to IEC/EN 62040 definition (%)	up to 99%										
DIMENSION AND WEIGHT											
Height (mm)	1950										
Width (mm)	500		750		1000		1250		2000		2650
Depth (mm)	900										
Net Weight (kg)	370		510		725		990		1550		2275

*Conditions apply

Figure 13 UPS system E technical specification

EC DECLARATION OF CONFORMITY

VERTIV.

We, (Importer Manufacturer)
Vertiv S.r.l.
Address
Bologna plant: Via Fornace, 30
Castel Guelfo (BO), 40023, Italy

Declare that

Product:	UNINTERRUPTIBLE POWER SUPPLY
Model:	Liebert 80-eXL Product range 100kVA, 120kVA, 160kVA, 200kVA, 300kVA, 400kVA, 500kVA, 600kVA, 800kVA, 1000kVA, 1200kVA
Brand Name:	Vertiv

The object of the declaration is conformity with the relevant Union harmonization legislation:

- Low Voltage Equipment: Directive 2014/35/EU
- Electromagnetic Compatibility: Directive 2014/30/EU

The tests have been made in accordance to the procedures required by the European Council Directive and standards:

Conformity to the following standards:

For Low Voltage Directive (LVD): EN 62040-1:2008+A1:2013

For Electromagnetic Compatibility (EMC): EN 62040-2: 2006

The last two digits of the year in which the EC marking was affixed: 17'

This product(s) which are defined herein was (were) manufactured under the conditions of the European Union directive and standards.

Name and address of the Person authorized to compile the technical file:
Graziano Galuppi
Vertiv S.r.l.
Bologna plant: Via Fornace, 30 Castel Guelfo (BO), 40023, Italy

Signature .....
Name & Surname: **Graziano Galuppi**
Title: **Technical Director** Date: 03/04/2017

Figure 14 UPS Certification

5.1.3.3 ELE3 |INHW] Replacement of Power Distribution Units

Active Equipment and Racks will be connected with IEC 60309 complaint plugs, as shown on the next picture:

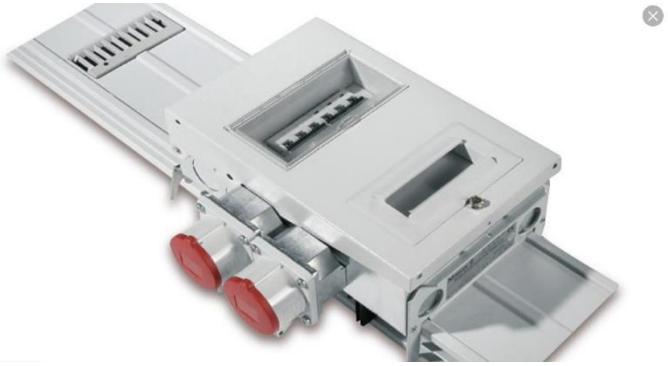
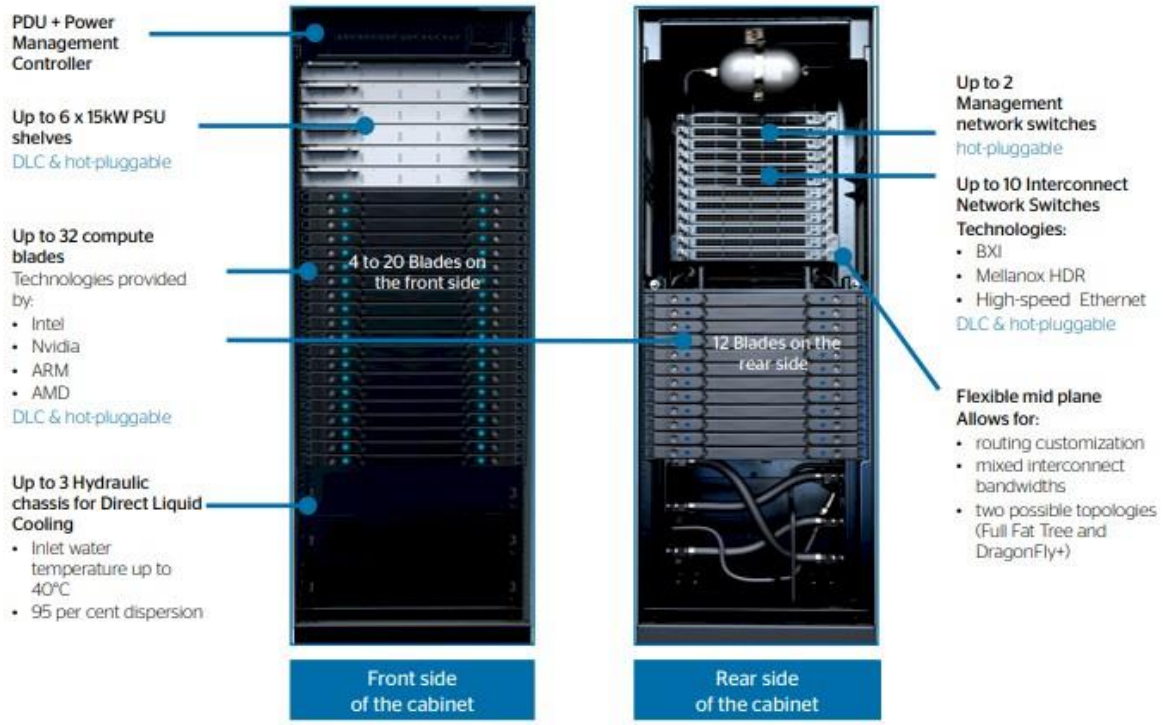


Figure 15 IEC 60309 compliant plugs

Each of three separate PDU units in Vega HPC rack is connected to separate outlet to provide possibility to replace PDU unit without interruption of operation of the rack.



BullSequana XH2000 has also internal redundant and hot plug ability to replace internal PDU and PSU with no need of shutting down the whole rack.

5.1.3.4 ELE4 |INHW] Electrical Infrastructure Documentation

We have described electrical infrastructure in chapter *5.1.1.2 SUP2 [INHW] Electrical Power Design of Components* of this document.

The proposed Electrical infrastructure is described with high level overview general block diagrams and description of main components. Due to limitations of size of pages for submissions, we are not attaching detailed descriptions of each components (pdf's), however all documentation is available for further clarification.

Atos will deliver a solution complying with supply of power and cooling needs for the proposed HPC Solution. All electrical components will comply with tender requirements and with local regulations and certifications.

Atos and its subcontractor will, for the purposes of Datacenter installation works, submit a complete detailed project documentation as follows:

- Intention plan (Idejna zasnova) - IDZ
- Execution project (Projekt za izvedbo) - PZI
- Project of executed works (Projekt izvedenih del) – PID
- Final documentation

These documents will contain all required details and will be validated by IZUM before any operations on site.

5.1.4 Cooling Requirements

5.1.4.1 COO1 |INHW] Free Cooling Requirement

HPC VEGA will operate in free cooling regime for most of the year. Main CPU part of the HPC VEGA will operate for up to 99% of time in freckling. Minor, storage part will operate up to 75% of time in free cooling regime.

Power of the CPU/GPU part of VEGA HPC:	830 kW
Free cooling temperature	up to 33 °C
Total year No. Of hours	8760 h
Total time of freckling:	8751 h
Percentage of freckling time	99,9%
Power of the support part of VEGA HPC:	108 kW
Free cooling temperature at normal operation	up to 17 °C
Total year No. Of hours	8760 h
Total time of free cooing:	6559 h
Percentage of free cooling time	74,9%
Free cooling temperature at emergency operation	up to 13 °C
Total year No. of hours	8760 h
Total time of free cooling:	3561 h

Percentage of free cooling time	59,4%
Maximum external temperature	according ASHRAE, up to 40 °C
Pumps and other motors:	all with frequency speed regulation

Temperature Regimes

Table 4 Temperature regimes

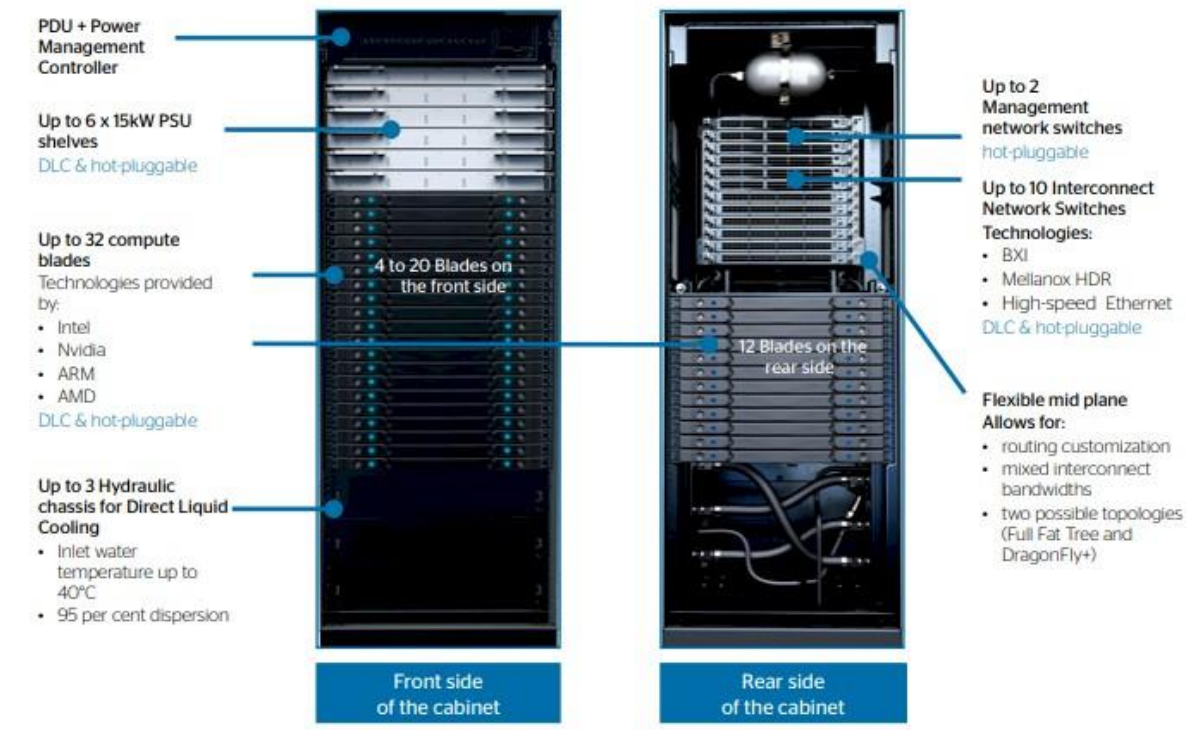
Cooling cicle	Operation	Cooling principles	Temperature regimee
Primary cooling of CPU/GPU	Normal operation	Water cooling	35/47 °C
Cooling of CPU/GPU air losess	Normal operation	Air cooling	20/26 °C
Cooling of storage/support systems	Normal operation	Air cooling	20/26 °C
Cooling of CPU/GPU air losess	Emergency operation	Air cooling	15/20 °C
Adiabatic cooling of water	Normal operation	Water cooling	35/50 °C

5.1.4.2 COO2 |INHW] Cooling System Maintenance

The HPC Vega cooling system is designed in such a way that the whole cooling system is equipped with additional pumps, CRACKs and pipes.

A replacement of single unit and single node is possible without interruption of working of other nodes. Same functionality and capabilities are utilized to change single rack or CRACK CDU.

The BullSequana XH2000 racks also integrate 3 Hydraulic Chassis for Direct Liquid Cooling of the racks. They operate as 2+1 redundant with Hot Plug replacement possibility.



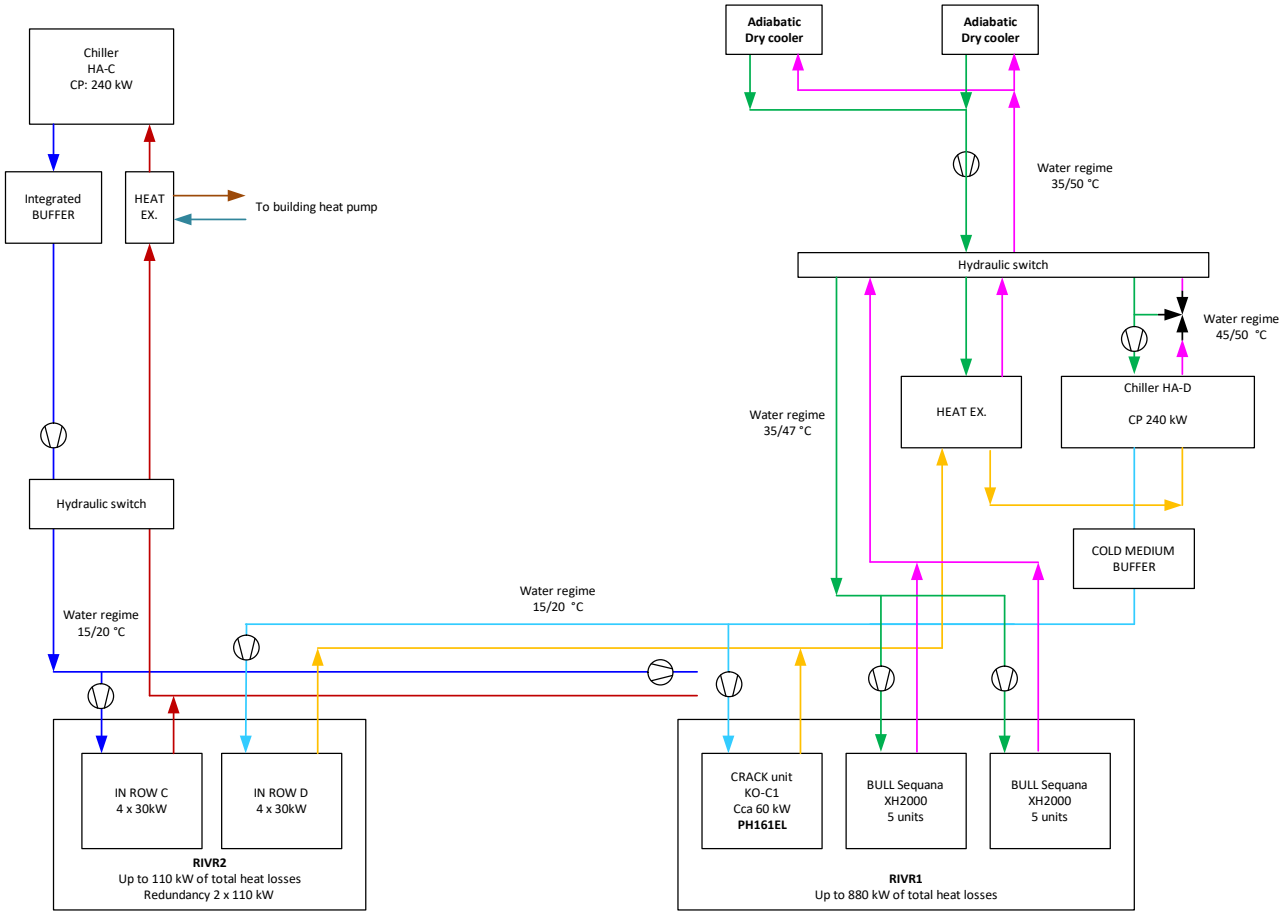


Figure 16 Cooling block diagram

Key Equipment Calculation

Table 5 Cooling power of key cooling infra

Equipment ID	Description	Type	Temperature regimee	Cooling power	Electrical power Working/total
DryCool1	Adiabatic Drycooler	ThermoKey SJGH2480	35/50 °C	629 kW	6912/15600 kW
DryCool2	Adiabatic Drycooler	ThermoKey SJGH2480	35/50 °C	629 kW	6912/15600 kW
HA-C	Freecooling chiller	Vertiv FB023	15/20 °C	216 kW	90 kW
HA-D	Liquid cooler	Vertiv WHT2043	47/50 to 20/26°C	240 kW	80 kW
HO D1	CRACK	Vertiv PH161EL	20/26°C	68 kW	5 kW
HO D2.1 to D2.4	IN ROW CRACK	Vertiv CR060RC	20/26°C	27,8kW	3 kW
HO C2.1 to DC.4	IN ROW CRACK	Vertiv CR060RC	20/26°C	27,8kW	3 kW
OČ D1	Pump with EC motor	Grundfoss			5 kW
OČ D2, D3	Pump with EC motor	Grundfoss			3,5 kW
OČ D4	Pump with EC motor	Grundfoss			4,5 kW
OČ D5, D8	Pump with EC motor	Grundfoss			3,5 kW
OČ D6, D7	Pump with EC motor	Grundfoss			4 kW
OČ C1	Pump with EC motor	Grundfoss			5 kW
OČ C2 and C3	Pump with EC motor	Grundfoss			3,5 kW

5.1.4.3 COO3 [INHW] Design Conditions for Cooling System

The cooling of the HPC Vega solution is designed to be efficient 365 days in the year no matter of external conditions.

All temperature design conditions of VEGA cooling system are prepared in accordance with ASHRAE design conditions and ASHRAE temperature profile for location Maribor,

Slovenia. The overall solution contains total redundant cooling in RIVR2 room with capacity up to 120 kW. In RIVR1 non-redundant cooling is designed.

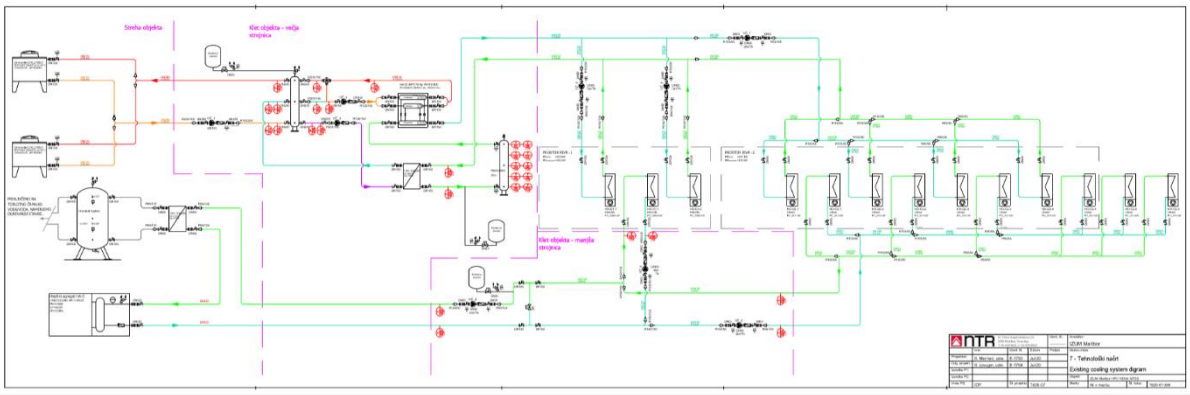


Figure 17 Existing Cooling System diagram
(See Annex 10.2.6-Existing cooling system digram.pdf

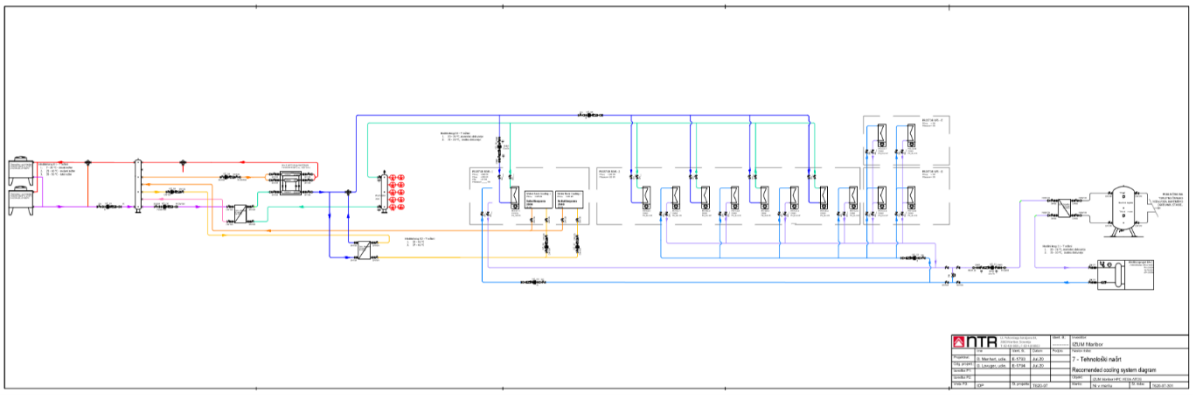


Figure 18 Recommended Cooling System diagram
(See Annex 10.2.7-Recomended cooling system diagram.pdf)

5.1.4.4 COO4 |INHW] Environmental Compatibility of Additives

In the cooling loops Melaled Solar additive will be used. It is nontoxic and biodegradable additive for freezing prevention. All necessary safety measures (e.g., leakage preventing catch basins) will be considered

Cooling Fluid:	Melaled® Solar
Bio-degradable:	Yes
Dangerous to people:	Not at normal usage
Dangerous to environment:	No

5.1.4.5 COO5 |INHW] Admissible Room Air Temperature

The overall cooling solution will be able to operate with a data center room temperature on cold side of 20°C to 26°C.

Admissible cold air room temperature: 20-26°C or according ASHRAE A1 class (18-32°C)

5.1.4.6 COO6 |INHW] Cooling System Management and Power Measurement

The entire HPC Vega supply infrastructure system will be integrated into the existing DCIM control system. This will make possible to connect and transfer data between existing systems, both dedicated to HPC and common to the entire facility, and consequently transfer data between individual systems.

These are the basic conditions for effective information transfer and control and management of operations. Part of the systems is common, and part is separate between the HPC and the rest of the facility. And the operation of some systems affects the operation of others. Uniform management ensures efficient management and automatic action in the event of unexpected events and alarms. Such a design of management and control work enables the operation of an automatic consumption adjustment system, which is necessary on the existing facility due to energy, space and other limitations.

The management of the cooling system is also an integral part of the DCIM control system. It monitors and detects the operation of all devices and in the event of emergencies, failures or faults via the power adjustment system, if necessary, reduces or even shuts down the entire HPC processor part in exceptional situations. This ensures the necessary protection of equipment and devices against damage that could result from errors, external events or other exceptional situations.

Within monitoring and remote-control systems, protection functions will be installed. They provide protection of IT equipment in case of any error, high temperature or malfunction of cooling systems.

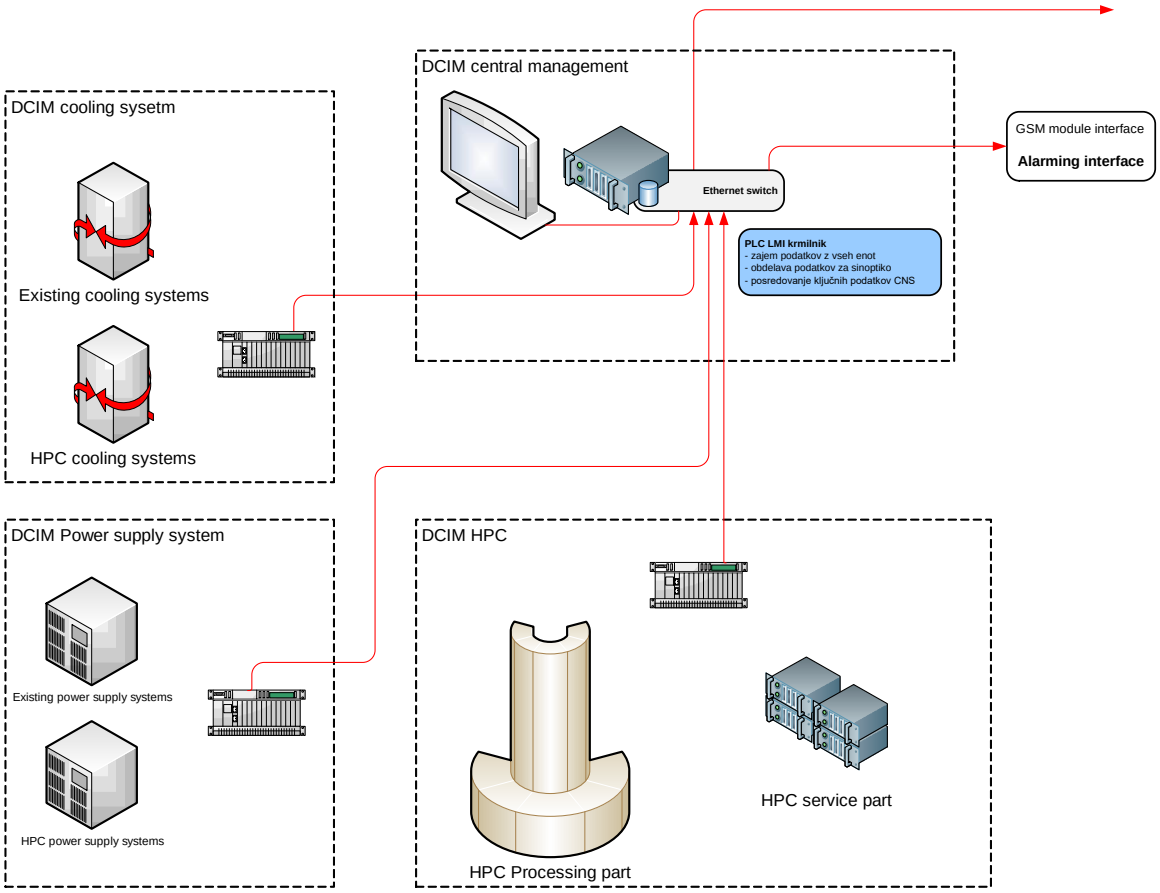


Figure 19 Block diagram of HPC management system

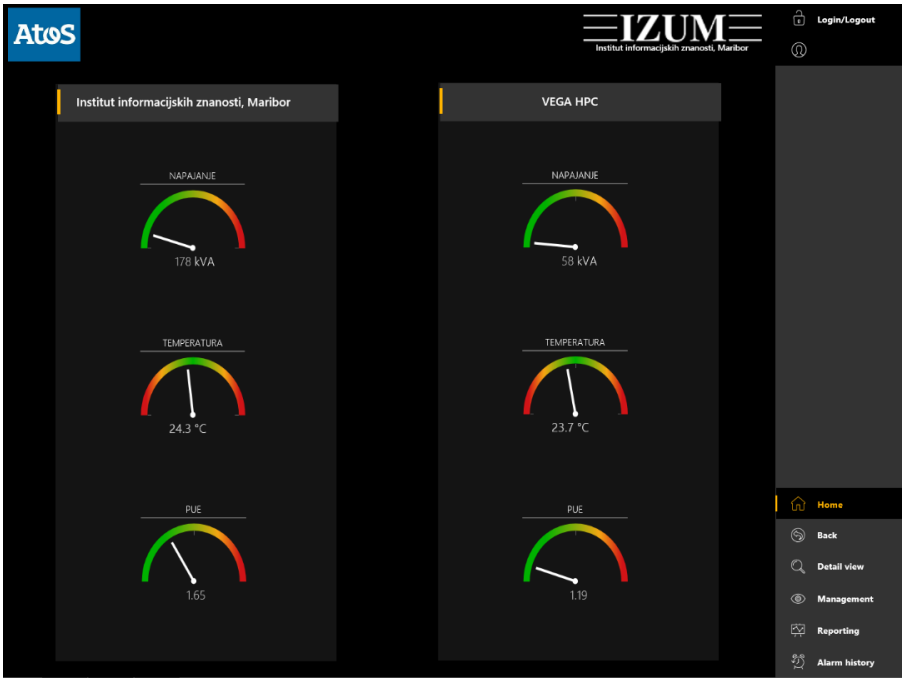


Figure 20 Cooling system and Power Measurement

Further details are also described in chapter 5.2 regarding the energy management system.

5.1.4.7 COO7 |INHW] Freeze Protection

With Melaled Solar additive we will provide adequate freezing protection, below minus 20 °C. All external cooling systems will be protected against freezing with appropriate antifreeze liquid.

Cooling Fluid:	Melaled® Solar
Recommended concentration:	38%
Freezing point:	- 21°C

5.2 Energy efficiency and power management and capping requirements

Controlling energy consumption is one of the main roadblocks on the path to Exascale, in addition to be a major environmental issue. Our ultra-energy-efficient platform called BullSequana XH2000 (and at the core of the proposed solution for the Vega system) lies at the heart of Bull's Exascale development program for HPC.

To complete the hardware, our R&D department developed a software product dedicated to power management for HPC clusters, **Bull Energy Optimizer (BEO)**, which is part of our proposition.

Bull Energy Optimizer includes a SLURM plugin integration that allows the correlation of Job and Energy consumption. The BEO Dashboard provides easy monitoring.

It is mainly designed to:

- **collect and store** power and temperature metrics for the hardware components of a HPC cluster (nodes, switches, storage, ...)
- **compute** the energy consumption and cost for selectable time periods
- **display** the collected and computed data as curves, charts, gauges or raw time series, with a component-oriented or job-oriented approach
- **monitor** the power, energy and temperature evolution by raising alerts on defined thresholds
- **limit** the power consumption by capping the power of components (when possible)

BEO is now in version 3.0 which brings several new features:

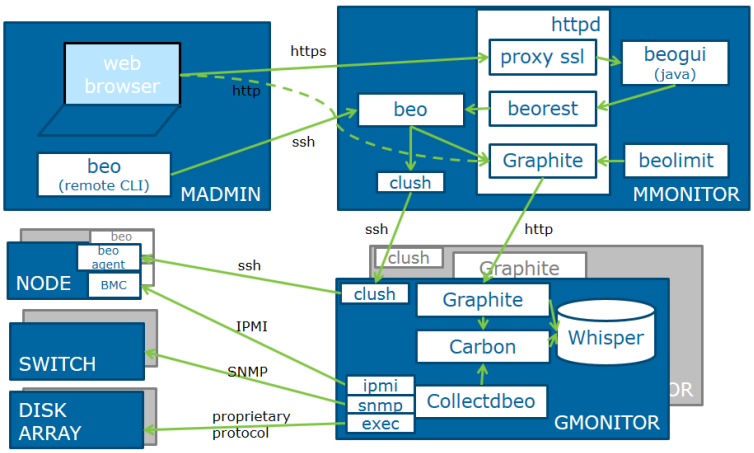
- Introduction of a GUI (based on eXtreme Factory XCS framework) offering:

- secured access web interface with 2 roles (admin/user)
- 3 predefined dashboards (hardware inventory, power dashboard, job dashboard)
- configurable graphical elements in dashboards to explore the collected data

- Power capping enhancements (CLI)
  - better display of power capping capabilities/status of components
  - enhanced accuracy of capping values
  - capping status reported in slurm *node features* (scontrol)
  - preventing from applying a power capping on nodes currently in use in a job

- New CLI commands
  - beo metrics: show collectable metrics for each component model
  - beo get energy: display energy time series
  - beo get temp: display temperature time series

- Support of new hardware
  - bullx R423e4m Sequana management node
  - BullSequana hydraulic pump (HYC)



Bull Energy Optimizer (BEO) Architecture

BEO is based on two main components:

- Front-end
  - GUI client (browser)
  - remote CLI
- Back-end (3 levels)
  - Monitoring
    - GUI web server (XCS)
    - BEO REST API Server

        - CLI + monitoring
        - Graphite server
    - Monitoring Gateway
        - Graphite/carbon server
    - Managed components
        - BMC / SNMP server / BEO agent

Bull Energy Optimizer (BEO) Added Value

Bull Energy Optimizer (BEO) is a software product dedicated to power management for HPC clusters.

Globally, BEO collects and stores Power and Temperature metrics in the format of time series for almost all the hardware components of the HPC clusters.
This is done completely "Out of Band": there is no perturbation of running applications.
The collects are done thanks to the following protocols:

- IPMI for compute nodes
- SNMP for interconnect switches
- Exec of proprietary command for disk arrays

Based on that information, BEO can calculate another metric which is the Energy consumed over the time.

When SLURM job scheduler is used, BEO can provide energy consumption related each job and includes the consumption of the compute nodes, the switches and the disk arrays.

Finally, BEO also includes an infrastructure-oriented Power Capping feature which lets the Administrator run the cluster or part of the cluster **under Power constraint.**

Bull Energy Optimizer (BEO) Packaging

BEO is an Atos proprietary closed source package. BEO is delivered in the frame of our overall model where we contribute to Open source communities, assemble, deliver and support Open source components and differentiate on targeted proprietary components.

BEO is delivered under a specific license agreement and sold as an annual subscription fee including L2 and L3 support service.

How to use BEO

BEO is accessible in two ways:

- A command Line Interface (CLI);
- A Graphical User Interface.

CLI can:

- Provides a hierarchical view of the components of the cluster (tree view).
- Provides power, temperature and energy metrics related to any subset of the cluster. Such a subset can group Compute Nodes, Service Nodes, Switches, Chassis, Rack, Islet, and cluster.
  - Provides also aggregation features (sum, min, max, avg)
  - Provides historical data (time series)
- Provides energy consumption related to any set of Slurm jobs.
  - Includes compute nodes, switches and disk arrays consumption
  - Includes all the job consumption, not only the consumption of the different steps of a job as provided by Slurm.
- Provides alerting functionality based on thresholds definition on a given set of metrics;
  - Alerts are reported into the syslog file of the management and can be easily adapted in an external monitoring tool.
- Provides ability to report energy costs, based on a configured description of power costs;
- Provides a Power Capping feature
  - Global power limit can be set to any subset of the cluster. Power Capping is done through RAPL-like solution on the computes nodes and consider the power consumption of all the different equipment to equitably share the power budget.

```
# beo report energy slurm94
| job     | state     | job.nodes.energy | job.switches.energy | job.da.energy | job.energy | job.trust | job.cost |
=============================================================================================================
| slurm94 | COMPLETED |       300.236 kJ |          25.0174 kJ |          0  J | 325.253 kJ |  100.00 % | 0.0106 € |
```

The GUI is accessible as admin but as regular user as well.

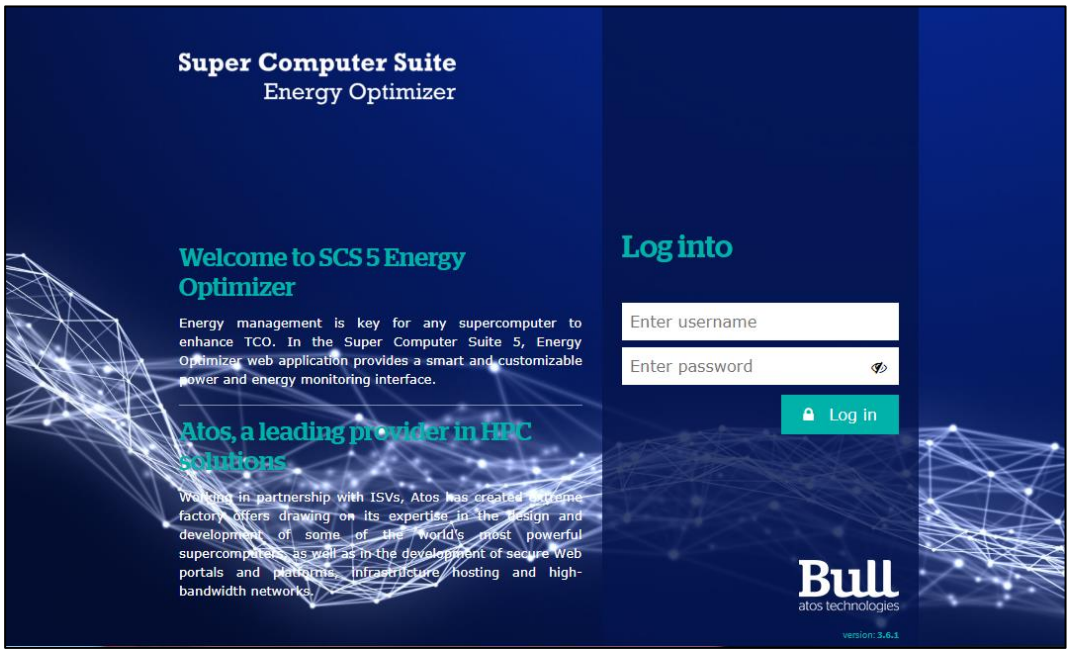


Figure 21: GUI Dashboard

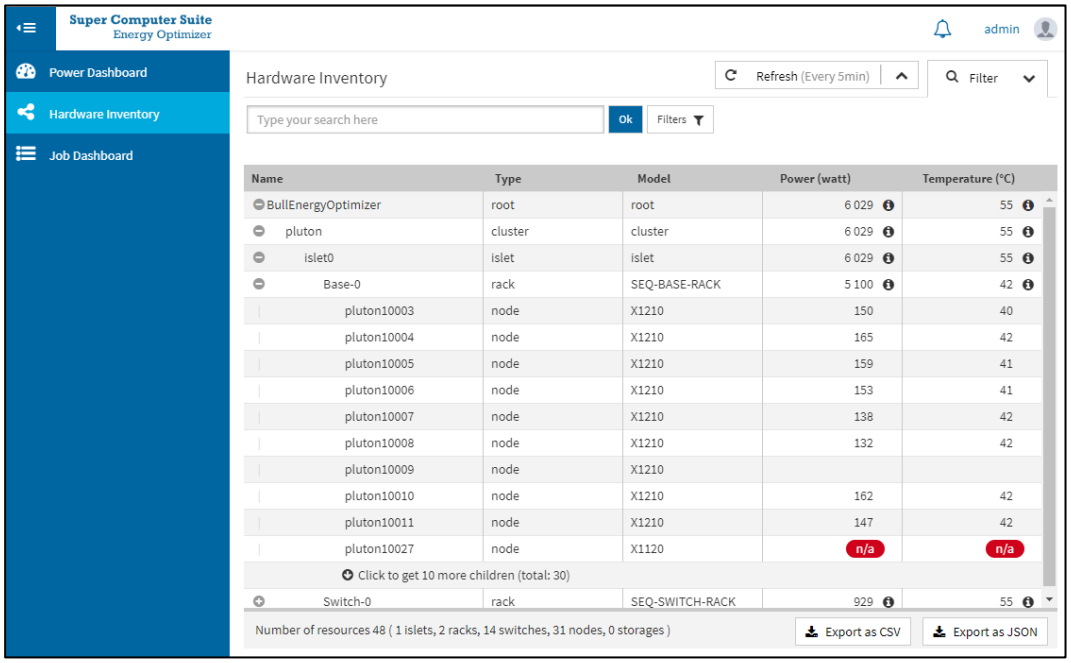


Figure 22: Hardware inventory

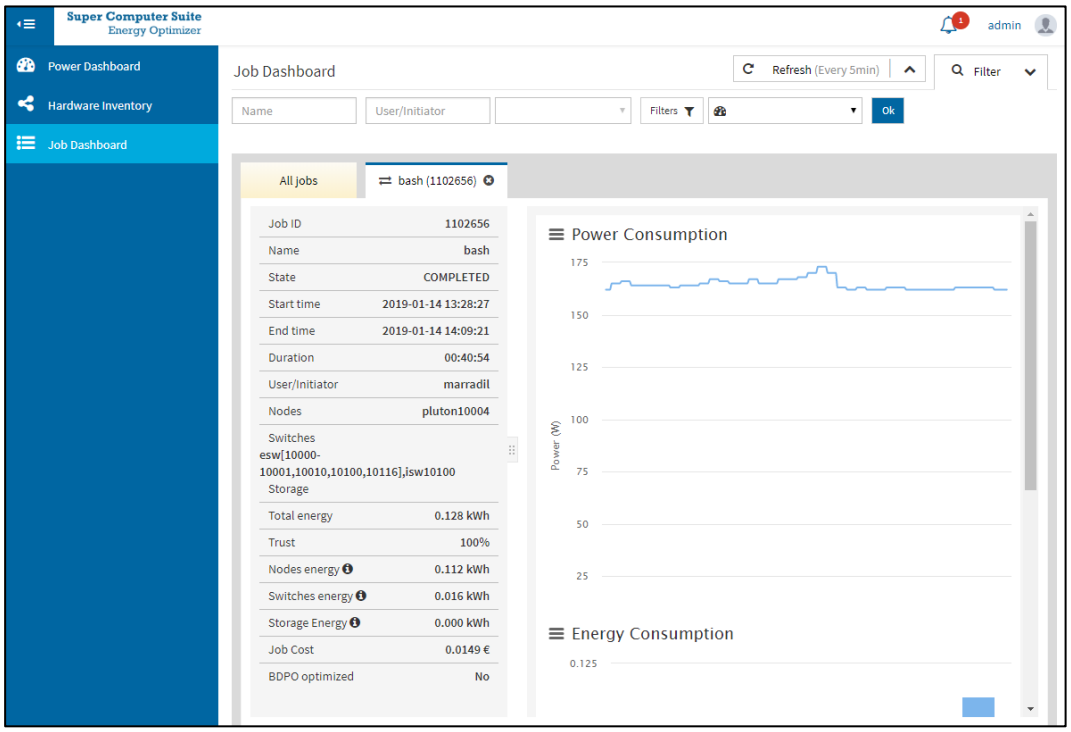


Figure 23: Power consumption information per job

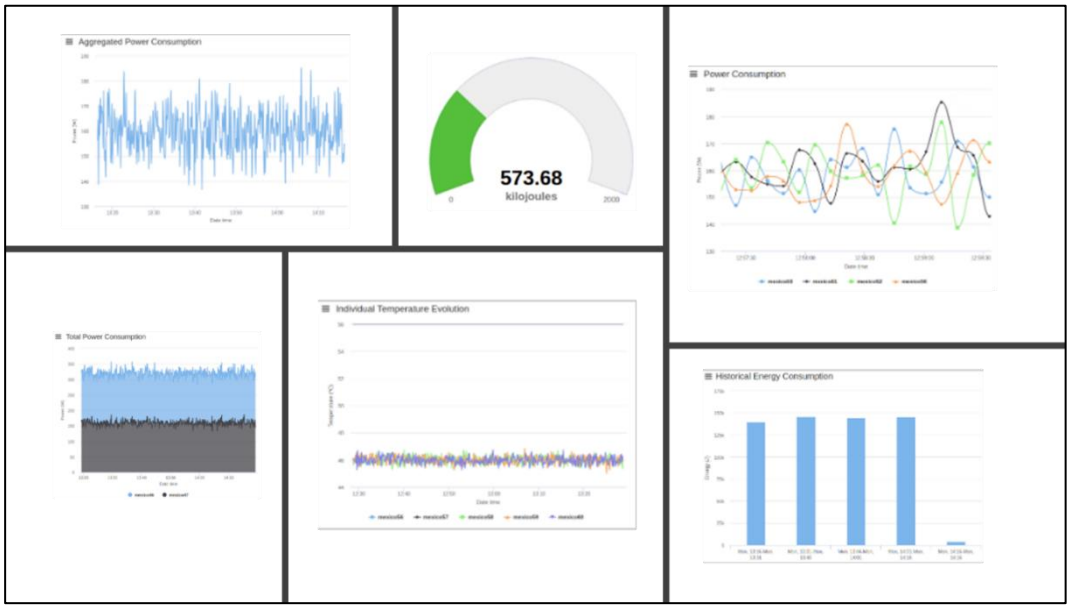


Figure 24: Different types of graph generated

Bull Energy Optimizer (BEO) Performance

The collecting of the metrics data is done out of band: there is no performance loss for the running applications.

Bull Energy Optimizer (BEO) Flexibility

By using standard protocols such as SNMP and IPMI BEO can easily support new hardware.
The low coupling design for getting SLURM information let's BEO be easily pluggable with another job scheduling system.

Bull Energy Optimizer (BEO) Security

In the current version BEO command line requires root access.

Bull Energy Optimizer (BEO) Storage Prerequisites

The storage needed for storing BEO metrics depends directly of the retention policy. With the default retention policy, the storage needed for:

- 10 000 nodes: 120 Gb
- 1 500 nodes: 20 Gb

5.2.1 ENE1 [INHW] Nodes power and energy measurement capabilities

Nodes power and energy measurement capabilities are possible with Bull Energy Optimizer included in our proposal.

5.2.2 ENE2 [INHW] Job energy accounting

Job energy accounting is possible with Bull Energy Optimizer included in our proposal.

5.2.3 ENE3 [INHW] Energy analysis capabilities

Energy analysis capabilities are possible with Bull Energy Optimizer included in our proposal.

5.2.4 ENE4 [INHW] Power- and energy-aware scheduling and resource management

Power- and energy-aware scheduling and resource management are possible with Bull Energy Optimizer included in our proposal.

5.2.5 ENE5 [INHW] Total power consumption per partition

Knowing the total power consumption per partition is possible with Bull Energy Optimizer included in our proposal.

Total average and maximum power consumption of both partitions together will not exceed specified power consumption limitations. We will perform HPL benchmarks for compute and GPU partition separately to measure the maximum power consumption per partition when performing final acceptance tests.

5.2.6 ENE6 [INHW] Power management and capping

The power capping can be applied to a compute node by using the BEO activate/deactivate powercap command, the SLURM user is informed by the *beo feature tag* of this node. The *beo feature* is built from the prefix '**Beo_powercap_**' followed by the **capping_name**.

Basic example:

```
# beo activate powercap Envelop-1 -c pluton10078
Power capping will be activated on: pluton10078
Do you confirm? [y/N] y
Power capping activated
success: pluton10078
```

```
# sinfo -o '%.5a %.10l %.6D %.6t %N %f'
AVAIL TIMELIMIT NODES STATE NODELIST AVAIL_FEATURES
up infinite 2 drain* pluton[10009,10055] (null)
up infinite 1 down* pluton10033 beo-capping-1
up infinite 1 idle* pluton10081 Test_Envelop-1
up infinite 1 comp pluton10057 (null)
up infinite 2 resv pluton[10004-10005] (null)
up infinite 4 alloc pluton[10007-10008,10052,10054] (null)
up infinite 1 idle pluton10029 Budget-1,Envelop-1
up infinite 1 idle pluton10078 Beo_powercap_Envelop-1
up infinite 1 down pluton10053 (null)
```

The upper limit for the power consumption is adjustable at various levels, including rack, cell and node levels. The *beo create powercap* action defines a new power limit that the node (or a set of nodes) or container and its children is not allowed to exceed. The unit of the cap is Watts.

Basic example:

```
# beo create powercap mondayCap_bullx[16-20] --limit 265
Powercap 'mondayCap' created
```

In case of aggregation (sum), the limit value represents the sum of the component's power. Each component will have its power capped to a specific value in order to enforce this global capping.

Basic example:

```
# beo show caprange power bullx[19,20]
| component.caprange.metric | min | max |
=================================================
| bullx19.caprange.power | 185 W | 353.25 W |
| bullx20.caprange.power | 180 W | 300 W |
```

```
# beo show caprange power rack1
| component.caprange.metric | min | max |
=====================================================
| rack1.caprange.power | 0 W | 0 W |
| bullx01.caprange.power | 180 W | 300 W |
| bullx02.caprange.power | 183.5 W | 307 W |
| bullx03.caprange.power | 185 W | 353.25 W |
```

Bull Energy Optimizer also includes an infrastructure-oriented Power Capping feature which lets the Administrator run the cluster or part of the cluster under Power constraint.

Global power limit can be set to any subset of the cluster. Power Capping is done on the compute nodes and consider the power consumption of all the different equipment to equitably share the power budget.

BEO can power cap only components which provides mechanism to do it. It means that compute nodes may be capped but, for example, most of the Ethernet switches or storage array do not provide any capping feature. The accuracy is very high until the global power consumption of the node is based on the components which are under control.

Power Capping tools on AMD Rome

AMD does not provide a RAPL interface such as on Intel CPUs, but it supports at least ACPI Core-States (C-States) and Power-States (P-States):

- C-States provide a way to put low-consumption on unused cores (idle, sleep, deep sleep, etc.);
- P-States provide a way to limit power consumption by setting a combination of voltage/frequency.

P-States for power-capping is the second real alternative after RAPL or RAPL-like solution. To use P-States for efficient power-capping it is required to provide a matrix of ‘state level’ / ‘estimated power consumption’ as it is not as accurate as RAPL methodology. We also have less available levels of settings since P-States are limited to a few (4 normal and 2 for boosting).

For that reasons we are working closely with AMD to see how to enhance power management features to provide RAPL-like solution:

- APML (Advanced Platform Management Link) for which tests are on-going;
- Energy manager, under discussion with AMD engineering (specific solution under definition).

We already know that the BMC will provide an interface for the power capping (it is still possible to get and set a power limit through the BMC).

Power Capping tools on Nvidia Next-gen accelerator

Nvidia provides its own *Management Library* (NVML) to set/get power management values [*nvmlDevice[Set|Get]PowerManagementLimit()*]. This is a low-level control of GPUs, included as part of the driver, header being part of CUDA Toolkit / DCGM.
Our BEO product will integrate Nvidia primitives to manage power capping features as it is done for CPUs. Still, many questions are open regarding policies (global node capping, CPU+GPU capping, GPU-only capping, etc), and regarding the ratio to set between CPU and GPU.

We will undoubtedly use *Data Center GPU Manager* (DCGM) API to ease integration and guarantee standardization. The integration with DCGM (or some lower level calls, NVML, if more relevant) will be part of BEO support by the time of the cluster installation. It provides:

- Dynamic power capping (single or group of GPU / nodes);
- Fixed clocks;
- Synchronous clock boost.

Atos Bull Energy Optimizer (BEO) will implement DCGM APIs to take benefit of:

- Multi-GPU per node support,
- Multi-node support,
- Simplified APIs,

through standardized data structures (ex. *dcgmDevicePowerLimits_v1*).

If for any reason we need to interface NVML it will be done instead of DCGM, especially if DCGM requires a local daemon on each node (which must be confirmed), as it is something, we want to avoid due to system noise generation).

Regarding the settings (whatever the component, CPU, GPU, others) they are defined through an automated process and wrote on local node configuration. If the node is re-installed BEO will re-apply the configuration on this node. Concerning BEO, nearly all the configuration is located on management nodes. No configuration is embedded on compute nodes. Persistent power cap will be implemented in the release available at the time of the installation (Q4 2020).

5.3 Connectivity to other data centres and top-level Ethernet network

5.3.1 Common technical requirements for routers and top-level Ethernet switches

The proposed solution for routers and top-level Ethernet switches is based on CISCO Nexus N3K-C3636C-R (routers) and CISCO Nexus N3K-C3408-S switches (top-level).

The following diagram represents the logical design of the proposed architecture:

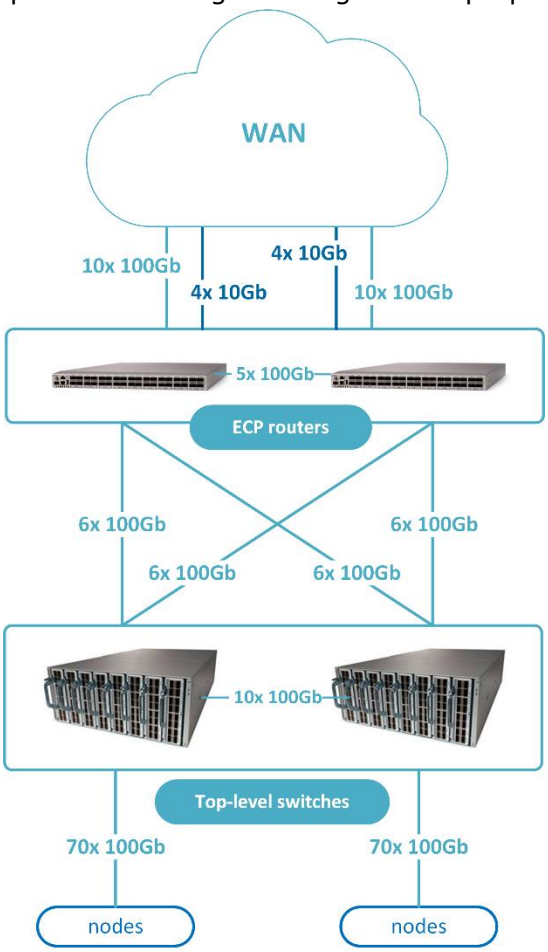


Figure 25: logical design of the proposed architecture for connectivity of VEGA with other data centers

5.3.1.1 RTS1 [INHW] Routers and top-level switches redundant AC power supplies, fans and rack mounting

Both CISCO Nexus N3K-C3636C-R (routers) and CISCO Nexus N3K-C3408-S switches (top-level switches) come with dual-redundant power supplies. Each device will be connected to two independent power sources and are mounting suitable for 19" racks.

5.3.1.2 RTS2 [INHW] Routers and top-level switches operating system and licenses and support for orchestration

Operating system and needed licenses together with updates and upgrades for whole maintenance period of procured system are provided.

Routers

The Cisco Nexus 3636C-R switch provides programmability with support for the Cisco NX-API, Linux containers, XML, and JavaScript Object Notation (JSON) APIs, the OpenStack plug-in, Python, and Puppet and Chef configuration and automation tools.

Top-level switches

CISCO Nexus N3K-C3408-S switches come with Cisco NX-OS software.
Cisco NX-OS is a data center-class operating system built with modularity, resiliency, and serviceability at its foundation. Cisco NX-OS helps ensure continuous availability and sets the standard for mission-critical data center environments. The self-healing and highly modular design of Cisco NX-OS makes zero-impact operations a reality and provides exceptional operational flexibility. Focused on the requirements of the data center, Cisco NX-OS provides a robust and comprehensive feature set that meets the networking requirements of present and future data centers. With an XML interface and a Command-Line Interface (CLI) like that of Cisco IOS® Software, Cisco NX-OS provides state-of-the-art implementations of relevant networking standards as well as a variety of true data-center class Cisco innovations.
The Cisco Nexus 3400-S provides:

- Wire-rate layer 2 and 3 switching on all ports, with up to 25.6 Terabits per second (Tbps) with 7.2 Bpps at ingress and 10 Bpps at egress.
- **Robust programmability, with support for Cisco NX-API, Linux containers, XML, and JavaScript Object Notation (JSON) APIs, Python.**
- High performance and scalability, with a four-core CPU, 32 GB of DRAM, and 70 MB of dynamic buffer allocation, making the switch excellent for massively scalable data centers and big data applications.

5.3.1.3 RTS3 [INHW] Routers and top-level switches support for traffic filtering ACLs

The characteristics of routers and top-level switches for traffic filtering ACLs are:

Routers

Feature	Verified Limit
IPv4 ingress access control entries (ACEs)	RACL-2000, PACL-1024 (without TCAM Carving)
IPv6 ingress access control entries (ACEs)	RACL-1000, PACL-1024 (without TCAM Carving)
ACL Scale	12000 (with TCAM Carving)

Top-level Switches

Feature	Verified Limit
Software Features Level 3	4096 ACL entries
	ACL: Routed ACL with layer 3 and 4 options to match ingress and egress ACLs
	ACL entries, ACLs with L3/L4 options for matching traffic

5.3.1.4 RTS4 [INHW] Routers and top-level switches support for routing protocols and orchestration

Provided devices support Static routes, OSPFv2 and OSPFv3, BGP with BGP load sharing and multipath.

Routers

Routers come with NX-OS operating system with comprehensive, proven innovations including a complete Layer 3 unicast and multicast routing protocol suites are supported, including Border Gateway Protocol (BGP), Open Shortest Path First (OSPF), Enhanced Interior Gateway Routing Protocol (EIGRP), Routing Information Protocol Version 2 (RIPv2), Protocol Independent Multicast Sparse Mode (PIM-SM), Source-Specific Multicast (SSM), and Multicast Source Discovery Protocol (MSDP).

Feature	Verified Limit
IPv4 host routes	750,000
IPv6 host routes	62,000

Top-level Switches

Feature	Verified Limit
Software Features Level 3	Routing protocols: Static, RIPv2, EIGRP, OSPF, and BGP
IPv4 Unicast (hosts) routes	98,304
IPv6 Unicast (hosts) routes	48,128
OSPF/OSPFv3 LSA/LSDB size	250, 000
OSPF/OSPFv3 areas	15
OSPFv2 neighbors	1,000
OSPFv3 neighbors	1,000
OSPF/OSPFv3 LSA/LSDB size	100,000
OSPF/OSPFv3 areas	100

5.3.1.5 RTS5 [INHW] Routers and top-level switches support for VRF

Both routers and top-level switches have support for VRF. The number of VRFs is:

Routers

Feature	Verified Limit
VRFs	3,967

Top-level Switches

Feature	Verified Limit
VRFs	256

5.3.1.6 RTS6 [INHW] Routers and top-level switches additional support

Routers

Feature	Comment
Configurable SPAN or ERSPAN sessions	32
Active SPAN or ERSPAN sessions	32
Active localized SPAN or ERSPAN session per line card	32 sessions across ports on single line card
Active localized SPAN or ERSPAN session (Rx and Tx, Rx, or Tx)	32 sessions, 128 sources and 1 destination
Destination interfaces per SPAN session	1

Source VLANs per SPAN or ERSPAN	6
SSH support	You can use the Secure Shell (SSH) server to enable an SSH client to make a secure, encrypted connection to a Cisco NX-OSdevice
RADIUS support	You can specify the following authentication methods for the AAA services: RADIUS server groups—Uses the global pool of RADIUS servers for authentication
	IEEE 802.1ab: Link Layer Discovery Protocol (LLDP)
	RFC 1305: Network Time Protocol (NTP) Version 3
	RFC 2338: VRRP
Jumbo Frames	Configurable Maximum Transmission Unit (MTU) of up to 9216 bytes (jumbo frames)
SNMP	NX-OS also supports Simple Network Management Protocol (SNMP) Versions 1, 2, and 3 MIBs
Ethernet link aggregation with LACP	IEEE 802.3ad: Link Aggregation Control Protocol (LACP)
Multi Chassis Link Aggregation	Virtual Port Channel (VPC) technology provides Layer 2 multipathing by eliminating the Spanning Tree Protocol. It also enables fully used bisectional bandwidth and simplified Layer 2 logical topologies without the need to change the existing management and deployment models
ECMP for IPv4 and IPv6	Equal cost multipaths (ECMPs): 8
QoS	You can classify traffic based on the Differentiated Services Code Point (DSCP) value in the DiffServ field of the IP header (either IPv4 or IPv6)

Top-level Switches

Feature	Comment
SPAN/ERSPAN	SPAN on physical, PortChannel, and VLAN
	ERSPAN Versions 2 and 3
Management Options	Switch management using 10/100/1000-Mbps management or console ports
	SSHv2

	RADIUS
	IEEE 802.1ab: LLDP
	VRRPv3
Jumbo Frames	Jumbo frame support (up to 9216 bytes)
SNMP	SNMP v1, v2, and v3
Ethernet link aggregation with LACP	LACP: IEEE 802.3ad, IEEE 802.1ax
Multi Chassis Link Aggregation	vPC
ECMP for IPv4 and IPv6	64-way equal-cost multipath (ECMP)
Qos	You can classify traffic based on the DSCP value in the DiffServ field of the IP header

5.3.2 Routers Technical Specifications

5.3.2.1 CPE1 [INHW] Number, type, height and location of routers

Our solution for routers is based on 2x CISCO Nexus N3K-C3636C-R model. The size of the proposed router model is 1 rack unit. The routers will be installed in IZUM's communication system room.

5.3.2.2 CPE2 [INHW] Ports per each router and management port

The characteristics of the proposed router model are summarized here:

Figure 26: Nexus N3K-C3636C-R switch

36x QSFP28 ports operating at 40 or 100 Gigabit Ethernet:

- 8 ports that can be used for MACSec
- Locator LED
- Environment LED
- Status LED
- Dual redundant power supplies
- 3 redundant fans
- One 10-, 100-, or 1000-Mbps management port (copper or fiber)
- One RS-232 serial console port
- One USB port

In addition, we also include in our proposition:

- 8x 10G SFPs with 5m optical cables included to connect to WAN
- 20x 100G SFPs to connect to WAN (no cables included)
- 5x 100G twinax optical cables of 5m length to interconnecting both routers
- 24x 100G twinax optical cables of 30m length to connect both switches

5.3.2.3 CPE3 [INHW] Routers support for big buffers

The total buffer size of the proposed switch model for routers is 16GB for 36 ports. Thus, each 100GbE port has at least 400MB buffer.

5.3.3 Top-level Ethernet switches technical specifications

5.3.3.1 TOP1 [INHW] Number, type and location of top-level switches

Our solution for top-level Ethernet switches is based on 2x CISCO Nexus N3K-C3408-S switches. The Cisco Nexus 3408-S is a 4x Rack-Unit (RU), 8-slot chassis with the flexibility of 100GbE or 400GbE Line-Card Expansion Modules (LEMs) offering 128x ports of 100GbE or 32x ports of 400GbE. The top-level switches will be installed in RIVR2 room.

5.3.3.2 TOP2 [INHW] Ports per each top-level switch

The characteristics of the proposed router model are summarized here:

Figure 27: Nexus N3K-C3408-S switch

The Cisco Nexus 3408-S has the following hardware configuration:

- 4RU, 8-slot chassis
- NXM-X16C LEM with 16 ports of Quad-Small-Form-Factor 28 (QSFP28)
- NXM-X4D LEM with 4 ports of Quad Small Form-Factor Pluggable – Double Density (QSFP-DD)
- Beacon LED
- Status LED

- Dual-redundant power supplies
- Redundant (2+1) fans
- Two 100/1000-Mbps SFP ports
- One RS-232 console port
- One RJ45 and one SFP Management port
- One USB port

Our proposal includes 2x Cisco Nexus 3408-S switches as top-level switches configured and cabled to that:

- Each top-level switch will be connected to each router using 6x 100GbE links (twinax cables)
- Each node in virtualization & service partition and login partition will be connected to one of the two switches. Those nodes will be dispatched over the two top-level switches and we will use twinax optical cables to make these connections.
- Each top-level switch will be direct connected to the other switch using 10x 100GbE twinax cables.

Each top-level switch has 1x RJ45 and one SFP management port and will be connected to the management network as described in section 6.8.

5.3.3.3 TOP3 [INHW] Top-level switch support for big buffers and RoCE

Cisco Nexus 3408-S supports RDMA. It is our understanding that support for big buffers requirement was dropped for top-level switches.

5.4 Maintenance and support of HPC infrastructure requirements

5.4.1 MAI1 [SW] Reboot time

The power on and reboot time of the whole system to a "ready for production" state is minimized but cannot be guaranteed to be under 30 minutes considering the storage components (DDN Lustre and Ceph).

For a complete power on of the solution, a sequence must be respected to ensure that the system is working properly for production. For example, the network infrastructure must be ready first, then the management servers, following by the storage components that will have to be mounted on the compute resources that will come at the end.

Considering each separated element, the target of 30 minutes seems very low to ensure that the solution will be ready for production.

The network will boot in about 1 to 2 minutes, then the management servers will take about 10 minutes to be ready depending on what will be installed on-top of the default installation. This duration considers that all storage and service on which the management relies are ready to serve. The storage systems can be turned on in parallel to the management components, anyway for storage we must consider potential system checking that will increase the time before availability. Even if we consider a clean filesystem, we must consider that the 'boot time' will be around 10 minutes at best, and certainly around 15-20 minutes.

Then the boot time per node mainly depends on the initial checking (especially memory) that can take longer for high-memory solution. We assume that the power-on cycle (i.e. not for the reboot) can be higher than 5 minutes to reach the boot loader, this value can be much higher for very high memory system (>1TB). Then the OS will have to boot (including potential file system check for diskfull servers). A normal boot time will take about 2-3 minutes to ensure network access (SSH) but again depends on which services are running at boot time, if the mount of remote file systems is doing well, etc. It is also important to notice that for diskless nodes it is not possible to boot all nodes in parallel (not only due to diskless as also some power constraints - power call). So, for booting nodes a sequence must also be put in place.

So, with the current configuration we assume that the boot time of the whole solution will not be possible in 30 minutes, Atos will do its best to enhance the total boot time without any guarantee to reach this requirement.

5.4.2 MAI2[SW] Health check of components

The proposed solution includes Node Health Checker (NHC) which is a well-known tool to check system health between job execution. This component is linked with the Slurm batch system to avoid the use of damaged (or unstable) components. NHC will be provided with a default configuration that can be adapted to on-site requirements for additional (or less) testing coverage.

5.5 Data centre integration requirements

5.5.1 DAT1 [SW] Reinstallation duration

A re-installation of the procured system software from scratch can be done in 5 working days considering that all the infrastructure and hardware is in stable condition.

5.5.2 DAT2 [SW] Reinstallation with software alternatives

Re-installation from scratch with an alternative software distribution provided by the hosting entity is a possibility as the equipment is based on standard component protocols and interface.

However, this requires in-depth understanding of all the compatibility element of the software and hardware component and will potentially impact the maintenance and support of the system.

A re-installation from scratch in such conditions should take 6 days and it is not provisioned in this proposal.

5.6 Data management requirements

5.6.1 DMG1 [SW] File system performance impact

Considering system noise that can be generated by remote I/O operations, we are considering that the impact of the Lustre file system is null when no operations are performed (stand-by mode).

With the current configuration, and if there are no I/O activities the impact to the workload will be imperceptible. This is true as Lustre and embedded Linux NFS are both running at kernel level: drivers for Lustre client, kernel daemons for NFS.

5.7 Programming environment and productivity requirements

5.7.1 PRO1 [SW] Compiler set

The proposed solution includes the following programming environment to fully cover the requirements on both CPU and GPU architectures:

- Intel Parallel Studio XE Cluster Edition

Please find below the proposed sizing of the associated licenses:

Product	License tokens
Intel Parallel Studio XE Cluster Edition Linux	10 Floating Seats

Intel Parallel Studio XE 2020

Intel Parallel Studio XE provides the tools to build C, C++ and Fortran HPC, enterprise, AI and cloud solutions, but also to maximize the software performance on compute platforms. Depending on the edition, the package includes compilers, numerical libraries, performance profilers and code analysers.

Intel Parallel Studio XE has three editions: Composer Edition, Professional Edition, and Cluster Edition. We propose here the Cluster Edition.

Product Contents

The following table shows which Intel Software Development Tools are present in each edition of Intel Parallel Studio XE 2020.

Tool	Composer Edition[1]	Professional Edition	Cluster Edition
Intel® C++ Compiler	X	X	X
Intel® Fortran Compiler	X	X	X
Intel® Distribution for Python*	X	X	X
Intel® Integrated Performance Primitives (Intel® IPP)	X	X	X
Intel® Math Kernel Library (Intel® MKL)	X	X	X
Intel® Data Analytics Acceleration Library (Intel® DAAL)[2]	X	X	X
Intel® Threading Building Blocks (Intel® TBB)	X	X	X
Intel-provided Debug Solutions	X	X	X
Intel® Advisor		X	X
Intel® Inspector		X	X
Intel® VTune™ Profiler		X	X
Intel® Cluster Checker **(For Linux* OS only)**			X
Intel® MPI Benchmarks			X
Intel® MPI Library			X
Intel® Trace Analyzer and Collector			X

For more information regarding this product, please refer to the following link: https://software.intel.com/en-us/parallel-studio-xe

The table below lists the product tools:

Tool	Version
Intel® Advisor	2020
Intel® C++ Compiler	19.1
Intel® Cluster Checker **(For Linux* OS only)**	2019 Update 6
Intel® Data Analytics Acceleration Library (Intel® DAAL)	2020
Intel® Distribution for Python*	2020
Intel® Fortran Compiler	19.1
Intel® Inspector	2020
Intel® Integrated Performance Primitives (Intel® IPP)	2020
Intel® Math Kernel Library (Intel® MKL)	2020
Intel® MPI Benchmarks	2019 Update 5
Intel® MPI Library	2019 Update 6
Intel® Threading Building Blocks (Intel® TBB)	2020
Intel® Trace Analyzer and Collector	2020
Intel® VTune™ Profiler	2020
Intel-provided Debug Solutions	

5.7.2 PRO2 [SW] Support for multiple executables

Bull OpenMPI 4.0 is supporting MPI_COMM_SPAWN_MULTIPLE, allowing the execution of multiple binaries with a single line of command. Process affinity settings could be specified per binaries and they will share the same MPI_COMM_WORLD.

5.7.3 PRO3 [SW] Optimized numerical libraries

The Intel compiler provides the Math Kernel Library (MKL) which can be used to replace the BLAS, LAPACK and ScaLAPACK libraries with optimized versions. It also includes a 2D/3D FFT library.
AMD provides BLIS, libM, libFLAME and FFTW.

5.7.4 PRO4 [SW] Development tools

Our solution includes Rogue Wave Totalview as parallel debugger for the VEGA system. License support is provided for 256 MPI ranks.

TotalView is a software debugger for complex C, C++, and Fortran applications that have hundreds to thousands of parallel processes running. Intuitively diagnose and understand your complex code, resulting in faster resolution of bugs, memory issues, and crashes at execution.

- **Reverse debugging**: Work backwards from failure and eliminate the need to repeatedly restart the application.
- **Multi-language applications debugging**: Easily analyze and debug apps written in both Python and C/C++.
- **Simultaneous debugging**: Get complete control over program execution within a single thread or within groups of processes or threads.
- **Pinpointing and fixing bugs**: Troubleshoot difficult problems that occur in concurrent programs that take advantage of threads such as OpenMP, MPI, GPUs, or coprocessors.

Mixed Language Debugging

Many modern HPC applications are leveraging the advantages of Python and C/C++ at the same time to build complex applications. However, debugging mixed language applications is not easy. TotalView allows developers to easily understand the flow of execution across language barriers and examine the data within both languages.

- Easily set up a Python debugging session.
- View an integrated Python and C/C++ stack trace.
- Examine and compare variables in Python and C/C++.
- Utilize TotalView's reverse debugging and memory debugging technologies.

Record & Replay Execution History

Reverse debugging records the execution history of a program and makes that history available for diagnosis. TotalView's ReplayEngine makes it possible to work back from a failure, error, or crash to find the origin without repetitive restarts and stops. It allows storage of the recording to investigate the error at any time. ReplayEngine reduces the amount of time invested in troubleshooting your code.

- Freedom to explore application execution either backwards or forwards.
- Step back through execution history and review all variables.
- Go back and look at functions and variables in the context of a crash.
- Easily follow the logic of unfamiliar routines.
- Set a watchpoint and run back to find the source of unexpected data.
- Enable recording during a debug session without detaching.

Multiprocess Program

Troubleshooting Building a multithreaded application or transitioning from a serial to a parallel application presents significant challenges. TotalView is a source code debugger for troubleshooting complex, multiprocess programs.

- Operate with equal ease on single thread/process or with groups of threads/processes.
- Easily establish interactive debugging sessions in clusters with Reverse Connect.
- Set breakpoints with thread or process width to synchronize or use barrier constructs.
- Control the execution of threads or processes individually or in groups.
- View program data and threads/processes with parallel backtrace.
- Troubleshoot deadlocks and race conditions.
- Work with automatically defined lockstep or custom groups.

Detect & Analyze Memory Errors

TotalView includes MemoryScape, a dynamic memory analysis tool that reduces time spent on memory debugging. The powerful memory error and analysis tool has a low performance overhead with an interface that allows for identification of heap memory within a program.

- No need to recompile applications.
- Detect leaks and errors in vendor libraries.
- Track allocated, deallocated, and leaked memory blocks.
- Detect memory leaks and corrupted memory early.
- Flag memory leaks and events before they crash your application.
- Analyze memory usage patterns.
- Support for multiprocess and hybrid applications in clusters.

5.7.5 PRO5 [SW] Licensing

Intel Parallel Studio XE, included in our proposition, includes licensing for 5 simultaneous users throughout the lifetime of the system. Validity extension is possible.

5.7.6 PRO6[SW] PAPI support

PAPI is part of the standard tools proposed for profiling purpose and will be installed on all relevant type of nodes.

5.8 System and application monitoring

5.8.1 MON1 [SW] Performance profiling

Atos lightweight profiler is included in our proposal. **Lightweight Profiler**, formerly MPI Prof, is a lightweight and modular tool for monitoring applications running on supercomputers. It is designed to collect different kinds of useful data like CPU metrics, memory utilization, energy consumption, MPI communication events, etc.
It is used as a wrapper for an application; the user invokes LWP explicitly via the command line. LWP is modular by nature, so each type of data is collected by a specific component.

LWP supports MPI (Bull OpenMPI and Intel MPI) and also OpenMP pragmas (including MPI+X). This tool provides statistics for all MPI calls with support of MPI-1, MPI-2 and MPI-3 standard.

Multi-levels:

- Global statistics
- Per process statistics with communication matrix
- Communications schemas
- Per MPI function time consumption
- MPI functions can be filtered
- Very low to low overhead
- Medium dataset (depends on the number of processes)
- Dual-matrix plots (transfer size, transfer time)
- Communication patterns
- Only 1to1, 1toA, Ato1 are printed
- Very low to low overhead
- Medium dataset (depends on the number of processes)

Figure 28: Example of per process histogram generated by Lightweight Profiler

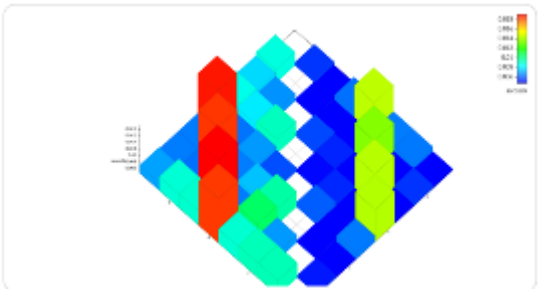


Figure 29: Example of dual matric plot generated by Lightweight Profiler

MPI Calls timeline

(Each node is an MPI communication and indicate function, from, to, message size, min, max, average time taken per this communication.)

- Formalize loops on set of MPI calls
- Higher overhead
- Very large data set (depend on # of processes to track)

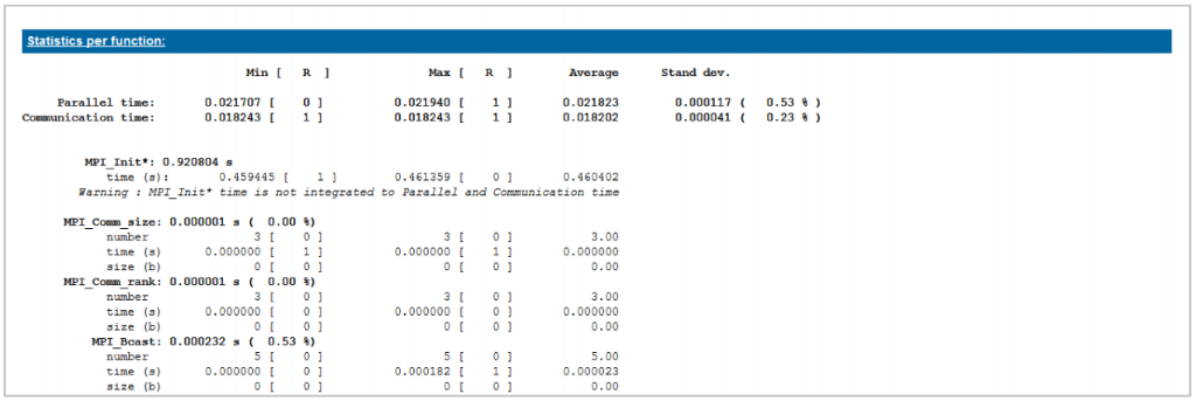
MPI functions statistics

A call path is characterized by a function name and line of source code where the function is called (the source code lines is not shown to the user).
MPI functions in the call path, having the same call parameters (e.g., function parameters such as: data type, count, destination, etc.) will be aggregated together. This feature is especially useful for loops.
MPI functions within the same application function but with different input parameters (e.g. destination rank) are displayed in the same call path section but are not aggregated together.
Computes the maximum, minimum and average time spent in the aggregated MPI calls.

HTML report

With Lightweight profiler, HTML report can easily be generated. It contains:

- Summary of application: name, number of MPI tasks, parallel time and communication time, etc.
- Statistics per function name.



Statistics per function:

	Min [R]	Max [R]	Average	Stand dev.
Parallel time:	0.021707 [0]	0.021940 [1]	0.021823	0.000117 (0.53 %)
Communication time:	0.018243 [1]	0.018243 [1]	0.018202	0.000041 (0.23 %)

MPI_Init*: 0.920804 s

time (s):	0.459445 [1]	0.461359 [0]	0.460402

Warning : MPI_Init time is not integrated to Parallel and Communication time*

MPI_Comm_size: 0.000001 s (0.00 %)

number	3 [0]	3 [0]	3.00
time (s)	0.000000 [1]	0.000000 [1]	0.000000
size (b)	0 [0]	0 [0]	0.00

MPI_Comm_rank: 0.000001 s (0.00 %)

number	3 [0]	3 [0]	3.00
time (s)	0.000000 [0]	0.000000 [0]	0.000000
size (b)	0 [0]	0 [0]	0.00

MPI_Bcast: 0.000232 s (0.53 %)

number	5 [0]	5 [0]	5.00
time (s)	0.000000 [0]	0.000182 [1]	0.000023
size (b)	0 [0]	0 [0]	0.00

Figure 30: MPI component HTML report - statistics per function name

Other relevant diagrams such as:

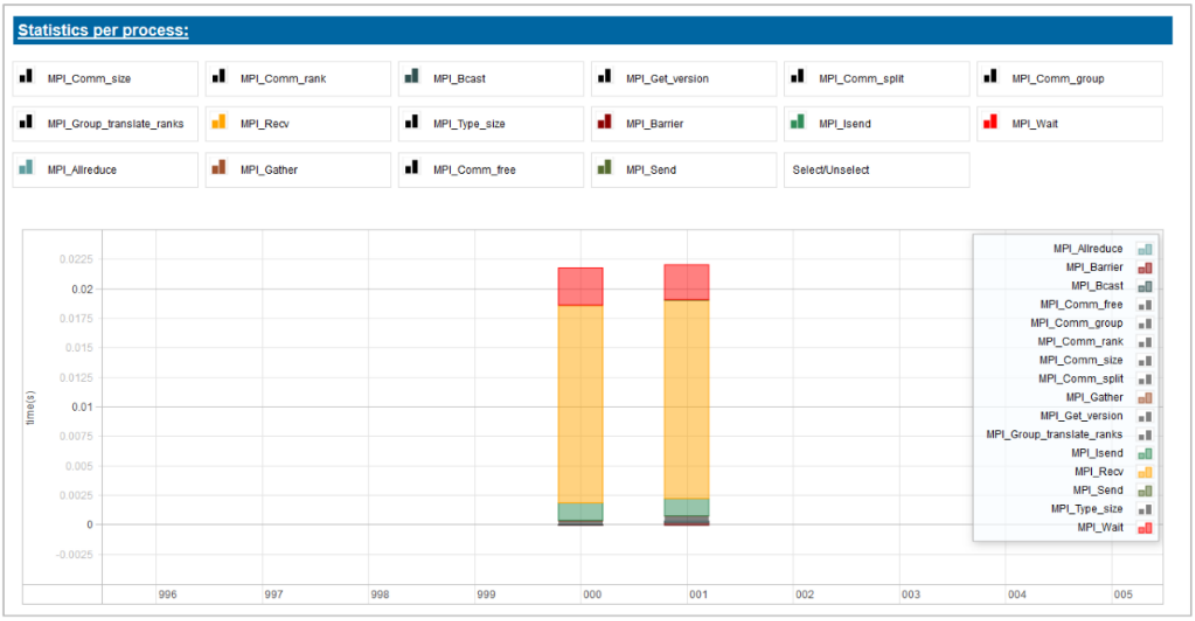


Figure 31: Statistics per MPI task

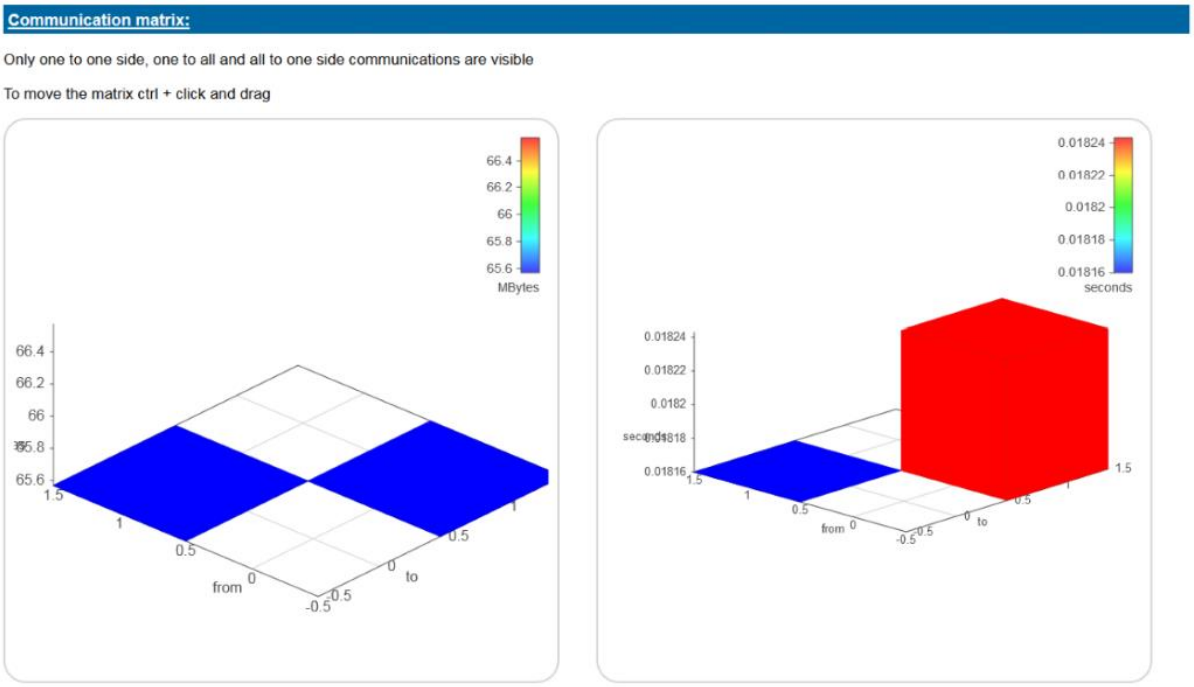


Figure 32: Communication matrix

Figure 33: Communication schema

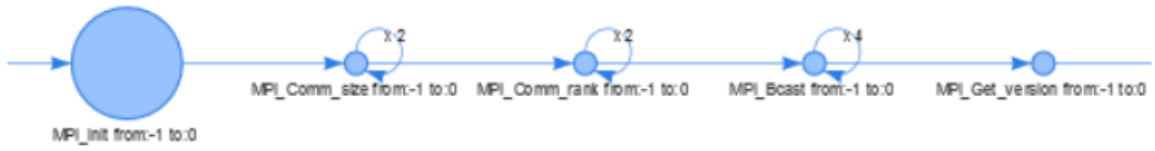


Figure 34: details of communications

Detailed ASCII report

Communication and computation overlap over non-blocking MPI point to point primitives.

For readability, the report is organized by the following MPI function categories: P2P, P2P requests, and generic.
The output report can be filtered by MPI rank and MPI function.

```
* Communication Details:

Call site:
\_ 0x400cfc
 \_ __libc_start_main
  \_ main
-----------------------------------------------------------------------------------------------------------------------------------
Type: [GENERIC]                         Elapsed time inside MPI function
ID:     Name            Hits:   Rank:   Min:        Max:        Sum:
0       MPI_Init        1       0       0.098879    0.098879    0.098879
1       MPI_Comm_rank   1       0       0.000000    0.000000    0.000000
2       MPI_Comm_size   1       0       0.000001    0.000001    0.000001
10      MPI_Init        1       1       0.095483    0.095483    0.095483
11      MPI_Comm_rank   1       1       0.000000    0.000000    0.000000
12      MPI_Comm_size   1       1       0.000001    0.000001    0.000001
Type: [P2P]                                                                 Elapsed time inside MPI function
ID:     Name            Hits:   Rank:   Count:  Type:       Dest:   Tag:    Comm:           Min:        Max:        Sum:
4       MPI_Isend       1       0       4096    MPI_INT     1       101     MPI_COMM_WORLD  0.000041    0.000041    0.000041
5       MPI_Isend       1       0       4096    MPI_INT     1       102     MPI_COMM_WORLD  0.000005    0.000005    0.000005
6       MPI_Recv        1       0       4096    MPI_INT     1       202     MPI_COMM_WORLD  0.000687    0.000687    0.000687
7       MPI_Recv        1       0       4096    MPI_INT     1       201     MPI_COMM_WORLD  0.000021    0.000021    0.000021
14      MPI_Recv        1       1       4096    MPI_INT     0       102     MPI_COMM_WORLD  15.000587   15.000587   15.000587
15      MPI_Recv        1       1       4096    MPI_INT     0       101     MPI_COMM_WORLD  0.000023    0.000023    0.000023
16      MPI_Isend       1       1       4096    MPI_INT     0       201     MPI_COMM_WORLD  0.000049    0.000049    0.000049
17      MPI_Isend       1       1       4096    MPI_INT     0       202     MPI_COMM_WORLD  0.000005    0.000005    0.000005
Type: [P2P REQUEST]                             Elapsed time inside MPI function    Overlapping communication/computation
time
ID:     Name            Hits:   Rank:   Match:  Min:        Max:        Sum:        Min:        Max:        Sum:
8       MPI_Wait        1       0       5       0.000007    0.000007    0.000007    15.001071   15.001071   15.001071
9       MPI_Wait        1       0       4       0.000001    0.000001    0.000001    15.001155   15.001155   15.001155
18      MPI_Wait        1       1       17      0.000112    0.000112    0.000112    0.000084    0.000084    0.000084
19      MPI_Wait        1       1       16      0.000017    0.000017    0.000017    0.000282    0.000282    0.000282

Call site:
\_ 0x400cfc
 \_ __libc_start_main
  \_ main
   \_ another_function
-----------------------------------------------------------------------------------------------------------------------------------
Type: [P2P]                                                                 Elapsed time inside MPI
function
ID:     Name            Hits:   Rank:   Count:  Type:       Dest:   Tag:    Comm:           Min:        Max:        Sum:
3       MPI_Send        1       0       1       MPI_INT     1       301     MPI_COMM_WORLD  0.000090    0.000090
0.000090
13      MPI_Recv        1       1       1       MPI_INT     0       301     MPI_COMM_WORLD  0.000143    0.000143
0.000143

===================================================================================================================================
                                                                [[ end of report ]]
===================================================================================================================================
```

Figure 35: Communication details generated by MPI component for an example MPI application

Information refinement

Full call path information will use GNU backtrace function.
User may opt to aggregate only by MPI function or parent function (caller); this disables GNU backtrace function and improves performance.

Support of optional full tracer capability

When activated, the full tracer will report every single MPI function call issued during the program execution without aggregating multiple occurrences of the same MPI event.

As a tradeoff, performance overhead and memory footprint are significantly increased.

5.9 Security requirements

As a European company Atos is fully committed in applying the European Regulations, and when applicable the related national laws.

For Atos, the protection of Personal Data is a topic of the utmost importance. The Processing of Data, including Personal Data, is part of its core activities, whether it does so as a Data Controller or as a Data Processor. Accordingly, compliance with Data protection laws and regulations is one of Atos' main priorities.
As a fundamental element of the first pillar, Atos has adopted a Group Data Protection Policy ("Group DP Policy"). Atos considers that the implementation of such a Group DP Policy raises awareness within the Group and participates to the demonstration of Atos' compliance with its legal obligations.
Atos applies the international standard ISO27001:2013 for the implementation of his Information Security Management System (ISMS).

Atos will ensure the availability of the latest software components. Emergency patches will be rolled out in collaboration with Hosting Entity's administration.

5.9.1 SEC1 [SW] Job resources usage isolation

Regarding user isolation, the solution provides a strong isolation between the management components and the final users environment with only as limited as possible services exposed to compute resources. In addition, SELinux is enabled on all management servers to reduce the scope of a security threat. Also, most infrastructure and all user related services are running as micro-services (containers) to also reduce the scope of a security issue to the service itself. SELinux is also available on compute resource but not enable by default due to performance impact on application.

5.9.2 SEC2[SW] Data security

Concerning security breaches and vulnerabilities, Atos has implemented a security patching mirror to minimize the reaction time required to install the security update, and Atos teams are working closely with hardware and software third parties to guarantee the best possible reactivity.

For data stored on disks, Atos will elect:

- Either to apply the "Non-Disk Return option" as part of the Maintenance and Support to ensure that there is no danger of any data being taken off site (In this case, the defective disks will remain on-site and the property of IZUM), or

- to delete all IZUM data before disks are taken off-site. In case a reliable data deletion is not possible (by means of a proper deletion software or deletion process), the device will be rendered unusable (physically destroyed)."

5.9.3 SEC3 [SW] Security patches and updates availability

Atos will provide security patches or workaround for all supported software including middleware and firmware. Regarding the delay, Atos has dependencies with various providers and aims to deliver security fixes which allow to keep the system running with the expected performance. It is why Atos is doing a validation for all fix and provide a fully integrated solution including low-level components (firmware, drivers, kernel, etc.) up to higher components such as libraries and frameworks. This statement is particularly true for all kernel-related security fixes which impact drivers such as Mellanox InfiniBand or Nvidia but also storage components such as Lustre.

Regarding the Operating System, it is also why we have made the choice to work closely with Red Hat and selected Red Hat Enterprise Linux which provides up-to-date security fixes in advance compared to the community. This is also true for a critical component regarding security, the container runtime, the security aspect has been a key element in the partnership with Sylabs.io regarding Singularity Pro runtime which is part of this proposal.

5.9.4 SEC4 [SW] Authentication and authorization

Considering authentication and authorization the SMC solution is based on a centralized directory server (389) included in Red Hat Enterprise Linux 8.

This solution will provide a service to declare and manage supercomputer users and groups information. The LDAP schema can be set to use eduGAIN user attributes. The authentication process can be achieved using standard 'password' mechanism or can be set to some other mechanism such as Kerberos or X509 certificates.

5.9.5 SEC5 [SW] Anomaly detection (aka SIEM)

Atos Big Data & Security has a dedicated Cybersecurity division. However, in the current proposition we do not propose any SIEM solution due to cost considerations.

What is included in our proposal is a monitoring environment that can be enhanced to better cover security patterns.

Also, as the solution is based on Red Hat Enterprise Linux 8, so it is strongly advised to set up and use OpenSCAP through Ansible playbooks. This way it will be possible to scan for security compliancy in an automated way.

6 Technical specifications of supercomputer

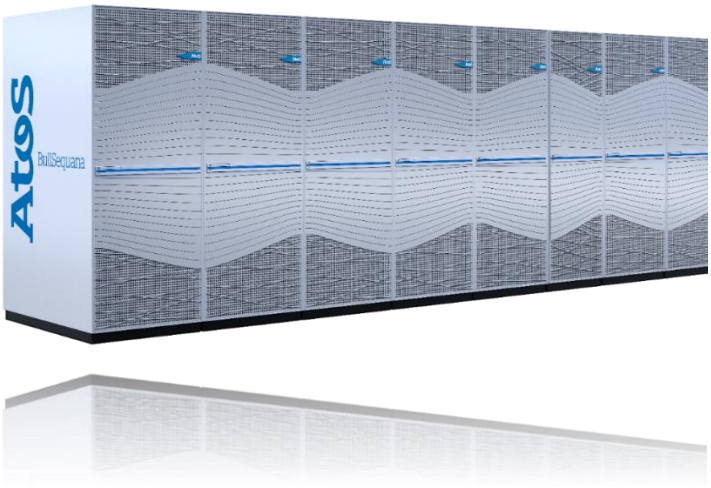
6.1 System architecture of supercomputer

The proposed solution to host the Vega CPU and GPU compute partitions has been built around our in Europe developed direct liquid cooled BullSequana XH2000 system.

The BullSequana XH2000 is an evolution of the BullSequana X1000 supercomputer which was introduced in 2016 and is in production use at many supercomputing sites.

The BullSequana XH2000 is compatible with existing and future compute blades and uses BullSequana X1000 components wherever possible to protect existing investments.

The BullSequana XH2000 platform is designed and based around European technology and development.



The BullSequana XH2000 preserves all the best features from the previous X1000 platform. It offers and enhances them in multiple directions:

- **Open and modular platform**. For IZUM this means that VEGA can be easily expanded and reconfigured, based on future demands and workloads;
- **Integration of current and future technologies**, preserving the design of the compute blades. For IZUM this means that there will be an investment protection in the case IZUM decides to move on to future technologies and new DLC platforms from Atos;

- **Ultra-energy efficient based on enhanced Direct Liquid Cooled (DLC) techniques** with up to 40°C for inlet water and 95% DLC of the complete system. For IZUM this means that with the BullSequana XH2000 VEGA will be a system with the highest possible energy efficiency available in the market;

- **Flexibility in the embedded interconnect**. For IZUM this means there is a choice in variety of cluster interconnects like InfiniBand, Ethernet or Bull eXascale Interconnect (BXI);

- **All-in-one compute rack design**. This means for IZUM an easy extension per rack and a highly scalable solution;

- **Smaller building blocks to address any size of requirement**. With these smaller building blocks IZUM can extend the environment by small steps of even large upgrades with high flexibility on overall capacity;

- **Larger systems with DLC**: With the BullSequana XH2000 solution, IZUM can easily expand the system from 96, up to 64.000 nodes;

- **New interconnect topologies and flexible pruning ratios** to better adapt end-user's needs;

- **New interconnect**: InfiniBand HDR100/200 (and existing Ethernet and BXI). This to make sure IZUM will profit from the latest technologies integrated in a DLC system;

- **Optimized cost options**: The Bull Exascale supercomputers have an open architecture and are based on industry standards (hardware and software), mainly developed in Europe. They offer users a large choice of technology options and have been designed to be compatible with successive generations of future processor technologies (CPUs, accelerators, low-power processors) and different interconnect technologies (BXI, Ethernet, InfiniBand), thus offering maximum investment protection.

Controlling energy consumption is the main roadblock on the path to EXAScale. The BullSequana XH2000, as a part of our EXAScale platform is ultra-energy efficient targeting PUE below 1.1.

To achieve this, all the components of BullSequana – including compute nodes (processor, memory and NICs), switches and power supplies – are cooled using an enhanced version of the European developed Bull Direct Liquid Cooling (DLC) solution. DLC, now in its fourth generation, is a proven cooling technology that minimises the global energy consumption of a system by using water with inlet temperatures up to 40°C with 95% efficiency. This technology, coupled with appropriate data centre design, propels BullSequana XH2000 into the lead with respect to energy use for a supercomputer of a given capability.

Cabinet design

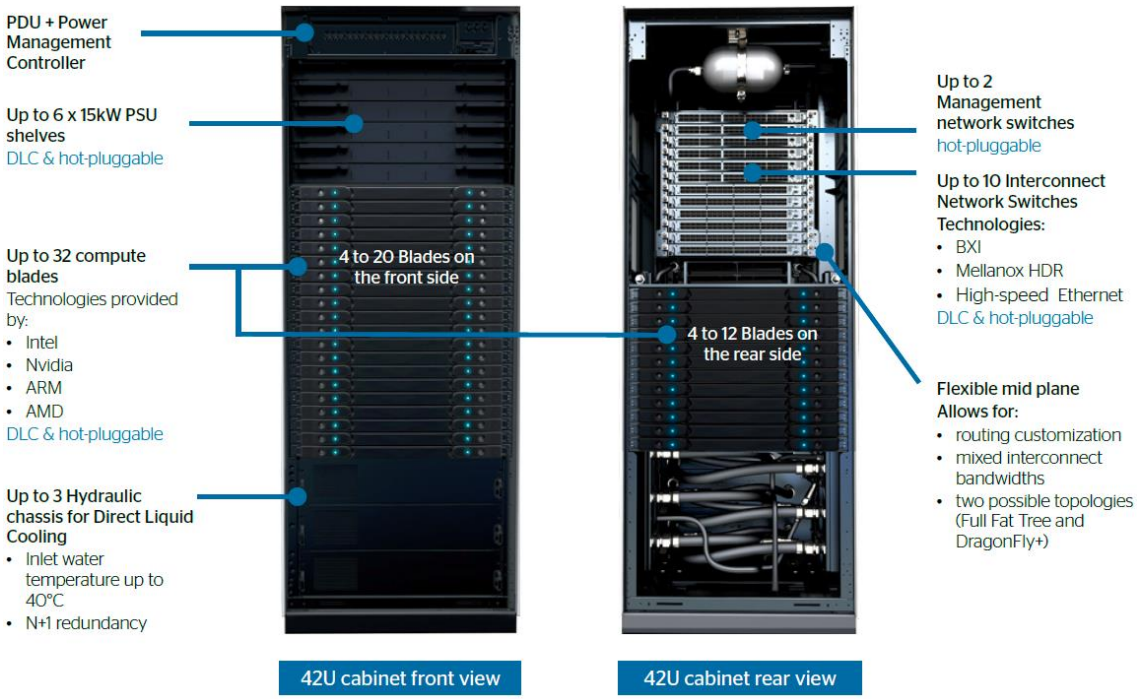


Figure 36. The BullSequana XH2000 flexible packaging.

The BullSequana XH2000 is designed in one 42U cabinet with:

- Up to 32 compute DLC blades/96 compute nodes
- 20 on front side, 12 on rear side
- Up to 6 liquid-cooled PSU shelves
- 2 HYCs, optional 3rd HYC for 2+1 redundancy
- 2 Leaf Ethernet modules
- Up to 10 Interconnect DLC Switches
- HDR100 & HDR200
- BXI and Fast Ethernet
- 1 Power distribution unit with 3x 63A tri-phase cables
- Power and signal connections at the top of rack
- Power and cooling capacity: 15 to 90kW

Compute node design

The BullSequana XH2000 is a blade-based system leveraging on the following technologies (note that support of technologies is subject to change, based on partners technologies availability):

- per socket and Apache pass support (when available)
- Intel Xeon Ice Lake / Cooper Lake with 3 x 2 socket nodes per blade and 8 memory channels per socket

- Cavium ThunderX2 (ARMV8) with 3 x 2 socket nodes per blade and 8 memory channels/socket
- AMD Epyc Rome / Milan with 3 x 2 socket nodes per blade and 8 memory channels per socket
- Intel Xeon Cascade Lake + 4 x Nvidia Volta V100 GPU's
- AMD Epyc Rome / Milan + 4 x Nvidia Volta-Next GPU's
- Bull eXascale Interconnect (BXI)
- InfiniBand HDR, HDR100.

The flexible nature of the BullSequana XH2000 architecture means that it is possible to mix blades within a system, allowing upgrades to meet customer's future requirements or the incorporation of future processor architectures in a straightforward manner.

Cell concept

With BullSequana XH2000, the computer resources are delivered in independent and integrated cells. Each cell tightly integrates:

- Compute nodes
- Interconnect switches
- Redundant power supply units
- Redundant liquid cooling heat exchangers.

A BullSequana XH2000 rack can be populated with up to ten (10) 200Gbps HDR 40-ports switches. Those switches can be used to build may kind of topology including Fat tree, Generalized Hypercube, All to All or DragonFly+ (DF+), a combination of Fat tree islands and All-to-all between the islands.
For Fat tree, those switches can be used as L3, L2 or L1 and we also can take advantage of the HDR100 setting that permit to have up 80 100Gbps ports per switch using Y-cable.

Proposal for Vega Supercomputer

We have chosen to provide IZUM an HDR InfiniBand Fabric in a Dragonfly+ topology (DF+).

The new Fully Progressive Adaptive Routing (FPAR) is implemented in the DF+ topology and fully supported by Mellanox. It provides a very efficient routing for which decisions are evaluated in every router. In addition, Adaptive Routing Notification (ARN) is implemented where messages are sent between routers to notify of distant congestion that can be resolved before it becomes an issue.

It is also scalable to a very large number of nodes and provide high bisection bandwidth. It is based on the concept of dividing nodes to groups while the groups are connected in an all-to-all way. A group is based on a multi-layer Fat tree. The nodes are connected to the so-called leaf switches which are then connected to a second layer and

a possible third layer, and so on. The top layer is called the spine switches which are used for the all-to-all connections to form a Dragonfly+ topology.

All compute nodes of the VEGA CPU and GPU partitions will be hosted in BullSequana XH2000 racks, while the service nodes are hosted in air-cooled racks. The components of the other partitions are also hosted in air-cooled racks.

The proposed VEGA GPU partition is based on compute DLC blades equipped with four NVIDIA GPUs and two AMD x86-64 CPUs.

The proposed VEGA CPU partition is made of dual-socket DLC compute nodes, AMD x86-64 based, homogeneous with the exception that some nodes differ in their memory configuration (following IZUM requirements further listed in this document).

Please find in the rest of the document a detailed description of the proposed solution for each of the VEGA partitions.

6.1.1 ARH1 [CPUP] Compute and GPU partition performance ratio

We understand that the performance of the proposed Compute and GPU partitions will be measured as the sustained performance during the High Performance Linpack benchmark (Rmax).

The sizing of each of the CPU and GPU partitions will then be made considering the efficiency of the proposed solution (choice of processor SKU and GPU model) on the HPL benchmark compared to its theoretical peak performance.

We also understand that Rmax performance between CPU and GPU partitions should be as equal as possible, with favour of the CPU partition if strict parity is not possible. Finally, Rmax performance for CPU partition will also be obtained using processors only of the GPU partition.

The proposed VEGA system will deliver the following sustained performance for Linpack double precision:

Partition	Sustained Linpack Performance
CPU Partition	3.8 Petaflop/s
GPU Partition	3.0 Petaflop/s
Total	**6.8 Petaflop/s**

6.1.2 ARH2 [CPUP] System design and architecture description

In addition to the compute solution, the VEGA system will also include (each of the following elements will be described in its corresponding section):

- A Large Capacity Storage (LCST) CEPH-based storage solution of **18TB usable capacity**, including CEPH storage nodes and an internal CEPH 25GB/s network (a pool of virtualization nodes – 8 recommended- will be dedicated to the RADOS gateway function),
- A High Capacity Storage (HCST) Solution implementing a temporary parallel Lustre file system of **1.1TB usable capacity** and offering a **400GB/s throughput**, based on a DDN appliance,
- A set of **4x InfiniBand / 100GbE data gateways** that will allow the exchange of data between the different partition/sub-systems of the proposed solution, with the required throughputs,
- A Login partition composed of 8x login nodes:
  - **4x login for the CPU partition** with the same processor configuration as for the CPU compute partition,
  - **4x login nodes for the GPU partition** with the same processor configuration and equipped with one of the same models as for the GPU partition.
- A virtualization partition composed of **30x virtualization nodes**,
- A fast interconnect network based on **Mellanox InfiniBand HDR technology**,
- A **1Gb Ethernet management network**,
- A **100Gb Ethernet network** to:
  - Enable WAN connectivity to ARNES NREN backbone with redundant connection provided by CPE (Customer-Premises Equipment) routers with prescribed technical specification requirements.
  - Provide enough bandwidth capabilities to other data centres and supercomputers within SLING network and other HPC centres in EU and Worldwide for large data transfers.
  - Provide quality, reliable and secure connectivity for all user communities.

The following figure schematically represents the global architecture of the proposed solution.

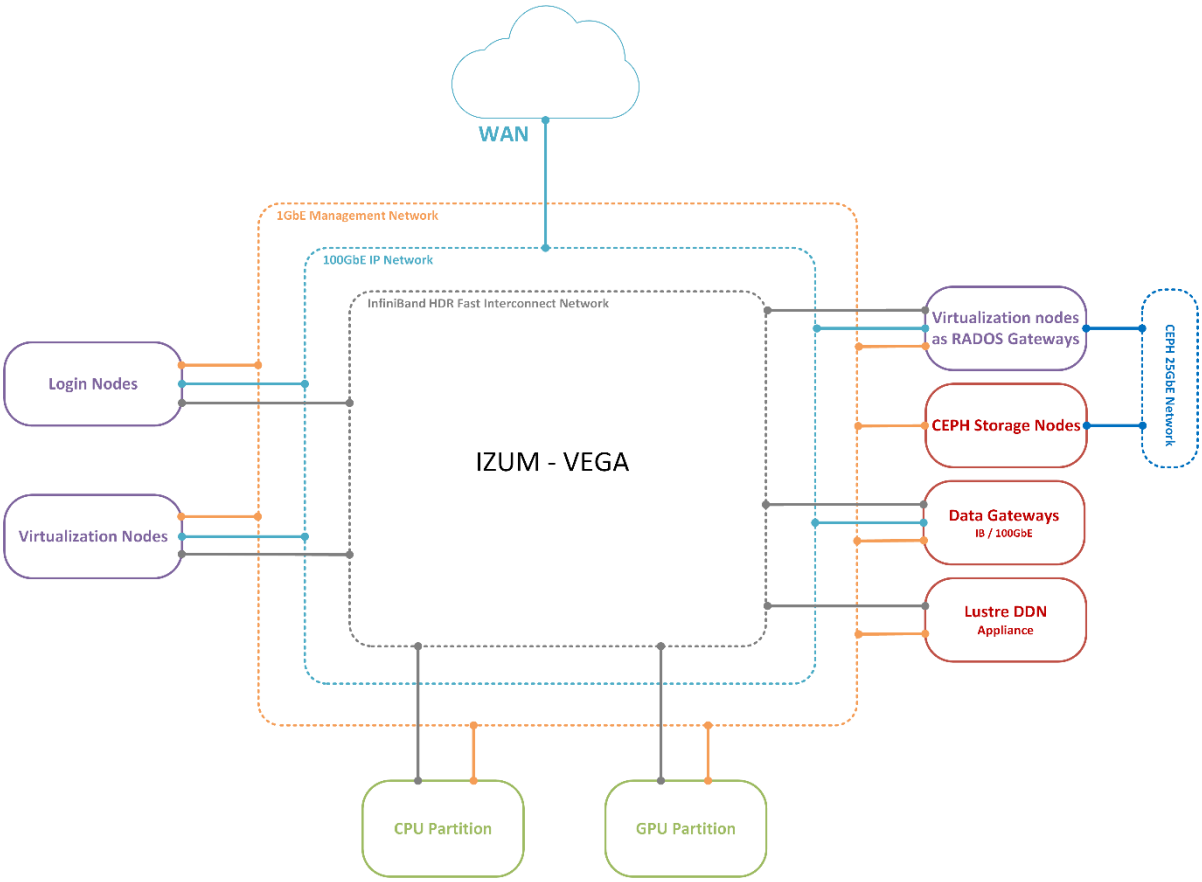


Figure 37: Global Architecture proposed for the VEGA System

6.1.3 ARH3 [CPUP] Bill of materials

Please find below a table summarizing the type and quantity of the major hardware components of the proposed solution for Vega system:

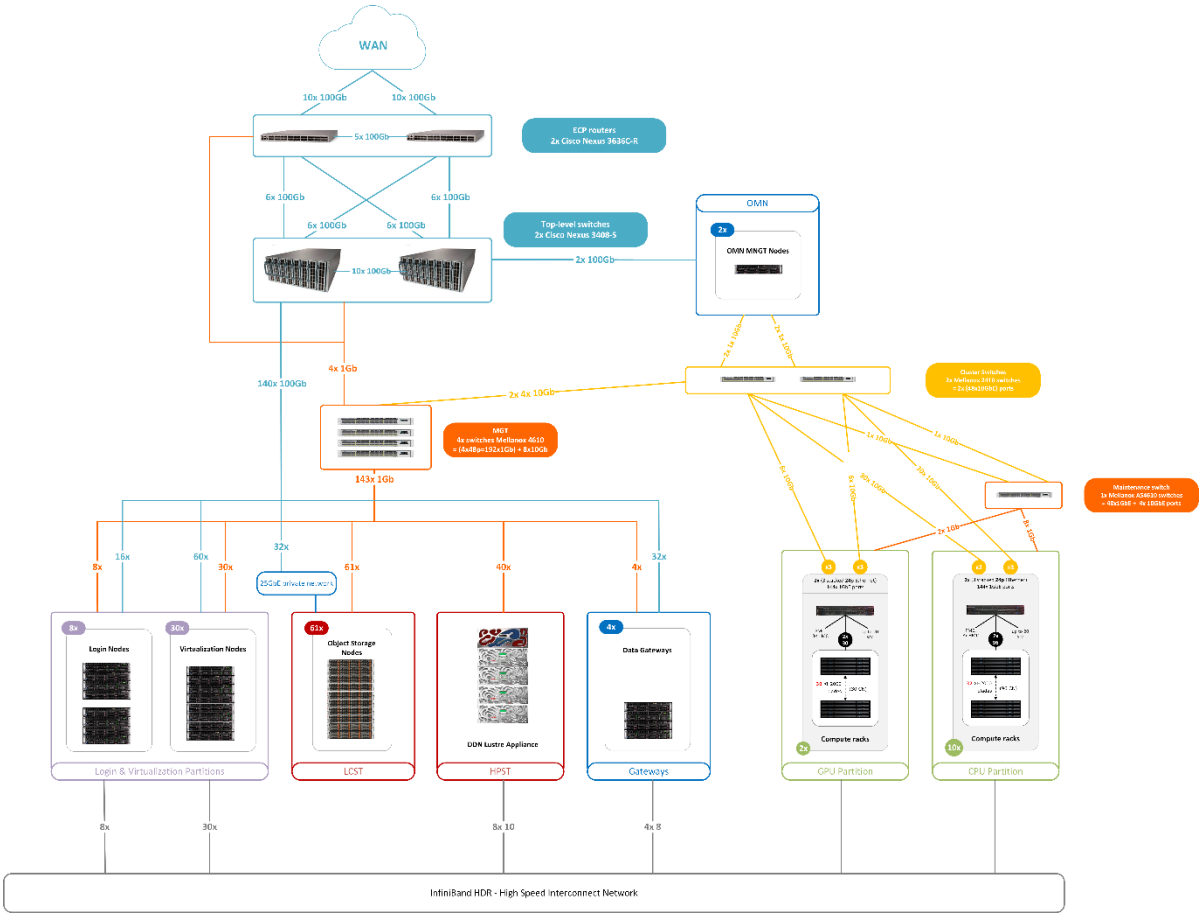
Partition	Category	Component	Quantity	Description
GPU partition	Infrastructure	Rack	2	XH2000 DLC rack with PSUs, HYC and IB HDR switches*
	Compute	GPU node	60	Per node: 4x Nvidia A100 (Nvlink) 2x AMD Rome 7H12 (64c, 2.6GHz, 280W) 512GB RAM 2x HDR dual port mezzanine 1x 1.92TB M.2 SSD

CPU partition	Infrastructure	Rack	10	XH2000 DLC rack with PSUs, HYC and IB HDR switches*
	Compute	CPU node Standard	768	201x blades of 3 compute nodes Per node: 2x AMD Rome 7H12 (64c, 2.6GHz, 280W) 256GB RAM 1x HDR100 single port mezzanine 1x 1.92TB M.2 SSD
	Compute	CPU node Large Memory	192	55x blades of 3 compute nodes Per node: 2x AMD Rome (64c, 2.6GHz, 280W) 512GB RAM 1x HDR100 single port mezzanine 1x 1.92TB M.2 SSD
CPU + GPU partition	Interconnect Network	IB switch	68	*40-port Mellanox HDR switch
	Service	Management node (OMN)	1	2x nodes in a 2U chassis. **Per nodes**: 2x Intel Skylake (12c, 2.2GHz-105W) 192 GB RAM 1x HDR100 single port mezzanine 2x 10GbE **Shared**: 9TB storage
	Compute Rack maintenance Network	Ethernet switch	1	1x Mellanox AS4610 Switch: 48x1GbRJ45 4x10GbSFP+
Login partition				
	CPU login	Login node	4	2U server 2x AMD Rome (64c, 2.6GHz-280W)

				512GB RAM 2x 7.6TB NVMe 1x HDR100 single port HCA 1x 100GbE dual ports HBA
	GPU login	Login node	4	2U server 2x AMD Rome (32c, 2.35GHz-155W) 1x NVIDIA A100 GPU 512GB RAM 2x 7.6TB SSD 1x HDR100 single port mezzanine 1x 100GbE dual ports HBA
Virtualization partition				
	Service	Virtualization node	30	2U server 2x AMD Rome (32c, 2.5GHz, 180W) 512GB RAM 6x 3.8TB NVMe 1x HDR100 single port HCA 1x 100GbE dual ports HBA
HPST				
	Storage	Flash-based building block	10	2U ES400NVX: ○ 1+1 redundant storage controllers and fully redundant IO-paths to the storage devices ○ Redundant Fans and redundant Power Supplies ○ 21 x 7.68TB NVMe devices (DWPD=1) ▪ 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.
LCST				
	Storage	Storage node	61	2U 2x Intel 4214R*, 12c, 2.4GHz, 100W 192GB RAM 1x 240GB SSD 2x 6.4TB NVMe 24x 16TB HDD 2x 25GbE 1x 1GbE IPMI
	Network	Internal Network	8	8x Mellanox SN2010. Per switch: 18x 25GbE + 4x 100GbE ports
Networks				
	Management Network	Service and IO nodes management Network	4	4x Mellanox 4610 switches, per switch: 48x 1GbE + 2x 10GbE ports
	Management Network	Top Management Network	2	2x Mellanox 2410 switches, per switch: 48x 10GbE ports
	Management Network	Data Gateways	4	4x 2U Mellanox Skyway IB to Ethernet Gateway Appliance. Per gateway: 8x IB and 8x 100GbE ports

	IP Network	IP Routers	2	2x CISCO Nexus N3K-C3636C-R
	IP Network	Top-level Switches	2	2x CISCO Nexus N3K-C3408-S



6.1.4 ARH4 [CPUP] Nodes description

The nodes description for the proposed system is provided in chapter 6.3.1 (CPU partition) and 6.3.2 (GPU partition) of this document

6.1.5 ARH5 [CPUP] SSD disk requirements

All SSD disks model procured for the Vega system will have at least 3 DWPD during warranty duration of 5 years

6.1.6 ARH6 [CPUP] RAS features in general

Adhering to enterprise RAS (Reliability, Availability, Supportability) standards, most hardware components (such as physical disks, power supply modules, and fan module) are redundant and hot-swappable.

- **Disks**
  - In nodes with more than one disk, RAID 1 or RAID 6 will be used. The failure of one disk will not have an impact on any other components of the cluster. During a rebuild, performance will be slightly affected but there will be no interruption to service.
- **PSU**
  - Each BullSequana XH2000 rack includes 6 x 15kW direct liquid cooled PSU shelves.
  - (1 single block of 6 shelves with a N+1 redundancy at 3kW PSU level)
  - All the other nodes include redundant Power supply
- **Fans for air cooled nodes**
  - Each air-cooled node has several fans that are redundant. All Fans run simultaneously and if one fails the others will turn faster if it is needed.
- **Hydraulic Chassis**
  - Hydraulic Chassis (HYC) contain the heat exchanger system that allows it to achieve 95% of heat transfer between the primary and secondary manifolds. 3 HYC are included in our proposition for N+1 redundancy. Third HYC is for N+1 redundancy only.
- **InfiniBand Network**
  - As you will see in the InfiniBand section, the cables are redundant between core switches and edge switches.
- **Memory modules**
  - All memory modules included in the nodes are delivered with ECC Memory to prevent from any memory error.
- **Services and management nodes**
  - Login: Solution includes 8x Login nodes to prevent any issue
  - Service nodes: the sizing has been made to guarantee operation even in case of failure of one of its components (fully redundant)
- **Software**
  - All critical software services are provided under RAS with high availability features and/or scalability features (up/down)
  - Linux operating systems provide RAS features through the RAS daemon service to provide centralized monitoring of event regarding hardware RAS (MCE, etc.)

6.1.7 ARH7 [CPUP] Out-of-band and remote management in general

All IT equipment of the proposed solution provide out-of-band monitoring/trending/alert capabilities. All nodes are also connected to the out-of-band management (BMC) network, providing access to console (Serial-over-LAN) and lights out management.

We propose to use the Bull Energy Optimizer (BEO) for remote power management (please refer to section 5.2 for more details about BEO).

6.2 Compute Partition

The CPU compute partition will provide supporting resources for workflows on Vega. The partition will consist of heterogeneous nodes with two x86-64 compliant CPUs. For supporting a mix of workloads while keeping costs at bay, the nodes will feature different memory sizes. The CPUs will be the same across the partition. The network is of high performance, allowing large-scale jobs.

Compute nodes of the CPU partition are built using BullSequana CPU blades.

Figure 38: BullSequana X2410 CPU blade view

The BullSequana CPU blade processors is a hot-plug blade, compatible with the BullSequana XH2000 Compute Cabinet architecture, packaging and interfaces.

Each blade contains 3 dual-socket AMD Epyc New generation (Rome or Milan), connected to the BullSequana XH2000 local interconnect network through one or two 100/200Gb/s ports.

The connection to the interconnect and to the management network is done through a mezzanine board to accommodate several types on a Network Interconnect Controller (NIC) or Host Channel Adapter (HCA). The BullSequana XH2000 interconnects currently supported are BXI, InfiniBand HDR or HDR100.

The BullSequana X2410 CPU blade is built upon a cold plate which cools all components by direct contact, except DIMMS for which custom heat spreaders evacuate the heat to the cold plate.

The BullSequana CPU blade contains:

- Three identical motherboards, one per node
- Three identical mezzanine boards for IO connection, one per node (type depends on interconnect selected)
- One common cold plate to cool all components in the blade (including the I/O).

6.2.1 CHW1 [CPUP] Compute node processor architecture

Compute nodes of the CPU partition (as well as those of the GPU partition) are equipped with AMD Epyc new generation processor code-named "Rome" (successor to current "Naples" first Epyc generation).

AMD® Epyc processor

AMD® Epyc™ second-generation processor (Codename: "Rome") is the next generation multiprocessor released mid-2019 by AMD®, serving as a successor to "Naples" and manufactured at 7-nanometer. It offers significant performance improvement from current generation production and the best performance per price and lowers TCO through an optimal balance of compute, memory, I/O and security. AMD® Epyc™ "Rome" is based on the "Zen2" microarchitecture.

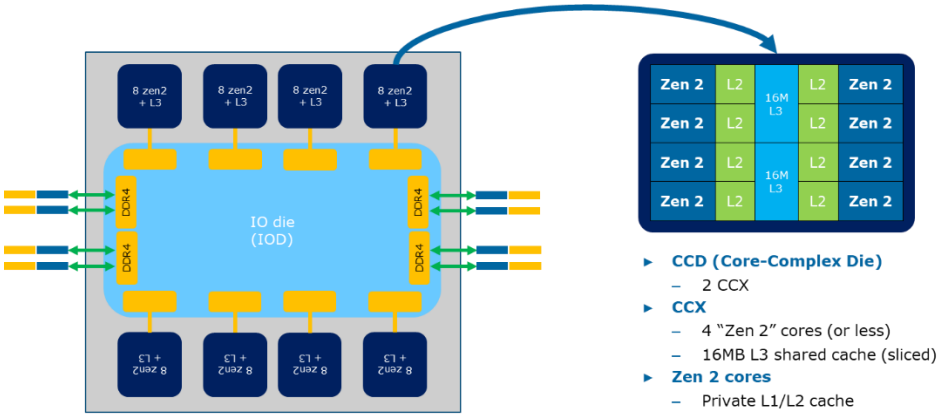
Characteristics are:

- Highly integrated SOC solution – NO Chipset
  - SP3 LGA package
  - 7nm process technology
  - TDP ranging from 120-225W
- Next generation x86 "Zen2" Core
  - Up to 32+ "Zen2" x86 cores with
    - L1 cache: 32KB for instructions, 64KB for data
    - 512KB of L2 (private to core)
    - 16MB L3 shared by a set of 4 cores (CCX)
- 8 channels 64-bit DDR4 with ECC
  - Up to 3200 MT/s (depending of the memory population rules)
  - Up to 2 DIMMs per channel (allowing maximum of 4TB/socket using 256GB DIMMs)
  - Support of RDIMM, LRDIMM, NVDIMM, Flash, 3DS
- Integrated I/O
  - 128 lanes of PCIe Gen3/4
    - Used for PCIe, SATA, Ethernet and Coherent Interconnect
    - Up to 32 SATA or NVMe devices

- 2nd generation of xGMI interconnect, extend the coherent memory system between sockets in a two-socket system
  - Running up to ~16Gbps
- High-performance Integrated Server Controller Hub (SCH)
  - USB3.0, LPC, GPIO, SPI, I2S, I2C, SMBus, UART
- AMD Security Processor
  - Trusted hardware validated boot
  - Hardware memory encryption

AMD® Epyc™ "Rome" is a server-class microprocessor System-On-a-Chip (SOC) multi-chip module (MCM). It is built using a mixed-technology chiplet approach - up to eight core/cache complex dies (CCD) and a single I/O die (IOD). The architecture of AMD® Epyc™ "Rome" is highly hierarchical:

- Each socket is built using up to 8 CCD
  - Each CCD contains 2 CCX (Core-Complex)
    - Each CCX contains 4x "Zen2" x86 cores for a total of 2MB L2 and 16MB L3 per CCX



"Zen2" Core supports up to 2 threads (Simultaneous Multi-Threading) per core. "Zen2" offers twice of the performance (FLOP/s) compare to previous or current "Zen" Cores, with 16x 64-bit floating point operations (FLOP) per cycle, which could be decomposed as:

- 2x FMUL/MAC over 256-bits vector: 8 FLOP/cycle
- 2x FADD over 256-bits vector: 8 FLOP/cycle

"Zen2" Core supports up to AVX2 instruction set (not AVX-512). Moreover, microarchitecture improvement of "Zen 2" over "Zen 1" lead to an estimate IPC ("instruction per cycle") uplift of 10-15%.

The AMD® Epyc™ "Rome" socket supports 8 memory channels with a maximum memory speed of 3200MT/s. It leads to a theoretical memory bandwidth of 204.8 GB/s offering an uplift 20% memory bandwidth compare to the current "Naples" running at 2667 MT/s. Note that the memory performance, and the maximum speed supported by the memory controller is dependent of the population rules and the characteristics of the memory DIMMs used (capacity and organization).
Compare to "Naples", "Rome" offers a reduced system diameter (NUMA domain), minimum of one NUMA domain per socket, which certainly improve the memory performance (bandwidth and latency) over prior generation.

For the CPU partition of the VEGA system, we have selected the AMD EPYC Rome 7H12 processor:

Partition	CPU instruction set	# sockets per node	# cores per node
CPU partition	x86-64	2	128

Partition	CPU reference and specifications
CPU partition	AMD Epyc Rome 7H12, 64c, 2.6GHz, 280W

The compute partition is built on **960x BullSequana XH2410 compute nodes** (i.e. 320x BullSequana XH2410 compute blades) offering a sustained HPL performance of **3.58 PFlops**.

Note that the total CPU Rmax will be obtained using all processors from the CPU partition and also CPU processors only from the GPU partition as the processor SKU is the same for both partitions. In that case, the total CPU Rmax will be 3.8PFlops.

6.2.2 CHW2 [CPUP] Compute node memory capacity

With 128 CPU cores per CPU compute nodes, we equipped each of the compute nodes with at least 256GB of DDR4 @ 3200MT/s memory capacity, satisfying the minimum of 2GB of memory per CPU core.

6.2.3 CHW3 [CPUP] Fraction of High-memory compute nodes

Our proposal includes a total of 960x CPU compute nodes. 768x compute nodes will be equipped with 256GB of memory (80%) and 192x compute nodes (20%) will be dedicated to the fraction of high-memory compute nodes.

6.2.4 CHW4 [CPUP] High-memory compute node memory capacity

High memory nodes will be equipped with 1TB of DDR4 @3200MT/s. The following table summarizes the memory configuration of both "standard memory" and "high memory" CPU compute nodes:

CPU Partition	# memory DIMMs per node	Type of memory DIMM
standard memory nodes	16	16GB DDR4 RDIMM 3200MT/s DR
High memory nodes	16	64GB DDR4 RDIMM 3200MT/s DR

6.2.5 CHW5 [CPUP] Uniform compute node performance

For both standard-memory nodes and high-memory nodes, all memory channels are fully populated; only the memory DIMMs size changes. Hence, there will be no impact of the node memory configuration on the STREAM performance.

6.2.6 CHW6 [CPUP] Compute node dedicated storage

All BullSequana XH2410 compute nodes are equipped with a single 1.92TB NVMe local storage.

6.2.7 CHW7 [CPUP] Compute node sensors

A wide range of sensors are accessible from the compute nodes via InBand and OutOfBand. Please find below a listing of accessible sensors:

Sensor Name	Event Trigger
CPU0_TEMP	Upper and Lower Critical and Non-critical Temperature
CPU1_TEMP	
DIMMG0_TEMP	
DIMMG1_TEMP	
DIMMG2_TEMP	
DIMMG3_TEMP	
MB_TEMP1	
MB_TEMP2	
RISER1_TEMP	
VR_P0_TEMP	
VR_P1_TEMP	
VR_DIMMG0_TEMP	
VR_DIMMG1_TEMP	
VR_DIMMG2_TEMP	
VR_DIMMG3_TEMP	

LM5066_VIN	Upper and Lower Critical and Non-critical Voltage
P_12V	
P_1V0_AUX_LAN	
P0_VDD_18	
P1_VDD_18	
P_3V3	
P_5V	
P_5V_STBY	
P_VBAT	
P0_VDDCR_SOC	
P0_VDDCR_CPU	
P1_VDDCR_CPU	
P1_VDDCR_SOC	
P0_VDDIO_ABCD	
P0_VDDIO_EFGH	
P1_VDDIO_ABCD	
P1_VDDIO_EFGH	
VR_P0_VIN	
VR_P1_VIN	
VR_DIMMG0_VIN	
VR_DIMMG1_VIN	
VR_DIMMG2_VIN	
VR_DIMMG3_VIN	
VR_P0_VOUT	
VR_P1_VOUT	
VR_DIMMG0_VOUT	
VR_DIMMG1_VOUT	
VR_DIMMG2_VOUT	
VR_DIMMG3_VOUT	
LM5066_IIN	Upper and Lower Critical and Non-critical Current
VR_P0_IOUT	
VR_P1_IOUT	
VR_DIMMG0_IOUT	
VR_DIMMG1_IOUT	
VR_DIMMG2_IOUT	
VR_DIMMG3_IOUT	

Sensor Name	Event Trigger
CPU0_Status	IERR
	Thermal Trip
	Presence detected
	Processor disabled
CPU1_Status	IERR
	Thermal Trip
	Presence detected
	Disabled
Memory	Uncorrectable ECC Error
	Correctable ECC Error threshold reached
EventLoggingDisabled	LogAreaReset/Cleared
	All Event Logging Disabled
SystemEvent	Oem System Boot Event
	Timestamp Clock Synch.
OSBoot	C: boot completed
	PXE boot completed
	Boot completed

Sensor id	Sensor Type	Explanation / Source
MBCPLD	Power Detection	VDD_33_DUAL_PG
		PWRGD_BMC_ALL
		P0_VDD_18_DUAL_PG

		P1_VDD_18_DUAL_PG
		P0_VDD_SOC_DUAL_PG
		P1_VDD_SOC_DUAL_PG
		P12V_HP_PG
		PWRGD_PS_PWROK
		P0_VDD_VPP_ABCD_SUS_PG
		P1_VDD_VPP_ABCD_SUS_PG
		P0_VDD_VPP_EFGH_SUS_PG
		P1_VDD_VPP_EFGH_SUS_PG
		P0_VDD_VTT_ABCD_SUS_PG
		P1_VDD_VTT_ABCD_SUS_PG
		P0_VDD_VTT_EFGH_SUS_PG
		P1_VDD_VTT_EFGH_SUS_PG
		P0_VDD_18_RUN_PG
		P1_VDD_18_RUN_PG
		P0_VDD_SOC_RUN_PG
		P1_VDD_SOC_RUN_PG
		P0_VDD_CORE_RUN_PG
		P1_VDD_CORE_RUN_PG
		PSU_PWROK
		P0_SLP_S3_L
		P0_33_PWROK
		P1_33_PWROK
		VDD_5_RUN_PG
		VDD_33_RUN_PG
		P0_PWRGD_OUT
		P1_PWRGD_OUT
		PWR_GD_SLOT3_N
		PWR_GD_SLOT4_N
		PWR_GD_SLOT5_N
		PWR_GD_SLOT6_N
MBCPLD	Throttle	PWR_REDUCTION
		FAST_PROC_HOT_N_3V3_R1
		P0_PROCHOT_L
		P1_PROCHOT_L
MBCPLD	Presence Detection	PRSNT_SLOT3_N
		PRSNT_SLOT4_N
		PRSNT_SLOT5_N
		PRSNT_SLOT6_N
CWG CPLD	Power Default Detection	EPO detected
		Timeout Power Sequencing
		Redstone Power Good Status
		Redstone Management Status
		P0V9 VR Power Good Status for PCIe Switch 1
		P1V8 VR Power Good Status for PCIe Switch 1
		P0V9 VR Power Good Status for PCIe Switch 0
		P1V8 VR Power Good Status for PCIe Switch 0
		P5V VR Power Status for PCIe Switch 1
		P5V VR Power Status for PCIe Switch 0
		P12V Main Mezzanine Power Good Status
		P3V3 Power Good Status
CWG CPLD	Throttling Source	Throttling feedback request from BMC CER
		Over Temperature on CWG board
		Over Temperature on P0V9 VR controller for SW0
		Over Temperature on P0V9 VR controller for SW1
		Fast Proc Hot requirement from Rack System
CWG CPLD	Throttling Source	
CWG CPLD	Alert and Error Source	Redstone Alert I2C1 (GPU module)
		Redstone Alert I2C2 (Redstone Board)
		VR SW0 SMB Alert
		VR SW1 SMB Alert
		P12V Main Hotswap Fault

		SW0 PCIe Error
		SW1 PCIe Error
		Over Temperature on Redstone Board
		Power Input VR SW0 Alert
		Power Input VR SW1 Alert
		Under / Over Voltage / Current on P54V via Hotswap LM5066i, Temperature, CML …
CWG CPLD	FSM power control	EPO: Emergency Power off Redstone Boards
CWG CPLD	LM5066	event data2: LM5066 PMBus command event data3: Response value from LM5066
RDSTN	OVERT INFO	GPU1
		GPU2
		GPU3
		GPU4
RDSTN	Interrupt Status 1	GPU1 presence status
		GPU2 presence status
		GPU3 presence status
		GPU4 presence status
		GPU1 PWR_GOOD status
		GPU2 PWR_GOOD status
		GPU3 PWR_GOOD status
		GPU4 PWR_GOOD status
		GPU1 PCIE interface presence status
		GPU2 PCIE interface presence status
		GPU3 PCIE interface presence status
		GPU4 PCIE interface presence status
RDSTN	Interrupt Status 2	SysVR 1V8 power supply status
		HSC #1 power supply status
		HSC #2 power supply status
		HSC #3 power supply status
		HSC #4 power supply status
		Output HSC power supply status

6.2.8 CHW8 [CPUP] Compute node High-speed interconnect connectivity

We consider a 100 Gbit/s injection bandwidth as the most cost-effective design for the CPU partition. Hence, each CPU compute node is equipped with a single port HDR100 InfiniBand mezzanine:

Partition	# physical connectors per node	Type of High-Performance Network Adapter
CPU	1	Mellanox ConnectX-6 HDR100 InfiniBand Mezzanine

This configuration provides, per compute node of the CPU partition, a theoretical injection bandwidth of 100 Gbit/s.

6.3 GPU Partition

6.3.1 GPU partition hardware requirements

The GPU compute partition will provide a computational performance equivalent to one provided by the CPU partition. The partition consists of homogeneous nodes with 4x Accelerators and two CPUs.

The proposed solution for the GPU partition is based on our **BullSequana X2415 GPU Blade:** an AMD-based host node that houses four (4) Nvidia next generation Ampere A100 GPUs together with the "cold plate" for hot water cooling. The power supply (54V DC) is provided centrally by the compute rack for all blades.

Figure 39: BullSequana X2415 GPU blade view

The host node is a typical dual-socket system with 8 memory DIMMs per socket and one (optional) 2.5" SATA SSD M.2 or NVMe. The mezzanine cards include the necessary PCIe switches, the interfaces for the high-speed interconnect (HDR InfiniBand), the Ethernet connections and sideband management (BMC/IPMI). Up to four (4) HDR ports are made available to the four (4) GPUs of the node (up to one high-speed interconnect port per accelerator).

6.3.1.1 GHW1[GPUP] Amount of GPU node main processor cores

The GPU partition compute nodes are equipped with AMD Epyc new generation processor code-named “Rome” (successor to current “Naples” first Epyc generation) which architecture description have been provided in section 6.2

For the GPU partition of the Vega system, we have selected the AMD EPYC Rome 7452 processor and NVIDIA Ampere GPUs:

Partition	CPU instruction set	# sockets per node	# cores per node
GPU partition	x86-64	2	128

Partition	CPU reference and specifications
GPU partition	AMD Epyc Rome 7H12, 64c, 2.6GHz, 280W

The GPU compute partition is built on **60x BullSequana XH2415 compute nodes** (i.e. 60x BullSequana XH2415 blades) offering a sustained HPL performance of **3.0 PFlops**.

6.3.1.2 GHW2 [GPUP] GPU node main memory capacity

With 64 CPU cores per GPU compute nodes, the combination of the min 256GB memory capacity and at least 2GB per CPU core requirements leads to building GPU compute node with at least 256GB memory. However, with the proposed NVIDIA A100 GPU new architecture it is critical to balance memory capacity between the node and GPU units. Hence, the memory configuration of the GPU nodes will be optimized with 512GB capacity.

Each GPU node will be equipped with 16x 32GB DDR4 @3200MT/s DIMMs.

CPU Partition	# memory DIMMs per node	Type of memory DIMM
all nodes	16	32GB DDR4 RDIMM 3200MT/s DR

6.3.1.3 GHW3 [GPUP] Fraction of High-memory GPU nodes

We understand, from the Q&As between the contractor and different vendors that requirement for large memory GPU nodes has been relaxed, allowing vendors to propose “high-memory” GPU nodes with 512GB capacity.

Since, this capacity corresponds to what we believe to be the minimum requirement for the GPU nodes with the proposed NVIDIA GPU A100 technology, we won’t make any distinction between “standard-memory” and “high-memory” nodes.

All nodes of the GPU partition will have the same memory configuration as described above.

6.3.1.4 GHW4 [GPUP] High-memory GPU node main memory capacity

We understand, from the Q&As between the contractor and different vendors that requirement for large memory GPU nodes has been relaxed, allowing vendors to propose "high-memory" GPU nodes with 512GB capacity. Since, this capacity corresponds to what we believe to be the minimum requirement for the GPU nodes with the proposed NVIDIA GPU A100 technology, we won't make any distinction between "standard-memory" and "high-memory" nodes. All nodes of the GPU partition will have the same memory configuration as described above.

6.3.1.5 GHW5 [GPUP] Uniform GPU node performance

Both standard-memory nodes and high-memory nodes have the same memory configuration. Hence, there will be no impact of the node memory configuration on the STREAM performance.

6.3.1.6 GHW6 [GPUP] GPU node dedicated storage

All BullSequana XH2415 GPU compute nodes are equipped with a single 1.92TB NVMe local storage.

6.3.1.7 GHW7 [GPUP] Graphics processor memory capacity

To feed its massive computational throughput, the NVIDIA A100 GPU has 40 GB of high-speed HBM2 memory with a class-leading 1555 GB/sec of memory bandwidth—a 73% increase compared to Tesla V100. In addition, the A100 GPU has significantly more on-chip memory including a 40 MB Level 2 (L2) cache—nearly 7x larger than V100—to maximize compute performance. With a new partitioned crossbar structure, the A100 L2 cache provides 2.3x the L2 cache read bandwidth of V100.

To optimize capacity utilization, the NVIDIA Ampere architecture provides L2 cache residency controls for you to manage data to keep or evict from the cache. A100 also adds Compute Data Compression to deliver up to an additional 4x improvement in DRAM bandwidth and L2 bandwidth, and up to 2x improvement in L2 capacity.

Partition	# accelerators per node	IEEE compliance
GPU partition	4	All CUDA compute devices follow the IEEE 754 standard for binary floating-point representation, with some small exceptions. https://docs.nvidia.com/cuda/floating-point/index.html

Partition	GPU reference and specifications
GPU partition	NVIDIA Redstone board with 4x "Ampere A100", each with the following specifications: • 40GB high-speed HBM2 memory, 1555GB/s memory bandwidth • 108 Stream Processors • 6912 FP32 CUDA Cores • 6912 INT32 CUDA Cores • 3456 FP64 CUDA Cores • 432 New Tensor Cores

6.3.1.8 GHW8 [GPUP] Graphics subsystem description

The BullSequana XH2415 can host 2x processors and 4x NVIDIA GPUs. The node design has been optimized to provide optimal bandwidth between components:

- Processor-to-Processor: 4x Infinity fabric links (aka xGMI-2) providing 4x 32GB/s
- Processor-to-Device: one PCIe-4 x16 dedicated channel per device providing 32GB/s (16Gbps per lane)
- Device-to-Device: direct NVLINK3 connection providing 50GB/s (2x 25GB/s)
- Device-to-External-Device: one PCIe-4 x16 dedicated channel per device providing 32GB/s

The following diagram shows the block diagram of an XH2415 GPU node with the internal communication channels:

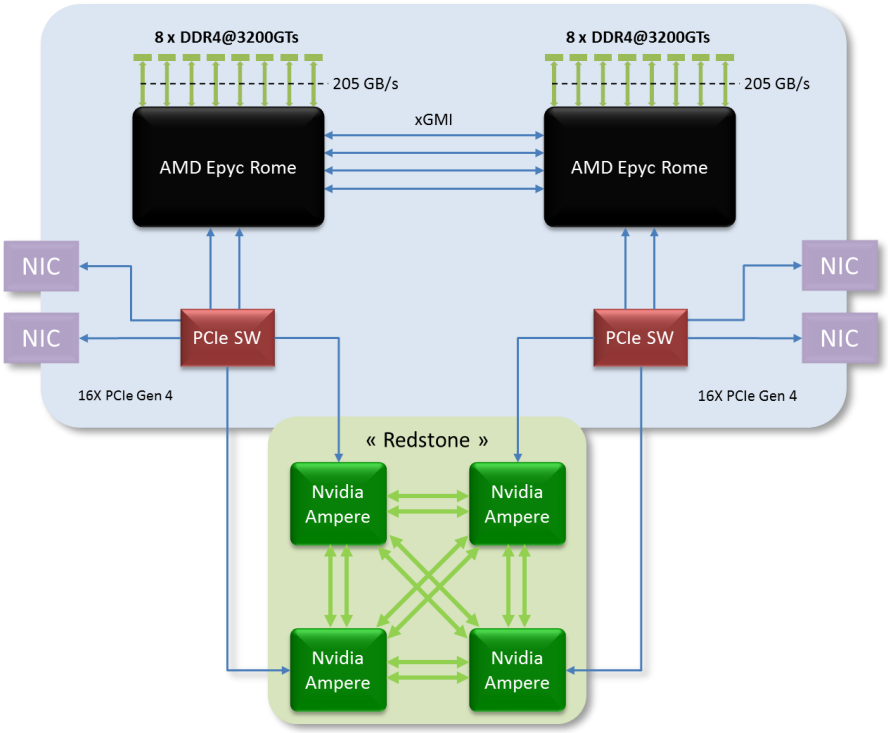


Figure 40. X2415 GPU node internal communication channels (4-NIC configuration presented).

6.3.1.9 GHW9 [GPUP] GPU node High-speed interconnect connectivity

Please find below the description of the solution proposed to fulfil these requirements:

Partition	# physical connectors per node	Type of High-Performance Network Adapter
GPU	2	2x Mellanox ConnectX-6 HDR InfiniBand Mezzanine

This configuration provides, per node, a theoretical injection bandwidth of 400 Gbit/s. We understand that it might seem oversized compared to the initial requirement, but this configuration and bandwidth are required to take full advantage of the NVIDIA A100 architecture and the BullSequana XH2415 node design.

6.3.2 GPU partition software requirements

6.3.2.1 GSW1 [GPUP] Graphics acceleration

The proposed GPU partition is composed of BullSequana XH2415 blades equipped each with 4x NVIDIA A100 GPUs. The NVIDIA A100 Tensor Core GPU delivers unprecedented acceleration at every scale for AI, data analytics, and HPC to tackle the world's toughest

computing challenges. As the engine of the NVIDIA data center platform, A100 can efficiently scale up to thousands of GPUs or, using new Multi-Instance GPU (MIG) technology, can be partitioned into seven isolated GPU instances to accelerate workloads of all sizes. A100's third generation Tensor Core technology now accelerates more levels of precision for diverse workloads, speeding time to insight as well as time to market.

The graphics acceleration features are summarized in the following table.

	NVIDIA A100 for NVIDIA HGX™	NVIDIA A100 for PCIe
GPU Architecture	NVIDIA Ampere	
Double-Precision Performance	FP64: 9.7 TFLOPS FP64 Tensor Core: 19.5 TFLOPS	
Single-Precision Performance	FP32: 19.5 TFLOPS Tensor Float 32 (TF32): 156 TFLOPS | 312 TFLOPS*	
Half-Precision Performance	312 TFLOPS | 624 TFLOPS*	
Bfloat16	312 TFLOPS | 624 TFLOPS*	
Integer Performance	INT8: 624 TOPS | 1,248 TOPS* INT4: 1,248 TOPS | 2,496 TOPS*	
GPU Memory	40 GB HBM2	
Memory Bandwidth	1.6 TB/sec	
Error-Correcting Code	Yes	
Interconnect Interface	PCIe Gen4: 64 GB/sec Third generation NVIDIA® NVLink®: 600 GB/sec**	PCIe Gen4: 64 GB/sec Third generation NVIDIA® NVLink®: 600 GB/sec**
Form Factor	4/8 SXM GPUs in NVIDIA HGX™ A100	PCIe
Multi-Instance GPU (MIG)	Up to 7 GPU instances	
Max Power Consumption	400 W	250 W
Delivered Performance for Top Apps	100%	90%
Thermal Solution	Passive	
Compute APIs	CUDA®, DirectCompute, OpenCL™, OpenACC®	

* Structural sparsity enabled

** SXM GPUs via HGX A100 server boards; PCIe GPUs via NVLink Bridge for up to 2 GPUs

6.3.2.2 GSW2 [GPUP] GPU node device drivers and software libraries

The proposed solution for the compute GPU partition comes with all appropriate device drivers and GPU specific software libraries:

Math Libraries

GPU-accelerated math libraries lay the foundation for compute-intensive applications in areas such as molecular dynamics, computational fluid dynamics, computational chemistry, medical imaging, and seismic exploration.

- cuBLAS: GPU-accelerated basic linear algebra (BLAS) library
- cuFFT: GPU-accelerated library for Fast Fourier Transforms
- CUDA Math Library: GPU-accelerated standard mathematical function library
- cuRAND: GPU-accelerated random number generation (RNG)
- cuSOLVER: GPU-accelerated dense and sparse direct solvers
- cuSPARSE: GPU-accelerated BLAS for sparse matrices
- cuTENSOR: GPU-accelerated tensor linear algebra library
- AmgX: GPU-accelerated linear solvers for simulations and implicit unstructured methods

Parallel Algorithm Libraries

GPU-accelerated libraries of highly efficient parallel algorithms for several operations in C++ and for use with graphs when studying relationships in natural sciences, logistics, travel planning, and more

- nvGRAPH: GPU-accelerated library for graph analytics
- Thrust: GPU-accelerated library of C++ parallel algorithms and data structures

Image and Video Libraries

- nvJPEG: High performance GPU-accelerated library for JPEG decoding
- NVIDIA Performance Primitives: Provides GPU-accelerated image, video, and signal processing functions
- NVIDIA Video Codec SDK: A complete set of APIs, samples, and documentation for hardware-accelerated video encode and decode on Windows and Linux
- NVIDIA Optical Flow SDK: Exposes the latest hardware capability of NVIDIA Turing™ GPUs dedicated to computing the relative motion of pixels between images

Communication Libraries

Performance-optimized multi-GPU and multi-node communication primitives.

- NVSHMEM: OpenSHMEM standard for GPU memory, with extensions for improved performance on GPUs.
- NCCL: Open-source library for fast multi-GPU, multi-node communications that maximizes bandwidth while maintaining low latency.

Deep Learning Libraries

GPU-accelerated libraries for image and video decoding, encoding, and processing that leverage CUDA and specialized hardware components of GPUs.

- NVIDIA cuDNN: GPU-accelerated library of primitives for deep neural networks
- NVIDIA TensorRT™: High-performance deep learning inference optimizer and runtime for production deployment
- NVIDIA Jarvis: Platform for developing engaging and contextual AI-powered conversation apps
- NVIDIA DeepStream SDK: Real-time streaming analytics toolkit for AI-based video understanding and multi-sensor processing
- NVIDIA DALI: Portable, open-source library for decoding and augmenting images and videos to accelerate deep learning applications

6.4 Storage subsystem partitions

6.4.1 Storage subsystem technical requirements for both tiers

6.4.1.1 STO1 [STOR] Storage subsystem high availability

HPST

The proposed storage solution is built from DDNs ES400NVX. Those High-Performance Storage building blocks are designed for building parallel filesystem solutions.

Each ES400NVX is equipped with:

- 1+1 redundant storage controllers
- 1+1 redundant power supplies
- 4+1 redundant fans per storage controller
- redundant data paths to the storage devices (DualPorted NVMe – one port to each controller)
- redundant network uplinks

With this configuration, the ES400NVX itself has no single point of failure. According to the clarification question and the according answer by IZUM, an ES400NVX is not considered a failure domain.

With the definition given for such cases: "In such cases, one path from the point of storage device (such as NVMe or SSD) to the point of shared network/interconnect shall be considered a failure domain. ", the following protection level is given:

- any two devices can fail without losing data or access to data (8+2p data protection)

- rebuild is done using a hot spare device. Due to the nature of flash a rebuild can be done very fast.
- In each ES400NVX a controller can fail without losing access to data (up to 50% of the controllers)
- If both controllers in a single ES400NVX fail access to data is interrupted
- A single controller (failure domain) controls ~60 TByte

LCST

Considering the failure domain definition, the biggest failure domain in the proposed system is a complete OSD node. This represents a capacity of 384 TB (24x 16 TB HDDs, please also refer to 6.4.2 below).

On LCST, data is protected by a high-level of redundancy: erasure coding. Erasure coding allows setting arbitrary ratios of original data and coding data. With a ratio of *m* parts of original data to *n* parts of coding data, the code can tolerate the loss of any *n* parts and regenerate the original *m* parts. For example, a code of *m*:*n* = 8:3 enriches every 8 parts of data with 3 parts of coding data and spreads the data across 11 (8 + 3) disks. This encoding can then tolerate the loss of any 3 disks and generates a redundancy blow-up of just (*m* + *n*)/*n* = (8 + 3)/8 = 1.375. Compare this with three-way quorum replication, which can tolerate the loss of only 1 or 2 disks and has a blow-up of a factor of 3.

In the event of a drive or complete node failure, Ceph will automatically initiate a recovery procedure, regenerating the lost data from the erasure coding to remaining drives and nodes in the system. The system remains operational during the recovery process, and all data is still usable while being regenerated.
Ceph will be configured so that any 2 nodes can fail without impacting data integrity by using an erasure coding pool in 8+3 mode and instructing it to place each of the erasure coding components on different nodes.

6.4.1.2 STO2 [STOR] Availability during failure

HPST

The whole storage solution is built in such a way that it has no single point of failure. The following scenarios are dealt with in the described way:

- Media error: All data is protected with a 8+2p parity scheme. Any two devices can fail without interruption to data access. In case a device fails the reconstruction of parity will start immediately with a Hot Spare media present in the system. The rebuild of a 7.68TB media is expected to take 5 to 6 hours in an idle system – to achieve this only about 20% of the devices performance is used. The impact in a system with user load can be controlled with setting the rebuild

priority accordingly. With a 36-hour time window the performance impact can certainly be kept at less than 15%.

- Network error: All (virtualized) storage servers have two network uplinks that are configured for redundancy. Losing one of these links degrades the performance slightly but certainly less than 15% of peak performance – and expected to be not noticeable for usual production.

- Controller: Controllers are 1+1 redundant in each ES400NVX. If one controller fails, the second takes over all storage devices managed by the failed controller to allow continuous access to data. The performance impact depends on the access pattern to the storage system. For a general-purpose high throughput scenario where many concurrent jobs are accessing a magnitude of files across the whole storage space the impact will be 10 to 15%. This is given as an average number and individual jobs can see higher/lower performance impacts.

For a single job that utilizes the whole peak performance of the storage system the performance degradation might be as high as 50% depending on the exact access pattern. For newly written data degradation can be limited to 10 to 15% if needed.

The replacement of a controller takes about 2 hours after the problem has been analysed and spare part is available. It can be done without taking the system down.

A controller replacement is an exceptional event and usually does only happen rarely in the lifetime of a system.

LCST

The ST16000NM008G Seagate drive provides a minimum of 120MB/s (approximately half of the maximum performance) in sequential read or write. Considering the loss of an entire node, and in the worst-case scenario where those drives are almost full at the time of the failure, each rebuilt drive would require writing the capacity of 1 drive spread across all remaining nodes. As a node contains 24 drives, this is 24 drives * 16TB = 384TB to write on the 60 remaining nodes (for the first node failure). As each node has a minimum potential write speed of 24 drives * 120 MB/s = 2880 MB/s, this means that recovery could be, at minimal drive speed, be finished in 384 000 000 MB / (2 880 MB/s * 60 nodes) = 2222 seconds or 37 minutes.

However in order to decrease the production performance loss below the 15% during this recovery process, some tuning on the Ceph configuration will need to be performed at installation time to adjust the speed of the recovery process (number of backfills, client and server op priority, number of osd recovery threads, …).

Considering that the full speed of each node is approximately 6.2GB/s (24 drives * 260 MB/s or 2x25Gb ethernet links / 8 bytes), the performance impact would be (2880 MB/s

* 60 nodes) / (6200 MB/s * 61 nodes) = 46%. Slowing down the recovery process so that it completes in 2.5 hours would decrease the performance impact below 12%.

6.4.1.3 STO3 [STOR] Storage subsystem reliability

HPST

All DDN storage systems use T10PI checksums between storage controllers and media (disks, NVMe) to protect data from modification. This T10PI information is checked upon every read for validity. Additionally, there is a background process running that cycles with low priority through all data and constantly checks all data.

If a mismatch is found data is recovered from parity (if possible).

Optional:
The checksums can be extended from the storage controllers up to the level of the filesystem client running on compute nodes to create true End2End-data-protection. By default, this mechanism is not enabled.

All performance data is given without this mechanism active. The impact (depending on the application) usually is between 5 and 15% but can exceed that for certain workloads.

LCST

Ceph integrates a light and deep scrubbing processes which will detect otherwise silent corruptions. Both processes can be scheduled depending on production constraints, preferably during light production time frames.

The light scrubbing will browse through metadata (the objects attributes), in order to detect inconsistencies and repair them. It is recommended to schedule this light scrub daily.

The deep scrubbing process will make sure that the content of all objects is readable and is consistent with checksums and erasure coding parities. Unreadable blocks are replaced, and any inconsistency detected is repaired.

6.4.1.4 STO4 [STOR] Storage subsystem memory per disk

HPST

Each ES400NVX controller contains 21 storage devices – and 4 VMs. All VMs combined have a RAM assignment of more than 100 GB. With that the memory is more than 2 GB per device.

LCST

Please find below the description of the solution proposed to fulfil these requirements:

Partition	# memory DIMMs per node	Type of memory DIMM	Memory Capacity per HDD in the node
LCST Object Storage Nodes	8	32GB DDR4 RDIMM 2933MT/s DR	10.6 GB

Each LCST Object Storage Node is equipped with 256 GB main memory capacity, i.e. 10.6 GB memory per drive (24x drives per server).
All memory channels are populated with the same speed and capacity memory DIMMs at the highest frequency available for the included CPU model.

6.4.1.5 STO5 [STOR] Storage subsystem file system

HPST

EXAScaler is based on a parallel filesystem that allows shared access to data from multiple clients. The filesystem used is POSIX compliant.

LCST

The Ceph File System, or CephFS, is a POSIX-compliant file system built on top of Ceph's distributed object store, RADOS. CephFS endeavors to provide a state-of-the-art, multi-use, highly available, and performant file store for a variety of applications, including traditional use-cases like shared home directories, HPC scratch space, and distributed workflow shared storage

6.4.1.6 STO6 [STOR] Storage subsystem Ethernet connectivity

HPST

In compliance with the answer to our clarification question, the EXAScaler solution (HPST) does not provide any 100GE connectivity.

LCST

Object storage nodes of the LCST proposed solution do not have direct 100Gb/s connectivity. However, through the internal CEPH network, the LCST is connected with 100GbE links to the IP network, as described section 6.4.2.

6.4.1.7 STO7 [STOR] Storage subsystem High-speed interconnect connectivity

HPST

The EXAScaler solution connects to the High-speed interconnect with multiple HDR100 links.

LCST

LCST is connected to the high-speed interconnect network through RADOS gateways (min 8 recommended) to be taken, as clarified from Q&As, from the pool of Virtualization nodes. Those nodes have HDR100 connectivity.

6.4.1.8 STO8 [STOR] Storage subsystem redundant power supply

HPST

All components of the EXAScaler solution are equipped with 1+1 redundant power supplies. These devices can be exchanged without taking the system offline (Hot Swappable).

LCST

All components of the LCST solution (servers, gateways, switches) are equipped with 1+1 redundant power supplies.

6.4.1.9 STO9[STOR] Data disks counted for storage capacity

HPST

ES400NVX units store their operating system and VM images on internal devices. Those devices are 1+1 protected and are not considered for providing the given capacity for the storage solution.

LCST

Only disks that contributes towards the total storage capacity of the LCST solution are considered to size the proposition, and disks for OS are excluded of the capacity calculation.

6.4.1.10 STO10 [STOR] Storage subsystem scalability

HPST

The offered EXAScaler solution for the HPST can be expanded with more building blocks to extend the capacity, extend the performance or both at the same time.

The system can be extended to multiple PBytes of capacity and several hundred GB/s sequential bandwidth.

LCST

The proposed CEPH architecture is highly extensible in terms of capacity and throughput (virtually with no limit). Up to 3 LCST Object Storage Nodes similar to the ones proposed in the configuration (see section 6.4.2 below) can be added with no need for network extension. Each added node brings 384 TB of raw storage in the Ceph infrastructure and an additional 6 GB/s of throughput.

Above 3 nodes, an additional Mellanox SN2010 first level switch will be added to support another group of up to 8 nodes, in the limit of available ports on top-level switches (2x 100GbE links connect each Mellanox SN2010 switch to each top-level switches).

Ceph supports addition of nodes and storage without downtime, so you can decide to extend your storage at any needed time.

6.4.2 Large capacity storage tier technical requirements

6.4.2.1 LCS1 [STOR] Large capacity storage tier design

The reference design of the proposed Large Capacity Storage solution is based on Ceph. The reference design uses 4 racks to achieve 18 PB of usable storage space. Data replication would primarily use erasure encoded (EC) pools with 8+3 scheme. Certain parts would still use a 3-replica scheme for pool redundancy.
The object storage hardware will be provided as air-cooled systems. The software of the object storage system will be installed and maintained by the hosting entity; this partition will be provided as hardware only with no need to integrate it into any cluster management stack.

The LCST CEPH design is based on a pool of Object Storage nodes that will host enough mechanical hard drives to provide the required **usable capacity of 18TB**, a pool of Object Gateway nodes (RADOS Gateways) that will be taken from the 30x provided virtualization nodes. The minimum recommended number of RADOS Gateways is 8x servers, allowing to reach the required bandwidth capacity of **200 GB/s measured as throughput for parallel streaming to and from High-performance storage tier**. Finally, an **internal 25GbEthernet network** composed of 8x Mellanox SN2010 switches (18x 25GbE + 4x 100GbE ports per switch) will allow connecting the LCST to the rest of the proposed solution.

CEPH Storage Nodes

The Ceph storage nodes (OSD) will consist of 2U servers hosting 24x mechanical hard drives (HDD) of 16TB storage capacity each. The total storage capacity of a single OSD node is 384TB, thus not allowing failure domain to exceed 400 TB of raw disk capacity.

Additionally, each node will have 2x NVMe drives, providing at least 3% of the combined storage capacity of all the hard drives in the node. The NVMe capacity needs will be supplied by at two drives, and the performance of the drives needs is such that if all access to the HDDs were routed through the SSDs they will not become a bottleneck, i.e. the combined theoretical performance is greater than the combined theoretical performance for the HDDs in the node.

The processors in the storage nodes will provide 1x core per HDD in the node (1x core per four disk devices is required). Each node will contain 256GB of DDR4 memory exceeding the 10 GB of memory per HDD requirement.

To fulfil all related requirements, we have selected a 2U form factor server: the Supermicro SuperStorage 6029P-E1CR24L. This is a dual-socket server with 24 Hot-swap 3.5" SAS3/SATA3 drive bays (+2 hot-swap 2.5" SATA3 drive bays at rear).
It provides the necessary connectivity (incl. IPMI over LAN) and redundancy with multiple fans and Titanium Level redundant power supplies.

Figure 41. Object Storage Node - LCST.

Main specifications:

- Dual Socket P (LGA 3647) - Intel Xeon Scalable Processors, 3 UPI up to 10.4GT/s
- 24 DIMM slots
- SAS3 (12Gbps) via Broadcom 3008

- SATA3 (6Gbps); RAID 0, 1, 5, 10
- 2 PCI-E 3.0 x16 slots
- 1 PCI-E 3.0 x8 slot
- 24 Hot-swap 3.5" SAS3/SATA3 drive bays
- 2 Hot-swap 2.5" SATA3 drive bays (rear)
- 2 Onboard NVMe M.2 (optional)
- IPMI 2.0 with virtual media over LAN and KVM-over-LAN support
- 1 RJ45 Dedicated IPMI LAN port
- 1600W Redundant Power Supplies with PMBus

To reach the required 18TB usable capacity of the proposed LCST system, our solution includes 61x of those OSD nodes.

All 24 drive bays of each LCST storage node are populated with 16 TB HDDs. Each server therefore contains 384 TB hard disk drive capacity.

Partition	# HDDs per Node	HDD specifications
LCST Object Storage Nodes	24	16 TB 3.5" SATA3 7.2K RPM 4Kn Seagate HDD-T16T-ST16000NM008G

Hence, the total LCST storage RAW capacity is:

Partition	# Object Storage Nodes	Total HDD Raw Capacity
LCST Object Storage Nodes	61	24.4 PB

Assuming an 8+3 scheme for the erasure encoded (EC) pools we expect this configuration to provide a minimum of 18PB of usable storage space.

Considering a complete LCST node would fail, the maximum HDD capacity loss would be 384TB (24x 16 TB HDDs).

Each Object Storage Node of the proposed LCST partition is also equipped with 2x NVMe SSDs providing a 12.8TB raw NVMe capacity in total per OSD node.

Partition	# SSDs per Node	SSD specifications	% of the capacity of the HDDs in the system
LCST Object Storage Nodes	2	6.4TB NVMe PCIe3.1 3D TLC 2.5" 3DWPD Intel DC P4610	3.33 %

The embedded NVMe drives provide 6400 MB/s sequential read and write (2x 3200 MB/s), which is over the 24 HDDs combined maximum theoretical performance of 6264 MB/s (24x 261 MB/s).

Finally, to connect the Object Storage nodes to the CEPH internal network, each node will be equipped with one 25GbE dual port card:

Partition	# physical connectors per Node	Type of Network Adapter
LCST Object Storage Nodes	2	25GbE DP Mellanox ConnectX-4 Lx EN (25 Gbit/s Ethernet - SFP28)

Virtualization nodes as CEPH Gateway Nodes

The proposed reference architecture for the LCST solution uses 8x nodes (from the Virtualization nodes pool) configured as RADOS gateways. These nodes will serve as the internet gateway to the object storage, and the gateways to the object storage from the compute part of the VEGA system.

The 8x Object/RADOS Gateways taken from the pool of Virtualization nodes are built with the BullSequana X430-A5 2U form factor chassis hosting 1x 2-socket node.

Partition	# of Nodes	Type of Chassis
LCST Object Gateway Nodes	8 (recommended)	2U – 1x 2-socket node BullSequana X430-A5

Please find below the main characteristics of the proposed BullSequana X430-A5 2U-1 node server:

- Dual AMD EPYC™ 7002 series processor family
- 8-Channel RDIMM/LRDIMM DDR4 per processor, 32 x DIMMs
- 2 x 1Gb/s LAN ports (Intel® I350-AM2)
- 1 x Dedicated management port
- Onboard 12Gb/s SAS expander

- 24 x 2.5" SATA/SAS hot-swappable HDD/SSD bays in front side
- 2 x 2.5" SATA/SAS hot-swappable HDD/SSD bays in rear side
- Ultra-Fast M.2 with PCIe Gen3 x4 interface
- x PCIe Gen4 x16 and x8 expansion slots
- 1 x OCP 3.0 Gen4 x16 mezzanine slot
- 1 x OCP 2.0 Gen3 x8 mezzanine slot
- Aspeed® AST2500 remote management controller
- 1600W 80 PLUS Platinum redundant power supply

It provides the necessary connectivity (incl. IPMI over LAN) and redundancy with multiple fans and Platinum Level redundant power supplies.

Figure 42. Object Gateway/RADOS Nodes - LCST.

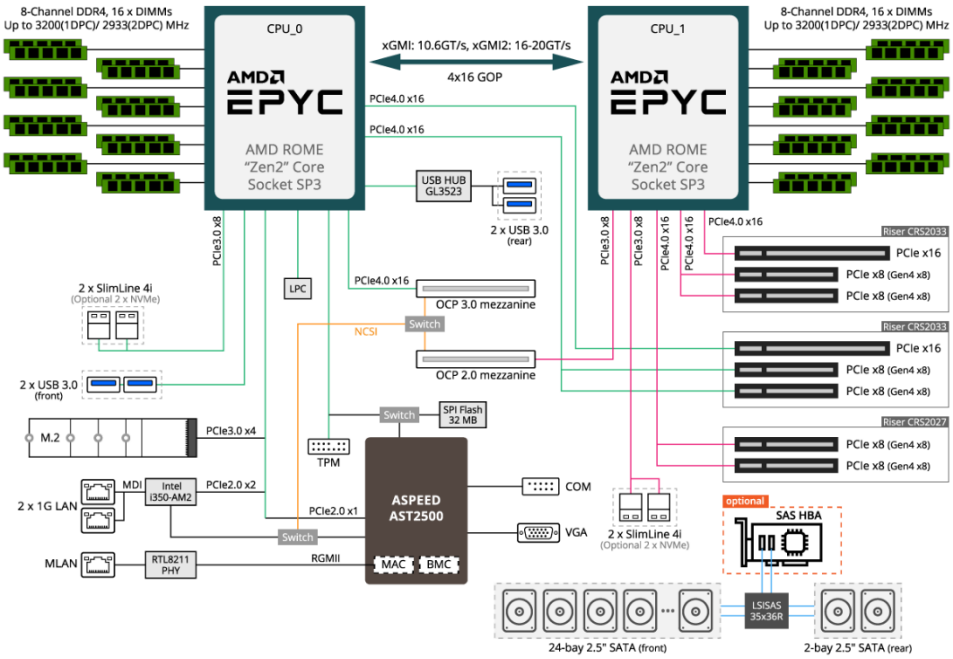


Figure 43: Motherboard design of the BullSequana X430-A5 2U-1N server

Those servers are equipped with 2x AMD Epyc ROME 7502 processors:

Partition	CPU instruction set	# sockets per node	# cores per node
LCST Object Gateway Nodes	x86-64	2	64

Partition	CPU reference and specifications
LCST Object Gateway Nodes	AMD EPYC Rome 7502, 32C (2.5GHz-180W)

And the memory configuration is:

Partition	# memory DIMMs per node	Type of memory DIMM	Memory Capacity per node
LCST Object Gateway Nodes	32	16GB DDR4-3200 ECC RDIMM DR	512 GB

All memory channels are populated with the same speed and capacity memory DIMMs at the highest frequency available for the included CPU model.

Partition	# physical connectors per Node	Type of Network Adapter
LCST Object Gateway Nodes	2	2x Mellanox ConnectX-6 SP HDR EDR card 100Gb QSFP56 PCIe3x16
	2	1x Eth adap 100GbE DP CNX5 PCIe4 x16 QSFP28

These components will provide 4 ports per Object Gateway Node (Virtualization node), 2 ports for the connection to the 100GbE network, 2 for the connection to the InfiniBand HDR network.

Partition	# SSDs per Node	SSD specifications
LCST	4	3.8TB 2.5" 7mm SATA3 SSD TLC 1.3DWPD

Object Gateway Nodes	2*	240GB 2.5" 7mm SATA3 SSD TLC 1.3DWPD (* proposed for the OS)

CEPH Internal Network

The internal network for the object storage system will provide connections to each of the storage and gateway nodes. The internal network is built out of 25GbE /100GbE top of rack switches that uses fanout cables to connect to the storage nodes.

We will use the proposed 100GbE network to connect the virtualization / RADOS gateway nodes to the internal CEPH network. Each RADOS gateway will be connected to two different switches, to ensure that the switches can be serviced without disconnecting the nodes.
By reserving 8x virtualization nodes to RADOS gateway function, with 2x 100GbE connections to the 100GbE switches, we ensure the minimum required bandwidth with the LCST of 200GB/s.

We propose to leverage on the Mellanox SN2010 Series switches to build the LCST Internal Network.
The SN2010 switch is the ideal top of rack (ToR) solution for hyperconverged and storage deployments. Packed with 18 ports of 10/25GbE and 4 split able ports of 40/100GbE, the SN2010 can deliver up to 1.7Tb/s aggregate throughput.

All SpectrumTM-based switches - including the SN2010 - support low-latency line rate traffic for all packet sizes and are ONIE bootable with support for Cumulus Linux, MLNX-OS and other open source operating systems with a broad installed base that drives the world's most innovative data center infrastructures.

SN2010 introduces low latency for 10/25GbE and 100GbE switching, features a robust implementation of data, control and management planes, and offers the most compact form factor and lowest power consumption.

The SN2010 design allows either stand-alone single switch installation, or side-by-side placement of two switches in a single 1RU slot of a 19" rack, delivering high availability to the hosts.

The proposed network architecture building block is the Mellanox SN3700C switch:

Mellanox SN2010 Specifications	
Connectors	18x SFP 25GbE / 4x QSFP 100GbE
Max. 100GbE Ports	4 (split able)
Max. 25GbE Ports	18
Throughput	1.7Tb/s
Packet Per Second	1.26Bpps

Latency	300ns
Serial Ports	1 RJ45
USB Ports	1
Hot-Swap Power Supplies	2 (1+1 redundant)

Figure 44. Mellanox SN2010.

CUMULUS

Switches are preinstalled with Cumulus Linux, taking the Linux user experience from servers to switches and providing a rich routing functionality for large scale applications.

Cumulus Linux embodies native Linux networking. Supercharged versions of the kernel and other networking-related packages encompass the latest industry thinking in networking while retaining compatibility with the full range of software available in Debian.

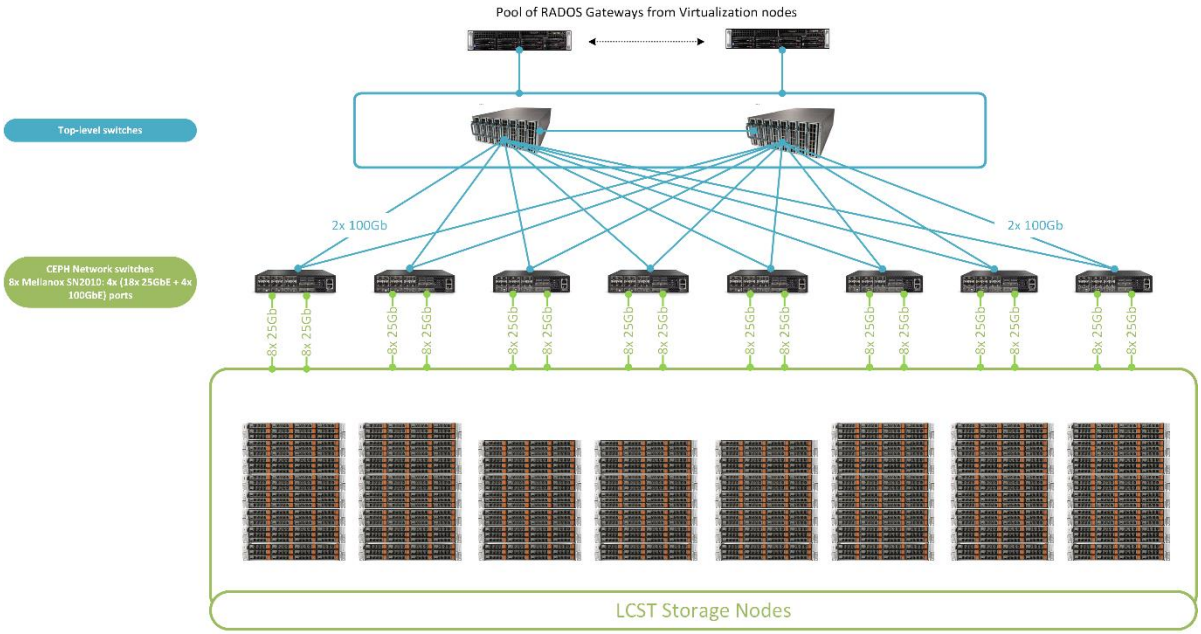


Figure 45. Network to connect VEGA LCST solution to the 100 Gb/s Network.

6.4.2.2 LCS2 [STOR] Large capacity storage tier storage capacity

We understand that this requirement has been relaxed to a desired usable capacity of 18TB. The proposed solution for the LCST solution provides 18TB of usable capacity.

6.4.2.3 LCS3 [STOR] Large capacity storage tier remote block storage access

Remote block storage access is provided using 8x of the virtualization nodes configured as RADOS gateways, the provided 100GbE network and the CEPH internal 25GbE / 100GbE network described above.

6.4.2.4 LCS4 [STOR] Large capacity storage tier I/O performance

All units of throughput depicted in this section are based on powers of ten (1 MB/s = 1 000 000 bytes/s; 1GB/s = 1000 MB/s).

The proposed system can sustain a throughput of 207 GB/s.

The median throughput performance of a single OSD drive is taken as the basis of this calculation.
The median performance of a single drive is considered the median value between its maximum throughput (obtained when accessing the outer most sectors of the drives) and its minimum throughput (when accessing the inner most sectors, which is usually half of the maximum throughput. The maximum throughput for the ST16000NM008G being 260MB/s, the minimum is 130MB/s and median performance is (260+130)/2 = 195MB/s.

The raw throughput provided by the system is then 61 servers * 24 drives * 195 MB/s resulting in 285.48GB/s. However, considering the erasure coding configuration of 8+3, the useful throughput is actually 285.48 * 8 (effective data) / 11 (real written data) = 207.6 GB/s.

The following procedure is proposed to measure this performance from the HPST to the LCST:

- Several HPST clients (determined during the benchmarks preparation) will each map a dedicated RBD (Remote Block Device),
- A local filesystem (ext4 or xfs) will be created on each of those RBDs and mounted on the HPST client nodes,
- A dedicated directory for each client node will be created on the HPST filesystem,
- A set of files will be created in these dedicated directories. This set of files can be replicated for each node and must be big enough to sustain a transfer for a few minutes. It is proposed to replicate a 1GB file the required number of times.

- The benchmark begins: copies are initiated on all the selected HPST client nodes from the set of files in the dedicated HPST directory to the locally mounted filesystem on the RBD
- The throughput is monitored via the "ceph iostat -p 5" command, which will display the read and write global effective throughput of the system. The benchmark is successful when a line showing superior or equal throughput to the commitment is shown.

The same procedure can be used to validate throughput from the LCST to the HPST, by copying files generated on the local filesystem to the dedicated directory on the HPST.

6.4.2.5 LCS5 [STOR] Large capacity storage tier core per disk

The Object Storage nodes are equipped with 2x 24-core processors. Thus, our solution offers more than the minimum core per disk requirement, i.e. 1x core per HDD instead of 1x core per 4x HDD.

Partition	CPU instruction set	# sockets per node	# cores per node
LCST Object Storage Nodes	x86-64	2	24 *(1 core per HDD)*

Partition	CPU reference and specifications
LCST Object Storage Nodes	Intel Xeon Silver 4214R*, 12 cores, 2.4GHz, 100W *: R stands for "Refresh"

6.4.3 High-performance storage tier technical requirements

6.4.3.1 HPS1 [STOR] High-performance storage tier design

DDN EXAscaler

A file system that is down, or one underperforms when running production workloads, can severely impact the overall effectiveness of the compute systems. It is estimated that large national lab scale systems installed in 2015 incurred a $150,000/day cost, including amortized hardware, software, power and cooling, personnel, etc. A National Lab scale system installed in 2020+ is forecast to cost $450,000 to $500,000 a day. If a storage system is unavailable, or has constrained performance, the impact on the whole system usability can result in enormous expense.

What is needed is a solution that is installed, accepted and running in production in 2-4 weeks rather than 4-6 months, one that handles the operational workloads running

currently as well as those that will run five years from now. A system of this complexity is best delivered by a vendor whose employees understand large-scale HPC and the importance of flagship computational laboratories. As the single point of continuity for ExaScale HPC systems, the storage subsystem is of critical importance. DDN feels strongly that our storage and expertise are a perfect match for this critical solution. This confidence is derived from a long history of providing storage systems designed and developed specifically for the HPC market at the largest scales.

DDN is unique in its heritage and capability and focus in at scale storage. DDN platforms have been developed and tried and tested in the very largest and toughest at-scale environments. DDN engineers and DDN systems have been already exposed to the myriad corner cases that arise in at-scale environments which contain 1000s of clients, millions of I/O threads, tens of billions of files, hundreds of storage servers and the full diversity of customer application behavior. It is that experience translated into solid filesystem engineering and methodologies and expertise that bring whole systems to production faster and deliver a service most effective for user workloads.

Only DDN is dedicated to building, installing and supporting at scale storage systems in this volume and at this size, and our business lives on our reputation. DDN is a company of nearly 1000 staff, running profitably and has the stability you need in a partner for the lifetime of your data.

With the richness and variety of current data generation and analysis technologies, Enterprises and Research Organizations alike have a greater opportunity than ever to create value from their data. The combined software ecosystem across the fields of Big Data, Large Scale Numerical computing (HPC), Analytics and AI brings unprecedented capability into the hands of those with large data sets. However, the associated challenges in data management are large as well. The diversity and sheer volume of these new data intensive workloads accentuates the need for a new way to manage data at scale; stronger and more comprehensive data security, better reliability and consistent high performance, but also a storage environment with the capability and flexibility to cater for cloud, multiple protocols, objects and files and integrates into a comprehensive data lifecycle approach. DDN’s EXAScaler meets these challenges and benefits from a core that has been hardened in the most demanding environments.

EXAScaler is highly scalable software for data storage and analysis and management, developed by DDN. EXAScaler runs in DDN appliances and in the cloud. At its core, is the fastest, most scalable open parallel file system, Lustre. Lustre itself is the most popular filesystem for scale-out computing and has been proven and hardened in the most demanding HPC environments for over 15 years. EXAScaler and Lustre are developed by a very active, dedicated and talented team, most of whom now work at DDN. DDN supports the open source community with Lustre through our Whamcloud organization and develops EXAScaler which builds on Lustre’s innate performance and scalability with a host of additional major features to greatly simplify deployment, storage and management at scale with differentiated integrations into emerging workloads. EXAScaler really helps when you need full spectrum performance across

workloads from deep learning through to Big Data and analytics. With special connections into NVIDIA DGX systems and other NVIDIA GPU based architectures, environments running on AWS, Azure and Google Clouds EXAScaler supports complete full range of modern workloads.

Major EXAScaler Themes

- Extreme Scalability and Performance: A highly parallel architecture for low latency workloads and high bandwidth applications alike
- Data Aware Intelligence: EXAScaler grows with your data, allowing limitless scale with the good economics by intelligently management your data across tiers
- Data Security and Integrity: Secure Multi-Tenancy, Encryption, End-to-End data protection, Replication all from a filesystem hardened at the largest scales
- Flexible access from any application, any architecture, anywhere: Integration with cloud, GPU, AI frameworks, containers, file and object

Simplicity is difficult to deliver at scale. At DDN, we've heard many troublesome tales of data storage systems that don't cope when presented with the onslaught of corner cases which are created when clients are in the tens of thousands, data is over 100PB and IO rates are measured in Terabytes per second. Much of the hard work of delivering strong RAS (Reliability, Availability, Serviceability) is down to our engineers' experience in developing and testing at the largest scales.

DDN EXAScaler appliances bring all the sophistication required to deliver fast, secure, stable, and simple storage at scale. EXAScaler can be delivered as all-NVMe and Flash-Hybrid options starting as small as 2 rack units and growing up to solutions that provide over 10PB per rack. No need for the complexities and headaches of Ceph and Big Data solutions that required multiple data copies, and deep domain knowledge to keep stable and performant. No complicated user experience in understanding your data architecture. No limitations on performance. Full Flexibility in building a data architecture that matches your business which grows and adapts alongside your data and workloads.

Proposed solution for IZUM

Our proposition contains storage components to build an EXAScaler based storage solution. The components are flash-based and are combined such that a storage solution with an extremely high performance is formed. The components achieve maximum performance in a minimal footprint.

This document only lists high level performance numbers. More detailed performance numbers are provided in a separate document.

The provided Solutions are summarized in the following Executive Summary.

Flash - OST	
FlashBased Building Block (FBB)	ES400NVX
# of FlashBased Building Blocks	10
Pool Configuration per Building Block	2 x 10/0 + 1 HS
Flash Device used for OST and MDT	7.68TB NVMe
FlashBased capacity – OST and MDT	~1200 TByte (7.68TB)
Flash based performance	
FlashBased performance (MAX read)	>480 GB/s
FlashBased performance (MAX write)	~400 GB/s
Physical	
#Racks	1
Power, Nominal	~18.0 kW
Power, Idle	~14.5 kW
Heat, Nominal	~61500 BTU/hr
Heat, Idle	~50000 BTU/hr
Max. Airflow	~1500 CFM

The main offering for IZUM achieves the required 200 GB/s from the 800TB flash partition. The useable capacity from disk drives is 1.2 PByte.

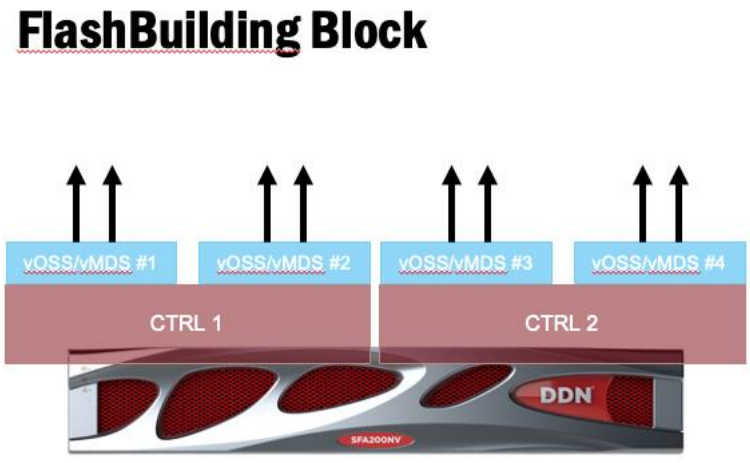
IZUM	
Filesystem	Lustre 2.12.x
Fabric	Mellanox HDR100
# Fabric Uplinks required	80 x InfiniBand EDR/HDR100
# Ethernet Ports (GE) required	20 x GigE HostAccess 10 x GigE BMC Network
# Building blocks	10 x ES400NVX
formattable Capacity after RAID	~1200 TByte (7.68TB)
Aggregated Filesystem Performance throughput (large block sequential IOR)	>480 GB/s read (Flash Based) ~400 GB/s write (Flash Based)

Power Consumption	
Power, Nominal	~18.0 kW
Power, Idle	~14.5 kW
Physical Specs	
Weight of Components	~400kg (rack and cables not included)
#Racks	Not included

ES400NVX Building Block

The ES400NVX building block consists of:

- The ES400NVX base enclosure with 1+1 redundant storage controllers and fully redundant IO-paths to the storage devices
- Redundant Fans and redundant Power Supplies
- 21 x 7.68TB NVMe devices (DWPD=1)
  - 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 itself is fully redundant and has no single point of failure.

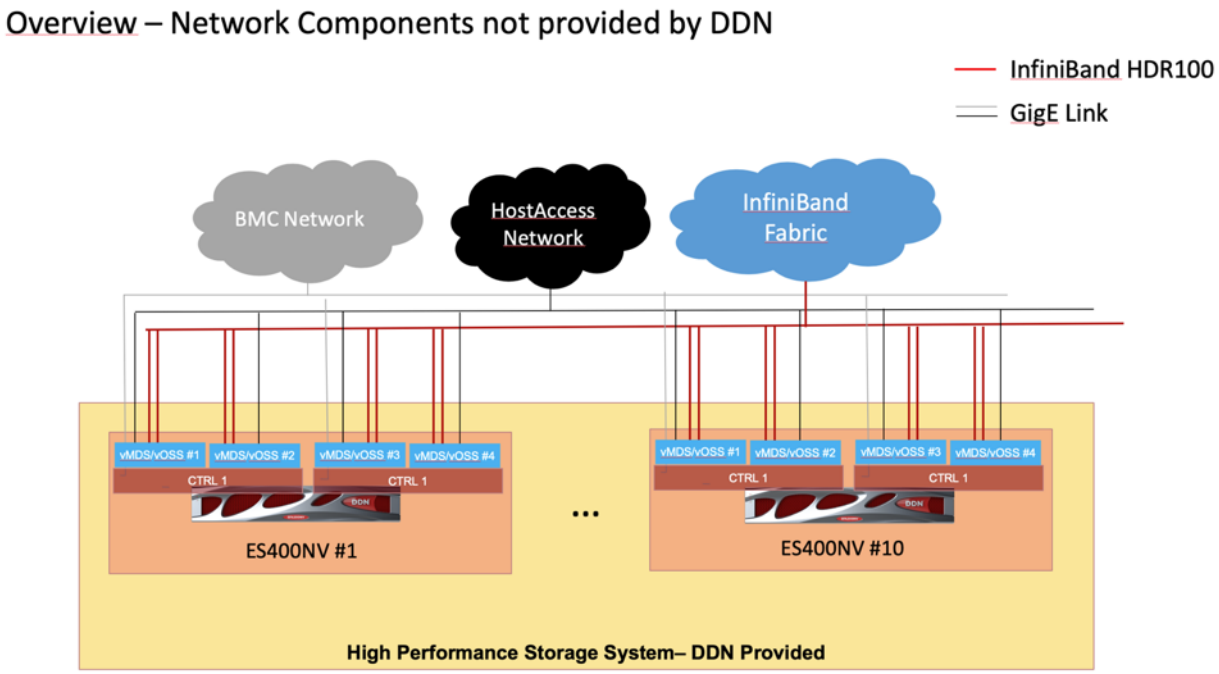
The building block provides 120 TB formattable capacity. This capacity will be split into:

- 4 TB for 4 x MDT (each 1TB in size)
- 116 TB for 4 x OST (each 29 TB in size)

Ten of these building blocks form the complete solution with:

~40TB of capacity for MetaData
~1160 TB of capacity for Data

After applying filesystem overhead, the useable capacity will be about 1.1 PByte for data and a maximum of more than 5 billion inodes in the system.



6.4.3.2 HPS2 [STOR] High-performance storage tier storage capacity

It is our understanding from the Q&A, that the 3PB usable capacity requirement has been relaxed to 1PB.

The proposed solution for HPST contains 10x ES400NVX storage building blocks providing **> 1PByte of useable capacity**. The exact capacity calculation can be seen in a separate document with in-depth details about DDN EXAScaler solutions.

6.4.3.3 HPS3 [STOR] High-performance storage tier I/O performance

The offered solution is meeting the requirements. The detailed description of the performance of the offered solution and how it is achieved is provided in Chapter 10 of this document.

6.5 Login partition

6.5.1 Login partition hardware requirements

The proposed solution for IZUM's VEGA system includes a set of 8x login nodes aimed to be used for software development and code compilation, data management including transfer in and out of the system, batch job submission and pre- and post-processing activities.

The proposed Login partition consist of several Login nodes decomposed as 4x Login nodes dedicated (but not restricted) to the CPU compute partition, and 4x Login nodes dedicated (but not restricted) to the GPU compute partition.

Thus, the 4x "CPU" Login nodes will be equipped with the same processor as the compute nodes of the CPU partition, and the 4x "GPU" Login nodes will be equipped with the same processor (with the same number of sockets) and exact GPU as those that equip the compute nodes of the "GPU" partition.

To build the two types of Login nodes, we have selected two servers from our BullSequana X400 series.

"CPU" Login nodes

The 4x CPU Login nodes of the proposed solution are built from Atos BullSequana X430-A5 server series.

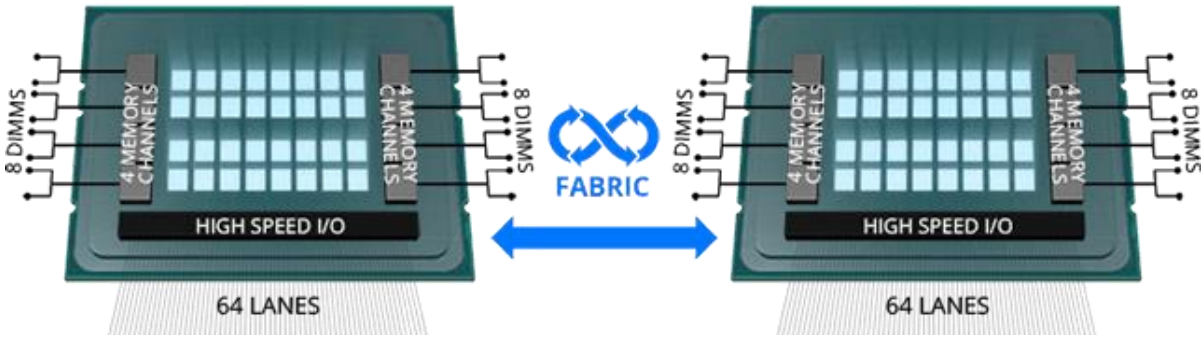
The BullSequana X430-A5 (2U-1N 2S) is a 2U rack-mounted 2-socket server. It is ideally suited as a service node, since its advanced connectivity features, extended storage options and redundancy features guarantee efficient and reliable cluster administration services.

Figure 46: BullSequana X430-A5 2U-1N 2S Server

Enhanced connectivity

The BulllSequana X430-A5 server is equipped with two AMD® EPYC® processors Scalable family.

Both single- and dual-socket CPUs come equipped with 128x PCIe lanes, providing extensive I/O options. This is a huge improvement in I/O vs. incumbent x86 offerings. Increased I/O allows for flexibility in extending your storage, networking or compute to a level previously unattainable with standard server motherboards.
In addition, the integrated nature of the processor – designed as a system-on-chip (SoC) - means that there is no need for a separate PCH, and less additional switches needed for expansion lanes.



The server includes several PCIe slots:

- 1 x PCIe x8 slot (Gen4 x8), Low profile half-length
- 1 x PCIe x8 slot (Gen4 x8), Low profile half-length
- 1 x PCIe x16 slot (Gen4 x16), FHHL
- 1 x PCIe x8 slot (Gen4 x8), FHHL
- 1 x PCIe x8 slot (Gen4 x8), FHHL
- 1 x PCIe x16 slot (Gen4 x16), FHHL
- 1 x PCIe x8 slot (Gen4 x8), FHHL
- 1 x PCIe x8 slot (Gen4 x8), FHHL
- 1 x OCP 3.0 mezzanine slot with PCIe Gen4 x16 bandwidth from CPU_0 Supported NCSI function
- 1 x OCP 2.0 mezzanine slot with PCIe Gen3 x8 bandwidth (Type1, P1, P2) Supported NCSI function

Enhanced memory capacity and throughput

The X430-A5 2U-1N-2S can support up to 32x DIMM slots.

It supports 8 channels 3200 MHz (1 DIMM per channel), 2933MHz (2 DIMM per channel) RDIMM, LRDIMM, 3DS RDIMM, NVDIMM Max. size 4TB/socket

High internal storage capacity

Depending on the option, the server can support:

- Front 24x 2.5" Hot-swap disk(4x U.2 with NVMe kit); Rear 2x 2.5" Hot-swap disk
- Front 24x 2.5" Hot-swap U.2 NVMe SSD only; Rear 2x 2.5" Hot-swap disk

Maximized reliability

Redundant Power Supply Unit with airflow and thermal control.

Detailed configuration of the CPU Login Nodes:

CPU Login Nodes configuration	
Server Model	BullSequana X430-A5 2U-1N 2Socket
Quantity	4
CPU model and quantity per node	2x AMD EPYC ROME 7H12 (64 cores @ 2.6GHZ – 280W)
Memory Configuration	16x 16GB DDR4 @ 3200MT/s – 256GB DDR4 total
Disk Configuration	2x 7.6TB U.2 15mm PCIe3 x4 TLC 3.1DWPD (5Y)
RAID Configuration	Only software RAID supported on NVMe storage
Connectivity	1x 100GbE Dual Port CNX5 PCIe4 x16 QSFP28 Adapter
	1x ConnectX-6 SP HDR EDR card 100Gb QSFP56 PCIe3x16
Ethernet Network	Dual 1GbE port (Intel I350) 1x Management LAN 10/100/1G
Power Supply	Redundant 1600W 80+ Platinum

"GPU" Login nodes

The 4x GPU Login nodes of the proposed solution are built from Atos BullSequana X410-A5 server series.

The BullSequana X410-A5 server is a 2U rack-mounted 2-socket server and a perfect fit for visualization, pre- / post- processing and GPU node purposes with up to 8x GPU dedicated PCIe slots and, 2x AMD Epyc Rome processors, 16x DIMM slots, and up to 4x 2.5" NVMe hot-swap HDD/SSD bays.

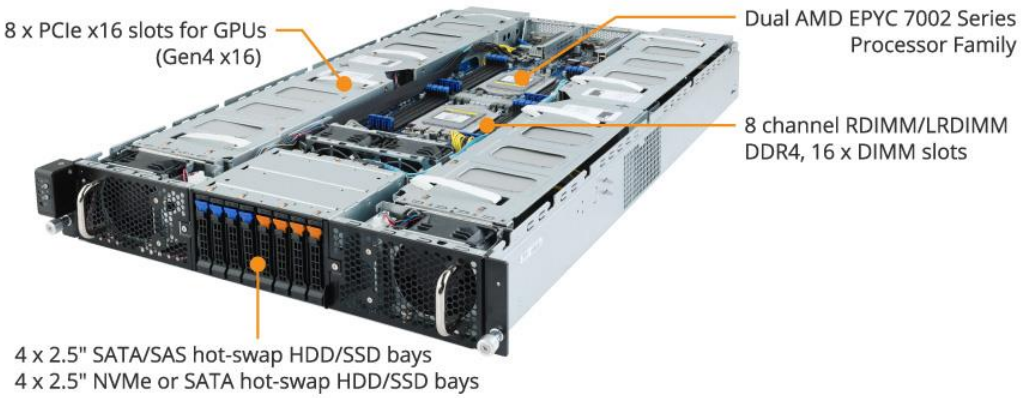


Figure 47: BullSequana X410-A5 2U-1N 2S Server

The BullSequana X410-A5 server series characteristics are:

- Supports up to 8 x double slot GPU cards
- Dual AMD EPYC™ 7002 series processor family
- 8-Channel RDIMM/LRDIMM DDR4 per processor, total 16 x DIMMs
- 2 x 10Gb/s BASE-T LAN ports (Intel® X550-AT2 controller)
- 1 x Dedicated management port
- 8 x 2.5" hot-swappable HDD/SSD bays
- 8 x PCIe Gen4 expansion slots for GPUs
- 2 x PCIe Gen4 x16 half-length low-profile slots for add-on cards
- Aspeed® AST2500 remote management controller
- 2+0 2200W 80 PLUS Platinum power supply

Detailed configuration of the GPU Login Nodes:

CPU Login Nodes configuration	
Server Model	BullSequana X410-A5 2U-1N 2Socket
Quantity	1
CPU model and quantity per node	2x AMD EPYC ROME 7452 (32 cores @ 2.35GHZ – 155W)
GPU Configuration	1x NVIDIA Ampere A100 PCIe GPU
Memory Configuration	16x 16GB DDR4 @ 3200MT/s – 256GB DDR4 total

Disk Configuration	2x 7.6TB U.2 15mm PCIe3 x4 TLC 3.1DWPD (5Y)
RAID Configuration	Only software RAID supported on NVMe storage
Connectivity	1x 100GbE Dual Port CNX5 PCIe4 x16 QSFP28 Adapter
	1x ConnectX-6 SP HDR EDR card 100Gb QSFP56 PCIe3x16
Ethernet Network	Dual 1GbE port (Intel I350) 1x Management LAN 10/100/1G
Power Supply	Redundant 2200W 80+ Platinum

6.5.1.1 LHW1 [LVSP] Amount of Login nodes

Our proposition includes **8x login nodes**:

- 4x "CPU" login nodes
- 4x "GPU" login nodes

6.5.1.2 LHW2 [LVSP] Login node processor and graphics processor requirements

The 4x "CPU" login nodes are built with the same processor SKU and quantity as the compute nodes of the CPU partition.

The 4x "GPU" login nodes are built with the same processor architecture (different SKU though) and quantity and host one identical GPU model (extendable up to 8x) as those that equip the compute nodes of the GPU partition.

Hence, the proposed Login partition will allow CPU and GPU binary compatibility with the CPU and GPU compute partitions.

6.5.1.3 LHW3 [LVSP] Login node memory capacity

CPU login nodes are equipped with 2x AMD EPYC ROME 64-core processors and GPU login nodes with 2x AMD EPYC ROME 32-core processors. Each node of the login partition is equipped with 16x 16GB DDR4 3200MT/s memory DIMMs, complying with both requirements.

6.5.1.4 LHW4 [LVSP] Login nodes local storage

Each node of the login partition is equipped with 2x 7.6TB U.2 15mm PCIe3 x4 TLC 3.1DWPD (5Y) NVMe storage drives. No hardware RAID is supported on NVMe storage on BullSequana X400-A5 server series, but a software RAID 1 solution can be implemented for mirroring operating system and node-local temporary scratch space.

6.5.1.5 LHW5 [LVSP] Login node Ethernet connectivity

Each node of the login partition is equipped with 2x 100GbE ports as a single 100GbE Dual Port CNX5 PCIe4 x16 QSFP28 Adapter.

6.5.1.6 LHW6 [LVSP] Login node High-speed interconnect connectivity

Each node of the login partition is equipped with 1x ConnectX-6 SP HDR EDR card 100Gb QSFP56 PCIe3x16 adapter for connection with the High-Speed interconnect network.

6.5.1.7 LHW7 [LVSP] Login nodes redundant power supply and location

Each node of the login partition is equipped with redundant 80+ Platinum power supplies. Login nodes will tolerate one power supply unit failing and will be connected to wo independent power sources. The login nodes are mounting suitable for racks and will be installed in the RIVR2 room.

6.5.2 Login partition software requirements

6.5.2.1 LSW1 [LVSP] Login nodes software compatibility with Compute/CPU nodes

The design of the login nodes allows native execution of the Compute/GPU node applications on the login nodes and vice versa.

6.6 Virtualisation and service partition

6.6.1 Virtualisation and service partition hardware requirements

The proposed solution for IZUM's VEGA system includes a set of virtualization/service nodes aimed to be used for general purpose services, administrative tasks and tasks that do not correspond to the compute, GPU and login partitions. This partition consists of 30x Virtualisation nodes.

To build the Virtualization nodes, we have selected Atos BullSequana X430-A5 server series.

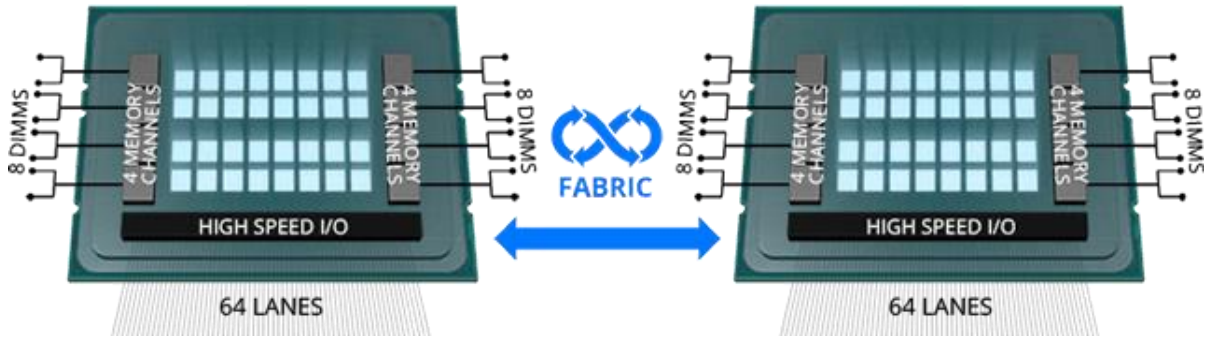
The BullSequana X430-A5 (2U-1N 2S) is a 2U rack-mounted 2-socket server. It is ideally suited as a service node, since its advanced connectivity features, extended storage options and redundancy features guarantee efficient and reliable cluster administration services.

Figure 48: BullSequana X430-A5 2U-1N 2S Server

Enhanced connectivity

The BulllSequana X430-A5 server is equipped with two AMD® EPYC® processors Scalable family.

Both single- and dual-socket CPUs come equipped with 128x PCIe lanes, providing extensive I/O options. This is a huge improvement in I/O vs. incumbent x86 offerings. Increased I/O allows for flexibility in extending your storage, networking or compute to a level previously unattainable with standard server motherboards.
In addition, the integrated nature of the processor – designed as a system-on-chip (SoC) - means that there is no need for a separate PCH, and less additional switches needed for expansion lanes.



The server includes several PCIe slots:

- 1 x PCIe x8 slot (Gen4 x8), Low profile half-length
- 1 x PCIe x8 slot (Gen4 x8), Low profile half-length
- 1 x PCIe x16 slot (Gen4 x16), FHHL
- 1 x PCIe x8 slot (Gen4 x8), FHHL
- 1 x PCIe x8 slot (Gen4 x8), FHHL
- 1 x PCIe x16 slot (Gen4 x16), FHHL

- 1 x PCIe x8 slot (Gen4 x8), FHHL
- 1 x PCIe x8 slot (Gen4 x8), FHHL
- 1 x OCP 3.0 mezzanine slot with PCIe Gen4 x16 bandwidth from CPU_0 Supported NCSI function
- 1 x OCP 2.0 mezzanine slot with PCIe Gen3 x8 bandwidth (Type1, P1, P2) Supported NCSI function

Enhanced memory capacity and throughput

The X430-A5 2U-1N-2S can support up to 32x DIMM slots.

It supports 8 channels 3200 MHz (1 DIMM per channel), 2933MHz (2 DIMM per channel) RDIMM, LRDIMM, 3DS RDIMM, NVDIMM Max. size 4TB/socket

High internal storage capacity

Depending on the option, the server can support:
- Front 24x 2.5" Hot-swap disk(4x U.2 with NVMe kit); Rear 2x 2.5" Hot-swap disk
- Front 24x 2.5" Hot-swap U.2 NVMe SSD only; Rear 2x 2.5" Hot-swap disk

Maximized reliability

Redundant Power Supply Unit with airflow and thermal control.

Detailed configuration of the Virtualization Nodes:

Virtualization Nodes configuration	
Server Model	BullSequana X430-A5 2U-1N 2Socket
Quantity	30
CPU model and quantity per node	2x AMD EPYC ROME 7502 (32 cores @ 2.5GHZ – 180W)
Memory Configuration	32x 16GB DDR4 @ 3200MT/s – 512GB DDR4 total
Disk Configuration	2x 7.6TB U.2 15mm PCIe3 x4 TLC 3.1DWPD (5Y)
RAID Configuration	Only software RAID supported on NVMe storage
Connectivity	1x 100GbE Dual Port CNX5 PCIe4 x16 QSFP28 Adapter

	1x ConnectX-6 SP HDR EDR card 100Gb QSFP56 PCIe3x16
Ethernet Network	Dual 1GbE port (Intel I350) 1x Management LAN 10/100/1G
Power Supply	Redundant 1600W 80+ Platinum

6.6.1.1 VHW1 [LVSP] Amount of Virtualisation nodes

Our solution includes 30x Virtualization nodes built on our BullSequana X430-A5 2U-1N 2-socket server series. 8x of those Virtualization nodes are intended to be used as RADOS gateways for accessing the CEPH LCST solution. Configuration of the Virtualization/RADOS Gateway nodes does not change depending of the aimed usage.

6.6.1.2 VHW2 [LVSP] Virtualisation node processor architecture

Processor architecture of the proposed virtualization nodes is described in section 6.2. It is compliant with this requirement.

6.6.1.3 VHW3 [LVSP] Virtualisation node processor requirements

The Virtualization nodes are equipped with 2x AMD EPYC ROME 7502 32-core @ 2.6GHz processors.

Partition	CPU instruction set	# sockets per node	# cores per socket
Virtualization Nodes	x86-64	2	32

Partition	CPU reference and specifications
Virtualization Nodes	AMD EPYC ROME 7502: 32 cores – 2.6GHz – 180W

6.6.1.4 VHW4 [LVSP] Virtualisation node memory capacity

Virtualization nodes are equipped with 2x AMD EPYC ROME 32-core processors. Each node of the virtualization partition is equipped with 32x 16GB DDR4 3200MT/s memory DIMMs (total of 512GB), complying with both requirements.

6.6.1.5 VHW5 [LVSP] Virtualisation node local storage

Each node of the Virtualization partition is equipped with 2x 7.6TB U.2 15mm PCIe3 x4 TLC 3.1DWPD (5Y) NVMe storage drives. No hardware RAID is supported on NVMe storage on BullSequana X400-A5 server series, but a software RAID 1 solution can be implemented for mirroring.

6.6.1.6 VHW6 [LVSP] Virtualisation node Ethernet connectivity

Each node of the Virtualization partition is equipped with 2x 100GbE ports as a single 100GbE Dual Port CNX5 PCIe4 x16 QSFP28 Adapter.

6.6.1.7 VHW7 [LVSP] Virtualisation node High-speed interconnect connectivity

Each node of the Virtualization partition is equipped with 1x ConnectX-6 SP HDR EDR card 100Gb QSFP56 PCIe3x16 adapter for connection with the High-Speed interconnect network.

6.6.1.8 VHW8 [LVSP] Virtualisation node redundant power supply and location

Each node of the Virtualization partition is equipped with redundant 80+ Platinum power supplies. Virtualization nodes will tolerate one power supply unit failing and will be connected to wo independent power sources. The Virtualization nodes are mounting suitable for racks and will be installed in the RIVR2 room.

6.6.2 Virtualisation and service partition software requirements

6.6.2.1 VSW1 [LVSP] Virtualisation support

Our proposal includes Red Hat Virtualization software stack on the virtualization nodes, including licences, subscriptions and support for the whole maintenance period of the procured system.

Red Hat Virtualization is an enterprise-grade virtualization platform built on Red Hat Enterprise Linux. Virtualization allows users to easily provision new virtual servers and workstations and provides more efficient use of physical server resources. With Red Hat Virtualization, you can manage your entire virtual infrastructure - including hosts, virtual machines, networks, storage, and users - from a centralized graphical user interface or REST API.

Red Hat Virtualization Key Components

The following table summarizes some of the key features of Red Hat Virtualization:

Red Hat Virtualization Key Components	
Red Hat Virtualization Manager	A service that provides a graphical user interface and a REST API to manage the resources in the environment. The Manager is installed on a physical or virtual machine running Red Hat Enterprise Linux.
Hosts	Red Hat Enterprise Linux hosts (RHEL hosts) and Red Hat Virtualization Hosts (image-based hypervisors) are the two supported types of host. Hosts use Kernel-based Virtual Machine (KVM) technology and provide resources used to run virtual machines.
Shared Storage	A storage service is used to store the data associated with virtual machines.
Data Warehouse	A service that collects configuration information and statistical data from the Manager.

For detailed technical information about Red Hat Virtualization, see the *Technical Reference* at https://access.redhat.com/documentation/en-us/red_hat_virtualization/4.3/html/technical_reference/

6.7 High-speed interconnect

6.7.1 HSI1 [NETW] High-speed interconnect network participants

We confirm. All Compute and Login, Virtualization, HCST and Gateway nodes of the proposed will be connected, with one or multiple ports, to the high-speed Interconnect. We use the free ports on the Level-1 fabric switches of our BullSequana XH2000 platform to connect, in addition of the compute nodes, all service and IO nodes.

6.7.2 HSI2 [NETW] High-speed interconnect performance

We comply with this requirement. As describes above, compute nodes of the CPU partition are equipped with 1x HDR100 port, whereas compute nodes of the GPU partitions are equipped with 2x HDR200 ports. Hence, the system design balances bandwidth with respect to the node performance, GPU node requiring more bandwidth than CPU nodes to take full advantage of the 4x embedded NVIDIA A100 GPUs.

6.7.3 HSI3 [NETW] High-speed interconnect topology

We comply with this requirement. As describes above, the system is divided into cells. Each CPU cell has a two-level non-blocking fat-tree topology. GPU cell has a small

blocking factor or 20-to-16 (1.25:1). All cells are interconnected using an all-2-all topology (18 downlinks and 12 uplinks per L2 switch), so that the full high-speed interconnect network is built on an optimized dragonfly+ topology.

6.7.4 HSI4 [NETW] High-speed interconnect bisection bandwidth

The High-speed interconnect provide a very high bisection bandwidth due to the all-to-all connections of the spine switches part of the Dragonfly topology. The current implementation provides 192x HDR (200 Gbps) links which give more than 4.8 TB/s of bisection bandwidth.

6.7.5 HSI5 [NETW] High-speed interconnect uniform performance

The proposed solution includes two types of nodes with a different number of InfiniBand cards:

- CPU Compute nodes with 1x HDR100 card per node
- GPU Compute nodes with 2x HDR cards per node

Regarding the bandwidth, no matter which core is used, the bandwidth will be the same. All communications will go through a card that will run at full speed.

For the latency, as GPU nodes have two cards connected to a different CPU, the latency will be homogeneous within the nodes and between two GPU nodes.
However, with CPU node, it will not be the case.

A PCIe card needs to be attached to a specific CPU. With only one card per node, a core located in the CPU attached to the IB card will have a lower latency than a core located in the other CPU.

We expect this latency to be minimal compared to the switch latency and it will not impact your performance.

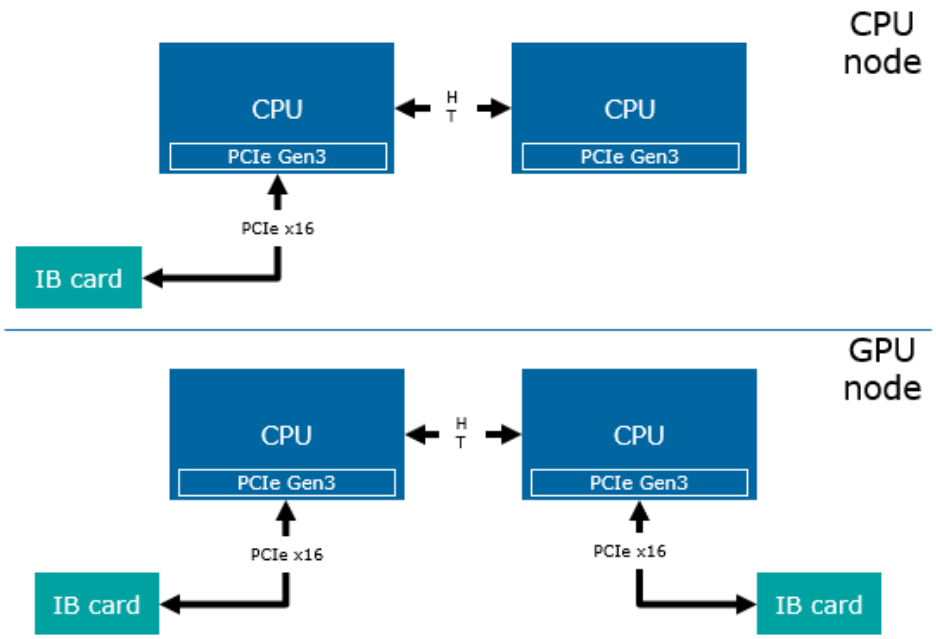


Figure 49: Connection between IB and CPU

6.7.6 HSI6 [NETW] High-speed interconnect IPoX stack

We confirm. The proposed High-Speed Interconnect can use the IP-over-IB protocol and thus sending IP packages via the High-Speed Interconnect network.

6.7.7 HSI7 [NETW] High-speed interconnect monitoring capabilities

Management and monitoring of the High-Speed Interconnect fabric is performed by IBMS. IBMS is a scalable monitoring tool for InfiniBand networks. It is validated on up to 10 000 IB ports, and on a variety of InfiniBand Hardware. IBMS Provides:

- Error detection, e.g. cabling error, misconfigurations, failing cables,
- Firmware update,
- Event tracking and visualization,
- Remote fabric monitoring.

Key features:

- SQL Data repository storing Hardware description, Topology, Versions, Counters, Events, InfiniBand & SNMP traps;
- Data display via ibview, a perl front-end providing a complete view of the fabric, reports on errors, performance & diagnostic counters, versions, InfiniBand traps, vital product data, temperature, fan speeds and more;

- Time series graphs via IBMS-Graphite;
- Event forwarding to external monitoring systems,
- Multi-level output plugins e.g. PostgreSQL, rrdtool, OpenTSDB, InfluxDB, carbon, csv, collectd

Part of IBMS are cli commands, such as Ibview. Ibview is a perl frontend to the IBMS database. It provides a complete view of the fabric (topology, equipment description with localization, link availability by cross-checking with the cluster database, reports on errors, performance & diagnostic counters, firmware & switch software versions, IB traps, vpd, temperatures, fan speeds etc.

Another tool for graphic analysis of a computing jobs is ibms-visualizer. This tool can:

- collect topology details from IBMS, job details form Slurm and metrics data from rrdcached daemon
- create graph entries in graphite-web GUI
- create XLS and CSV reports
- be used standalone and in Slurm epilogue

ibms-visualizer can be used with different graphing backends (Grafana, Graphite-web) and metrics storage backends (rrdtool's rrdcached, OpenTSDB, InfluxDB, Graphite-web acting as a proxy).

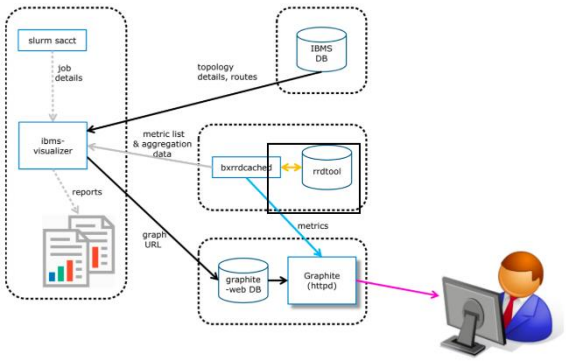


Figure 50: ibms-visualizer

6.7.8 HSI8 [NETW] High-speed interconnect mechanisms for congestion avoidance

The new Fully Progressive Adaptive Routing (FPAR) is implemented in the DF+ topology and fully supported by Mellanox and provides a very efficient routing for which decisions are evaluated in every router to minimize congestion. It is also very scalable to a very large number of nodes and provide full b-section bandwidth. It is based on the concept of dividing nodes to groups while the groups are connected in an all-to-all fashion. A

group is based on a multi-layer Fat tree. The nodes are connected to the so-called leaf switches which are then connected to a second layer and a possible third layer, and so on. The top layer is called the spine switches which are used for the all-to-all connections to form a DragonFly+ topology.
The InfiniBand spec defines 16 levels of QoS, with Mellanox's hardware supporting eight levels, as well as an additional administrative QoS level. The QoS is supported by the network hardware, with the ability to administratively define priority, resource availability and communication traits, such as support for adaptive routing, at each of these levels. This mechanism supports prioritizing the traffic based on Virtual Lanes (VL). Mellanox's hardware supports eight virtual lanes, plus VL 15, and 16 SLs.
Mellanox also supports an advanced arbitration mechanism in both the IB switches and NICs which provides a round robin deficit arbitration with rate limiting and flexible prioritization between traffic classes. The advanced arbitration also provides lower jitter for the arbitrated flows.
Mellanox InfiniBand also has the capability to re-route paths of over-utilised links to best use the available bandwidth. Adaptive Routing (AR) enables the switch to select the output port based on the port's load. AR supports two routing modes:

- Free AR: No constraints on output port selection
- Bounded AR: The switch does not change the output port during the same transmission burst. This mode minimises the appearance of out-of-order packets.

The Adaptive Routing Manager enables and configures Adaptive Routing mechanism on fabric switches. It scans all the fabric switches, identifies which switches support Adaptive Routing, and configures the AR functionality on these switches. The Adaptive Routing Manager currently supports three algorithms; LAG, TREE, and DF_PLUS. It configures AR groups and AR LFTs tables to allow switches to select an output port out of an AR group for a specific destination LID. The configuration of the AR groups depends on the selected algorithm:

- LAG: All ports that are linked to the same remote switch are in the same AR group. This algorithm suits any topology with multiple links between switches, and especially Hypercube/3D torus/mesh, where there are several links in each direction of the X/Y/Z axis.
- TREE: All ports with minimal hops to destination are in the same AR group. This algorithm suits tree topologies such as fat tree, quasi fat tree, parallel links fat tree etc.
- DF_PLUS: Algorithm designed for Dragonfly+ topology. DF_Plus has alternative credit loop avoidance logic, optimised for the topology. It also enhances adaptive routing with additional routing. Advantages of DF_Plus routing algorithm:
    - Improved Congestion Control with Adaptive Routing provides 1:1 throughput for random jobs placement and traffic patterns
    - No need for topology aware job-scheduler

DragonFly+ only requires two virtual lanes to avoid credit loops.

6.7.9 HSI9 [NETW] High-speed interconnect virtualisation support

We confirm. Single Root IO Virtualization (SR-IOV) is supported on the High-Speed Interconnect. This technology allows a physical PCIe device to present itself multiple times through the PCIe bus. This technology enables multiple virtual instances of the device with separate resources. Mellanox adapters can expose in ConnectX-4/-5/-6 adapter cards up to 128 virtual instances called Virtual Functions (VFs). These virtual functions can then be provisioned separately. Each VF is an addition device connected to the Physical Function. It shares the same resources with the Physical Function.

6.7.10 HSI10 [NETW] High-speed interconnect low level API

The Mellanox Quantum Evaluation Board (EVB) and Software Development Kit (SDK) are available to accelerate an OEM's time to market and for running benchmark tests. These rack-mountable evaluation systems are equipped with QSFP56 interfaces for verifying InfiniBand functionality. In addition, SMA connectors are available for SerDes characterization. The Mellanox Quantum SDK provides customers the flexibility to implement InfiniBand connectivity using a single switch device. The SDK includes a robust and portable device driver with two levels of APIs, so the user can choose their level of integration. A minimal set of code is implemented in the kernel to allow for easy porting to various CPU architectures and operating systems, such as x86 and PowerPC architectures utilizing the Linux operating system. Within the SDK, the device driver and API libraries are written in standard ANSI "C" language for easy porting to additional processor architectures and operating systems. The same SDK supports the Mellanox SwitchX®-2, Switch-IB®, Spectrum® and Mellanox Quantum switch devices.

Documentation of the API can be found here:
https://github.com/Mellanox/SwitchRouterSDK-interfaces/tree/master/SDK

6.7.11 HSI11 [NETW] High-speed interconnect IMB performance

For the proposed High-speed interconnect network, we evaluate the maximum latency between 2 random compute nodes as 1.35us + 0.15us per hop in the Dragonfly topology.

The worst-case scenario would then correspond to an IMB test between two nodes in different cells. In that case, the number of hops would be 3 and the maximum latency would be 1.8us.

6.7.12 HSI12 [NETW] High-speed interconnect documentation requirement

As previously described (in section 6.2), a BullSequana XH2000 rack can be populated with up to ten (10) 200Gbps HDR 40-ports switches with up to 16Tb/s aggregated switch throughput and sub-130ns port-to-port latency. Those switches can be used to

build many kinds of topologies including Fat tree, Generalized HyperCube (GHC) or Dragonfly+ (DF+). This latter is based on the concept of dividing nodes into groups while the groups are connected in an all-to-all fashion. A single group is based on a multi-layer Fat tree in which the nodes are connected to the so-called leaf switches (also Level 1 or short L1) which are then connected to a second layer (L2) and a possible third layer (L3), and so on. The top layer switches are the so-called spine switches which are used for the all-to-all connections between the groups to form a Dragonfly+ topology. This topology is very scalable to a very large number of nodes and provide full bi-section bandwidth.

The following illustrations show a few examples of Dragonfly+ topologies:

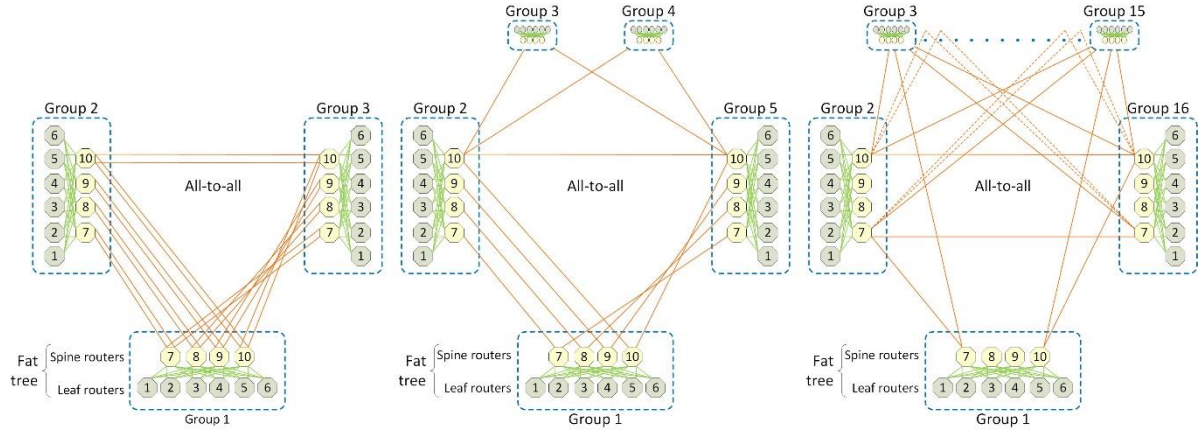


Figure 51: DragonFly+ topologies

After thorough analysis the most optimal interconnect for VEGA is a 2-layer fat tree topology in combination with a Dragonfly+ topology. In the design HDR100 is used to connect the nodes to the leaf switches (L1) and full HDR between the L1 and spine switches (L2). In the concept of the offered BullSequana XH2000 each group of nodes is called a cell. Each CPU and GPU compute partition comprises of 1 or more cells. The free ports of the L2 switches in each cell are used to connect all cells in a single fabric to create a DragonFly+ topology by linking the same numbered L2 switches from each cell (19⇦⇨19, 20⇦⇨20, etc.) together.
The building block for the High-Speed Interconnect is based on the latest generation Mellanox HDR Quantum switch that manages 40 ports in HDR, HDR100, EDR/FDR/QDR on a per port basis. Since each 200Gpbs port can be split into two (2) 100Gpbs ports (so-called HDR100 that utilizes two pairs of two lanes per port) we can therefore take advantage of HDR100 and have up to 80x 100Gbps ports per switch available using splitter-cables or use a combination of HDR and HDR100 ports in each switch. In the CPU partition the compute nodes are equipped with one HDR100 NIC (Network Interface Card) and in the GPU partition the nodes are equipped with 2x HDR NICs. The choice of different speed (at the NIC) level between the two partitions is explained by the need for the GPU nodes, equipped with 4x NVIDIA A100 GPUs, to be able to fully exploit the power of the 4x GPUs and to ensure optimized performance of real life GPU accelerated applications.

The VEGA system is built using 4 cells with:

- 3x cells of 3x racks of CPU compute nodes (cells 1, 2 and 3)
- 1x cell of 1 rack of CPU compute nodes and 2 racks of GPU compute nodes (cell 4)

CPU partition cells 1 to 3, as shown in Figure 52, are made of:

- 3x racks per cell and per rack:
  - 32x blades
  - 3x compute nodes per blade
  - 96x compute nodes
- Non-blocking Fat tree topology
- 17x HDR switches divided as 5-7-5 over the racks
  - 8x spine switches each with 12x HDR uplinks and 18x HDR L1-L2 interlinks
  - 9x leaf switches, each with 16x HDR L1-L2 interlinks and up to 32 compute node connections using 16x HDR ports split into 32x HDR100 ports
  - 8x HDR ports available on each leaf switch for service or I/O nodes
- 288x compute nodes
  - divided over 9x leaf switches (32 each)
  - using 1x HDR100 port each

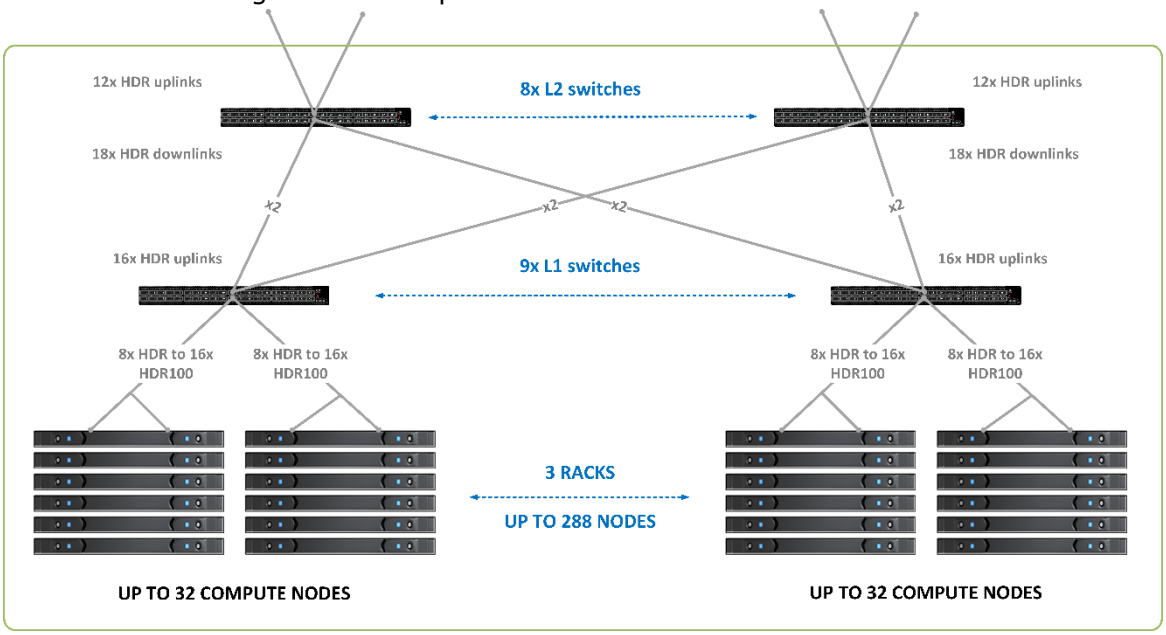


Figure 52: CPU only cells 1 to 3

Hybrid cell #4, as shown in Figure 53, is made of:

- 2x GPU racks and per rack:
  - 30x GPU blades
  - 1x GPU compute nodes per blade
  - 30x GPU compute nodes
- 1x CPU rack and per rack:
  - 32x CPU blades
  - 3x CPU compute nodes per blade
  - 96x CPU compute nodes
- Fat tree topology with a 20:16 (1.25) blocking factor for GPU nodes, non-blocking for CPU nodes
- 14x HDR switches divided as 7-7 over the racks
  - 8x spine switches each with 12x HDR uplinks and 18x HDR L1-L2 interlinks
  - 6x leaf switches, each with 16x HDR L1-L2 interlinks and up to and up to 10x GPU node connections (per L1 switch) using 20x HDR ports
  - 4x HDR ports available on each leaf switch for service or I/O nodes
  - 3x leaf switches, each with 16x HDR L1-L2 interlinks and up to 32x CPU nodes connections (per L1 switch) using 16x HDR ports split into 32x HDR100 ports
- 60x GPU compute nodes
  - divided over 6x leaf switches (30x each)
  - using 2x HDR port each
- 96x CPU compute nodes
  - Divided over 3x leaf switches (32x each)
  - Using 1x HDR100 port each

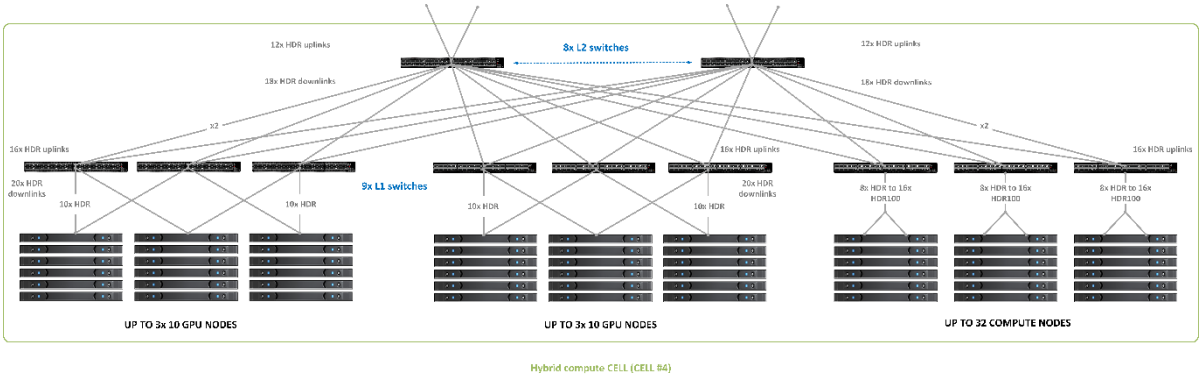


Figure 53: Hybrid cell #4

Figure 54 shows the complete High-Speed interconnect with the offered 4x cells. All cells are interconnected in the same DragonFly+ topology. To summarize the full VEGA solution comprises of:

- 4x cells
  - 3-rack CPU cells (cell #1,2, 3; design according to Figure 52)
  - 3-rack Hybrid cell – 2x GPU rack and 1x CPU rack - (cell #4; design according to Figure 53)
- DragonFly+ topology with 4x groups (cells) of nodes in which:
  - each CPU cell is a non-blocking 2-layer fat tree
  - the GPU cell is a 1.25:1 blocking factor 2-layer fat tree
  - all-to-all interconnection between cells
- Each comprises of 8x spine switches, each with 12x HDR uplinks connected to the same numbered L2 switch in each of the 3x neighbouring cells with 4x HDR interlink cable.
- In total there are 192x HDR interlink (L2-L2) cables between all the cells.

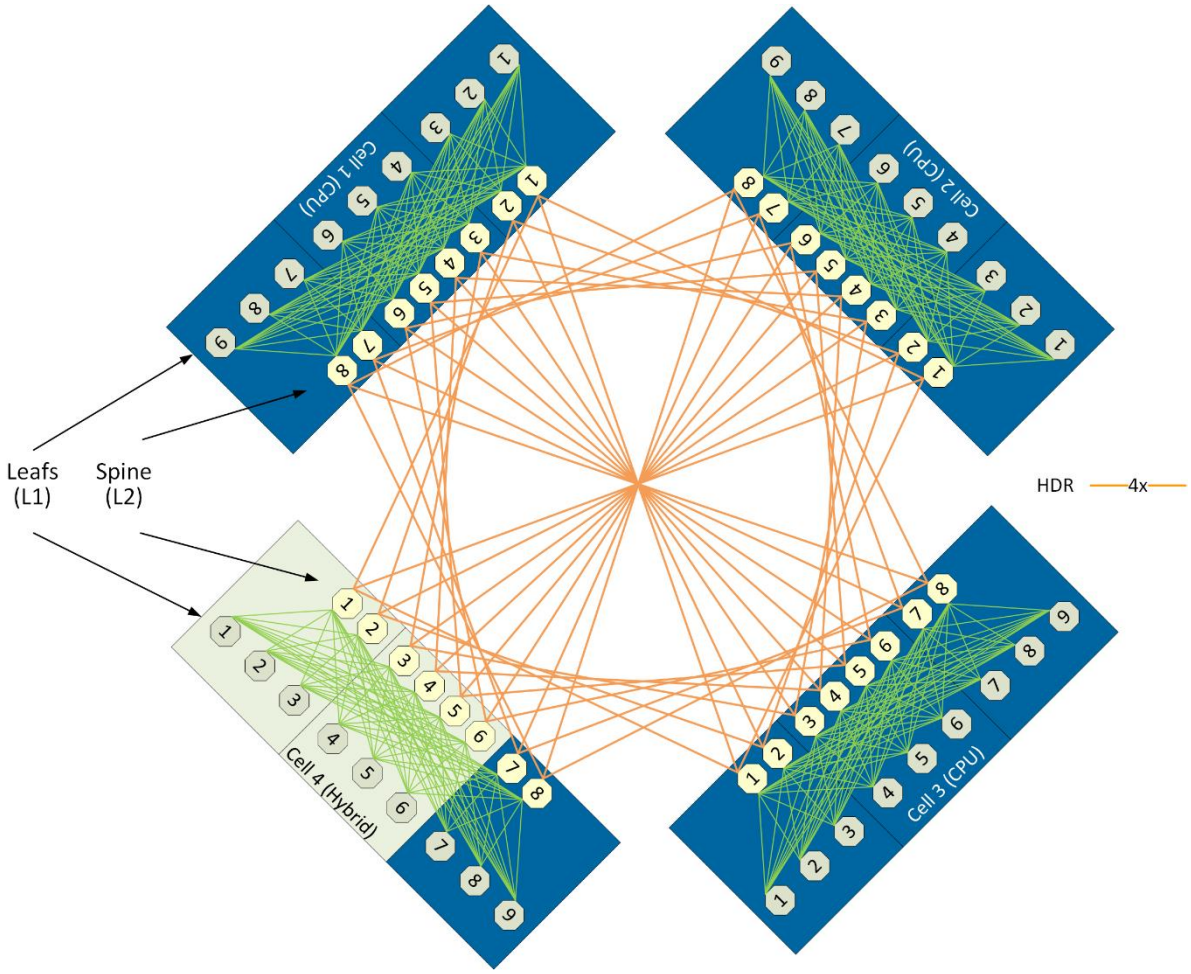


Figure 54: Complete VEGA High-Speed Interconnect

6.7.13 HSI13 [NETW] High-speed interconnect storage protocol support

The High-Speed Interconnect storage protocol is based on the LNET layer that is the basis of communication for Lustre. LNET is supporting RDMA on the InfiniBand standard.

6.7.14 HSI14 [NETW] High-speed interconnect redundant power supply

Each Mellanox Quantum HDR switch provided in our solution comes with 1+1 redundant & hot-swappable power.

6.8 Management network

6.8.1 MNG1 [NETW] Management network

The proposed HPC system for VEGA includes a two-level management network described below. Atos' proposal includes all necessary switches, cables and transceivers (when necessary) to build this network. The second-level switches of this management network (hereafter named "cluster switches") will be connected to the-Top level Ethernet switches of the IP network and isolated from user access.

6.8.2 MNG2 [NETW] Management network documentation

The full ethernet stack solution consists of a two-level management networks dedicated to specific resources including fully populated BullSequana XH2000 Cells. The management network is based on 1/10 Gbps Ethernet technologies and is completely independent of the HDR interconnection network, to avoid any contention between different types of MPI, data and administration streams.

The Ethernet network includes two levels of switches:

- Local rack switches called WELB (Leaf switches)
- Service, IO and other nodes (except compute nodes) leaf switches
- Cluster switches to connect all WELB and Service and IO nodes leaf switches

In addition, the proposed solution includes one management switch allowing updates on WELB switches of the BullSequana XH2000 racks.

The synthetic diagram of the proposed Management Network is represented in Figure 55.

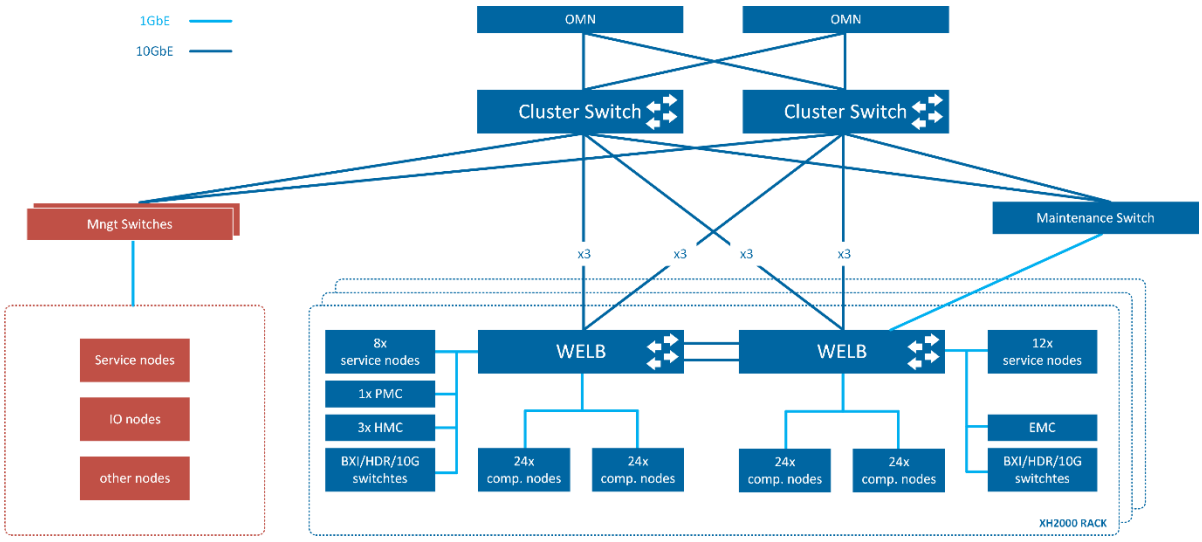


Figure 55: Proposed Management Network for VEGA

1St Level – Rack management network: WELB

Each BullSequana XH2000 rack is equipped with two integrated switches called WELB (sWitch Ethernet Leaf Board). These are custom switches especially designed by Atos to control and management the rack elements.

Each WELB implements three 24-port Ethernet switch instances (WEL0, WEL1, WEL2) and one Ethernet Management Controller (EMC):

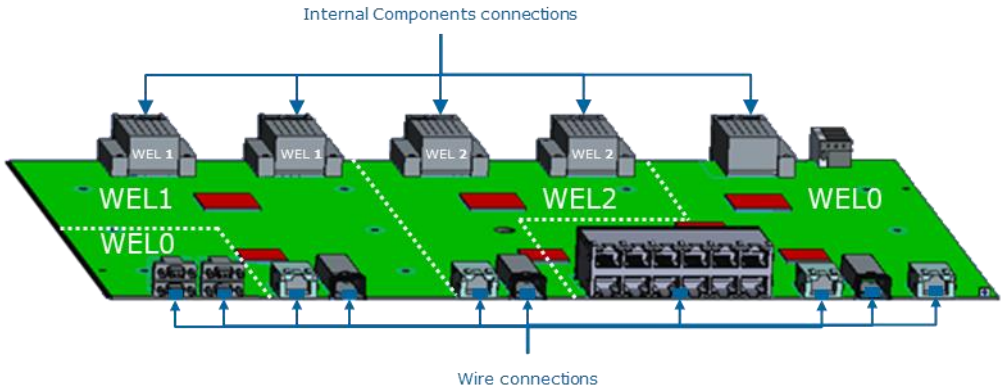
- WEL0: 24-port Ethernet switches for
  - External nodes
  - Power Management Controller (PMC) on WELB0 only
  - Hydraulic Management Controller (HMC) on WELB0 only
  - BXI, HDR or 10G switches management port
- WEL1: 24-port Ethernet switches for rack nodes
- WEL2: 24-port Ethernet switches for rack nodes
- EMC: Connected to all components BMC

The BullSequana sideband network implemented by the EMC is used to:

- Attribute topological addresses to management controllers,
- Control the component power,
- Fix management controller dysfunctions,
- Discover network components (presence/absence),
- Handle signal propagation for power over-consumption or for safety rack stop procedure

Internal and external connection to WELB

WELB have been designed to connect internal components and external components.



- Internal connections:
  - Compute nodes
  - Other WELB
  - Internal Infiniband, BXI and 10G switches management port.
- Wire connections
  - External nodes
  - External switches
  - PMC
  - HMC

As shown in Figure 55, it is the possible to connect up to 20 services, IO or any other external node per BullSequana XH2000 rack directly to the WELB switches. However, in the case of the VEGA system, the compute racks and the service racks will be located in two different rooms so to avoid too many long cabling between the two rooms, we prefer to connect all service nodes to a separate first-level set of ethernet switches.

Service, IO and other nodes first-level switches

Our solution also includes a set of 4x Mellanox AS4610 48x 1GbE + 2x 10GbE ports switches that will be used to connect the following equipment to the management network:

- 8x login nodes (8x connections needed)
- 30x virtualization nodes (30x connections needed)
- 72x Object Storage nodes (72x connections needed)
- 10x SFA400NVX Appliances (40x connections needed)
- 4x Data Gateways (4x connections needed)
- 4x links to manage IP routers and top-level switches

In total, then, 158 connections are needed and will be dispatched over the 4x Mellanox AS4610 switches.

BullSequana XH200 rack maintenance switch

In order to perform maintenance operation on the BullSequana XH2000 rack equipment, a separated 1GbE network is needed. One 1GbE link per BullSequana XH2000 rack is needed. We will use one Mellanox AS4610 switch for this function.

First level switches of the proposed Management Network		
BullSequana XH2000 Rack management	WELB	 • 3x 24 ports 1Gbs Internal connection • 12x 1Gbs external connection with RJ45 ports • 3x 1Gbs RJ45 Uplink for TOP switch connections • 3x 10Gbs SFP+ Uplink for TOP switch connections • 4x HMC connections • 1x PMC connection • 1x EMC RJ45 port
BullSequana XH2000 Rack maintenance	1x Mellanox AS4610	 • 48 x 10/100/1000BASE-T RJ-45 ports • 4 x SFP+ uplink ports, supporting 10 GbE (DAC, 10GBASE-SR/LR/ER/LRM) or 1 GbE (1000BASE-T/SX/LX)
Service, IO and other nodes management	4x Mellanox AS4610	

6.9 IP Network

The proposed IP Network has been designed so that all the following bandwidth requirements are fulfilled. The network is schematically represented in Figure 56 and built using the 2x 100GbE top-level switches, the 2x routers and 4x data gateways allowing IPoIB traffic and connecting over the IP network, the HPST to all other partitions.

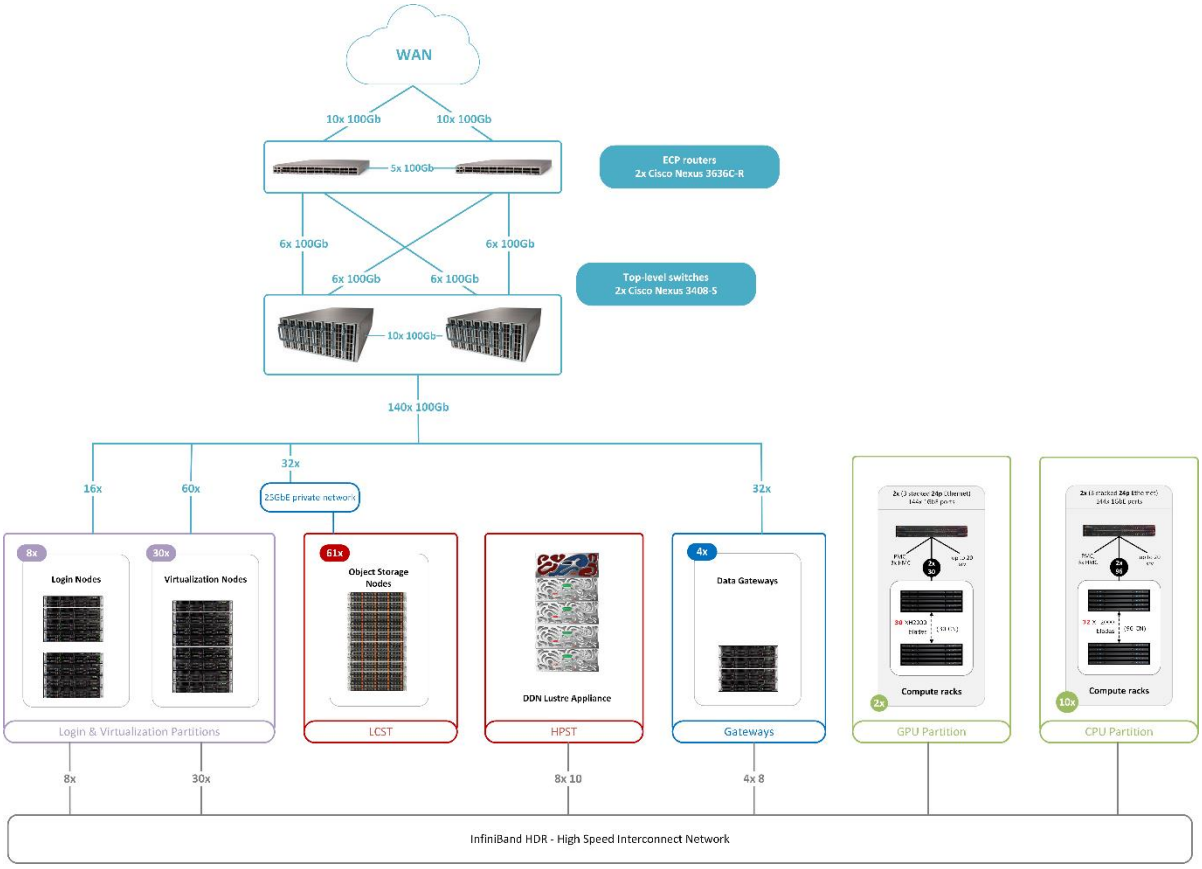


Figure 56: Proposed IP Network for VEGA

The proposed solution includes 4x Data Gateways built on Mellanox Skyway GA100 model.

Figure 57: Mellanox Skyway GA100 gateway

Mellanox Skyway GA100 is an appliance-based InfiniBand to Ethernet gateway, enabling Ethernet storage or other Ethernet based communications to access the InfiniBand datacenter. The solution, leveraging ConnectX's hardware-based forwarding of IP packets and standard IP-routing protocols, supports 200Gb/s HDR connectivity today, and is future-ready to support 400Gb/s NDR and 800Gb/s XDR speeds. Mellanox Skyway contains **8x ConnectX VPI dual-port Adapter cards** which enable the hardware-based forwarding of IP packets from the InfiniBand to Ethernet, and vice versa. Mellanox Skyway also includes the Mellanox Gateway Operating System (MLNX-GW), which manages the appliance and handles the high availability and load balancing between the ConnectX cards, and between gateway appliances.

Network-wide configuration handles Ethernet Link Aggregation (LAG) and implements high availability and load balancing. The solution supports a maximum bandwidth of 1.6Tb/s, utilizing 16 ports with each exceeding 100Gb/s traffic. Connectivity-wise, the InfiniBand ports can be connected to the InfiniBand network via HDR/HDR100 or EDR; and the Ethernet ports using 200Gb/s, 100Gb/s Ethernet or lower speeds.

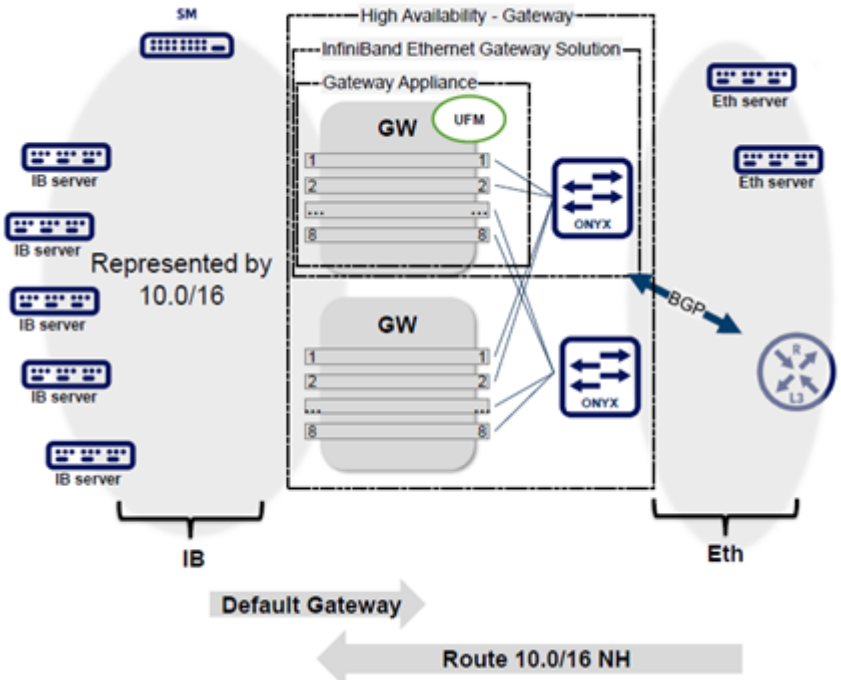


Figure 58: Mellanox Gateway Utilization Example

MLNX-GW Management

Mellanox Gateway operating system, MLNX-GW, is the management software that handles the Mellanox Skyway. It controls the ConnectX adapters, as well as the high availability and load balancing between cards and between gateway appliances.

UFM® Management

Mellanox Skyway is also managed by the external Unified Fabric Manager (UFM) software, which may reside either in the UFM appliance or on any host in the InfiniBand cluster. The UFM enables data center operators to efficiently provision, monitor and operate the state-of-the-art InfiniBand data center fabric

Key Features

- Standard 2U appliance
- 1.6Tb/s solution
- 8-port HDR/HDR100/EDR
- 8-port 200/100Gb/s Ethernet
- Dual Intel Xeon scalable processors
- 80 Plus platinum 1+1 redundant power supply

6.9.1 IPN1[NETW] Sustained IP traffic bandwidth between internal segments [GByte/s]

Sustained IP traffic bandwidth are here given as the minimal number of 100GbE links between the different partitions.

Bandwidth between LCST partition and HPST or Compute or GPU partition

The Large Capacity Storage system is connected to the IP network with 32x 100GbE links via the internal LCST network and traffic with HPST, Compute or GPU partitions is made through the 4x Mellanox Gateways and IPoIB protocol (32x 100GbE links ←→ 32x HDR 100Gbits/s links). Hence, the global bandwidth is here limited by the 32x 100GbE links connection LCST to IP network. The corresponding bandwidth is then 32x 100Gbits / 8 = **400 GBytes/s**.

Bandwidth between HPST partition and Compute or GPU partition

Communication between HPST and compute or GPU partition over the IP network will be made through the use of the 4x Mellanox InfiniBand/Ethernet gateways and IPoIB protocol. There is a total of 4x 8x 100Gb/E and 4x 8x HDR 100Gbits links on the Gateways. The corresponding bandwidth is then 32x 100Gbits / 8 = **400 GBytes/s.**

Bandwidth between Virtualisation and service partition and Compute or GPU or HPST or LCST partition

There is a total of 30x 100Gbits links between the Virtualization and service partition nodes and the IP network. Communication between Virtualization and service partition and compute or GPU partition or HPST over the IP network will be made through the

use of the 4x Mellanox InfiniBand/Ethernet gateways and IPoIB protocol. There is a total of 4x 8x 100Gb/E and 4x 8x HDR 100Gbits links on the Gateway. As stated previously, the bandwidth between LCST and IP network is 400BG/s. The corresponding bandwidth is then limited by the smaller 100Gbits links, i.e. between Virtualization and service partition and the IP network, and will be 30x 100Gbits / 8 = **375GBytes/s**.

Bandwidth between Login partition and any other partition

The bandwidth between the login partition and any other partition is here limited by the number of 100Gbits links between the login nodes and the IP network. There are 8x 2x 100Gbits links corresponding to a bandwidth of 16x Gbits / 8 = **200 GBytes/s**.

6.9.2 IPN2 [NETW] Sustained IP traffic bandwidth to and from routers [Gbit/s]

The bandwidth between each partition and the CPE routers are limited by the number of 100Gbits links between the CPE routers and the top-level Ethernet switches. The number of those links is 6, so the corresponding bandwidth will be 6x 100Gbits = **600Gbits/s** (for both IPv4 and IPv6).

6.9.3 IPN3 [NETW] Throughput on Local Area Network [Gbit/s]

There is a total of 10x 100Gbits and 4x 10Gbits connections per router to the local area network. Throughput on Local Area Network will be above 65 Gbit/s as described in the Benchmarks document.

6.10 Software requirements

6.10.1 System software requirements

6.10.1.1 SWR1 [SW] Integration of software and hardware in general

The proposed solution for VEGA system relies on a strong integration of software and hardware: the combination of BullSequana XH2000 compute system with Smart Management Center software solution offers the perfect integrated platform.

The Smart Management Center solution is a comprehensive, easy to deploy and robust HPC cluster management center solution that meets the requirements of deploying and running HPC management jobs. It is the result of Bull-Atos's long experience in deploying HPC software with continued efforts in Research & Development.

Smart Management Center is part of the Atos HPC software offering:

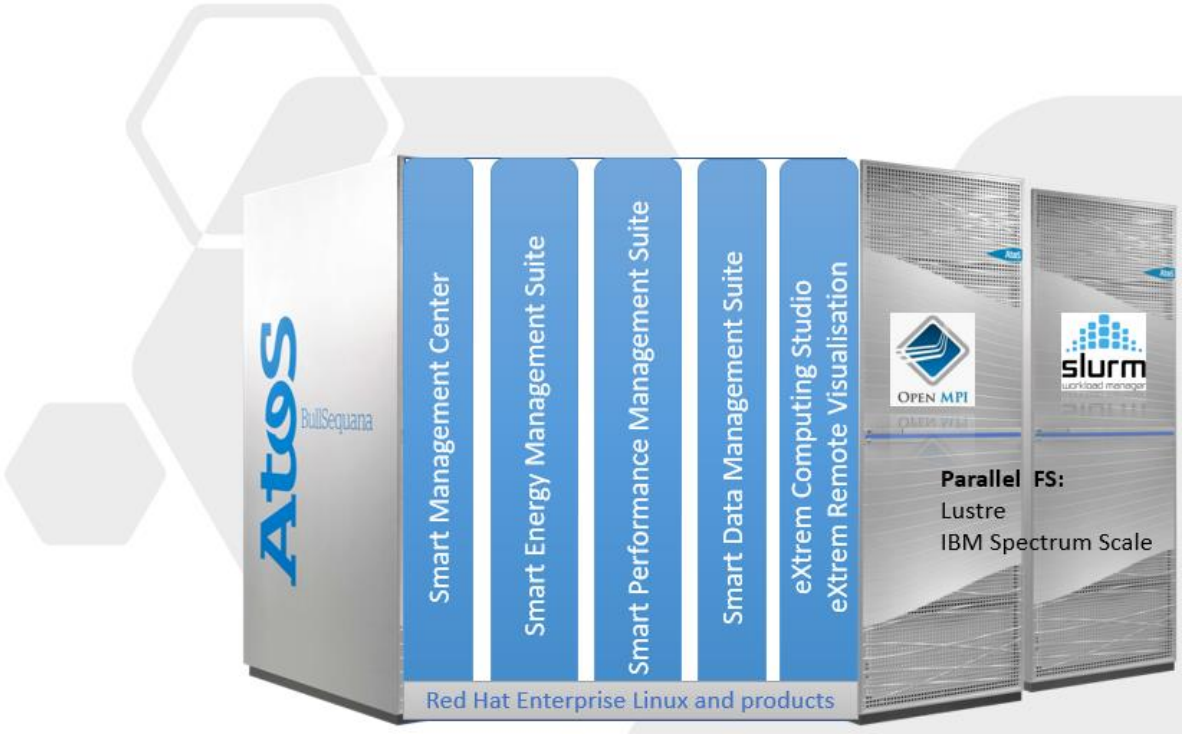


Figure 59: HPC Clusters Software ecosystem

Smart Management Center is targeted at small supercomputers with just a few tens of nodes to supercomputers with up to 1200 nodes.

Smart Management Center is the basis to build complete HPC environment. The software includes components that provide Cluster Management features and can be completed with parallel runtime library (MPI, OpenMP), resources management and scheduling (Slurm, Altair PBS, IBM LFS), Parallel File System (Lustre, IBM Spectrum Scale) and others added value software and hardware to enhance your HPC environment (IO, energy, etc.)

Smart Management Center is tightly integrated with the Atos BullSequana hardware proposition targeting the HPC environment and optimized to run on these platforms.

Key features

Smart Management Center services are installed as bare metal. High availability configurations are possible, and they are automated and supported.

The Smart Management Center targets Tier-2 and Tier-3 supercomputers. It focuses on small & medium market with main challenge on bootstrapping a complete HPC system in a very short timeframe (few hours to less than 3 days).

Taking off the unnecessary components related to very-high scalability (petaflopic or exaflopic class), which brings some complexity for installation and management, the solution has been developed to ease the day to day operation tasks and to provide effective frameworks to help administrators and users to understand what happen on the system.

Smart Management Center is designed to fulfil requirements of system up to 1200 nodes (service and compute) with standard (basic) infrastructure in terms of networks and storage.

Smart Management Center is working with all software components developed in Atos HPC such as OpenMPI library, Slurm resource/batch manager, etc. All those components are optional and can be replaced, or completed, by other alternatives coming from open-source or third parties (ex. Intel MPI, IBM Spectrum LSF, etc.) using Professional Services.

OS Provision	Hardware Control	Infrastructure Discovery
➢ Bare metal provisioning methods: diskful, diskless ➢ Redhat 8.1+ for Head nodes ➢ RedHat 7.7+ / 8.1+ for compute nodes ➢ x86_64 CPUs (Intel/AMD)	➢ Management protocols (IPMI) ➢ Remote console	➢ Static inventory
Scale	**Configuration management**	**Accelerators & interconnect**
➢ Small-medium scale(<1200 nodes) clusters	➢ Software maintenance ➢ Parallel command and file dispatch ➢ Multiple OS support	➢ NVIDIA GPUs support ➢ InfiniBand network support Mellanox OFED ➢ BXI network support (planned)

Figure 60: Smart Management Center - Key features

6.10.1.2 SWR2 [SW] GNU/Linux operating system in general

Atos Smart Management Center comes with RedHat Enterprise Linux (version 8.1+ for head nodes / version 7.7+/8.1+ for compute nodes). Licenses and subscriptions for the lifetime of the procured system are included.

6.10.1.3 SWR3 [SW] Standard operating system higher compatibility in general

The operating System proposed is a full and standard RedHat Enterprise Linux Edition and is fully compatible with the GBU/Linux distributions.

6.10.1.4 SWR4[SW] Resource and workload manager with job scheduler

The proposed Smart Management Center solution includes SLURM as the standard job scheduler.

Atos-Bull involvement

For more than 10 years Atos/Bull has invested in the Slurm community and has heavily participated in the development of the product. Atos/Bull is recognized by the community as an important contributor, with 15 employees having participated in the development. For instance, we provided more than 45 000 lines of code for version 15.08.

We of course enhanced Slurm to perfectly fit with the hardware we design for HPC market. For example, in the domain of power measurement we have improved the overall framework gathering data from BMC power sensor on Bull hardware to reach higher accuracy (up to 4Hz instead of classical 1Hz).

The design of Slurm offers a very high scalability through multi-threaded daemons and commands offering separated read/write locks on data structures. This is further enhanced, within the hierarchical design of the Bull cluster management suite which provides a mechanism to spread the load of epilog and periodic health-check to management units (ISMA) to guarantee scalability.

Atos service operations have been deploying Slurm based projects for years and have set up a complete chain of support skills, from first line to Level 3 support, including escalation to SchedMD, the company that acts as the community lead.

Add-on & value-add components

Atos contributes to the Slurm code base, like for example IB & Lustre accounting.

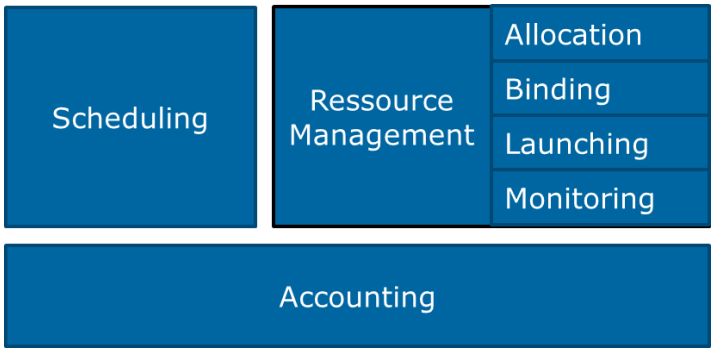
Atos value-add strategy is two-fold:
Early availability of advanced functionalities within Bull Slurm.
Advanced functionalities can be defined in collaboration with customers, or as part as the Atos strategy. Functionalities are discussed and designed with
the Slurm community. Then the development is performed by the Atos development team and is made available to our customers. When the formal approval process is performed by the Slurm community, the contribution is integrated in the reference code base.
Proprietary extensions. These value-add functionalities are delivered with
Bull Slurm package and are not planned to be integrated into community code base.

Features included in our management suite: SlurmTOP, HDEEM hardware-based energy measurement, End of job report, support for Singularity containers), slurmconf helper.

Bull Slurm is currently based on the community Slurm version 19.05.

Functionality

The three functional areas covered by the product are Scheduling, Resource management and Accounting.



Job submission requests specify the job to be executed and a specification of the needed resources (e.g. number of nodes, cores). Scheduling logic works in conjunction with Resource management in order to allocate the needed resources and plan the execution at a given time.

Resource management covers resource allocation, binding on the target execution platforms, launching of the job, and monitoring of the job execution.

The accounting system provides the functionalities allowing defining access control to cluster resources and gathering execution time information. During job execution, Slurm can collect a great deal of accounting information that is recorded in a dedicated accounting database. This accounting information can be used as a basis for implementing charge back schemes, and as an input to optimize resource usage.

Scheduling

- Job submission requests can be either for immediate execution (interactive) or for deferred execution;
- Job submission can specify a specific time slot for execution (startime, duration parameters);
- Job requests are managed per queue. A queue is associated to a partition, which is a subset of available compute nodes;
- Priority assignment to queues;
- Fine grained control over queues, e.g maximum number of cores, time slots when partition can be used;

- Quality of service (QOS) parameters can be associated to a queue, these parameters provide control on e.g. priority management, mode of pre-emption, limits;
- Several scheduling policies e.g. Builtin (FIFO), BackFill (allows out-of-order scheduling), hold (block job scheduling e.g. for maintenance period);
- Flexible priority management, per user, per project (bank account) where priority assignment can take several weighted parameters into account e.g. age, fair-share, job size, partition, QOS;

Resource Management

- Execution platforms: job can be scheduled on any kind of bootable CPU or Multi core; target system can be a bare system, a virtual guest, or a container;
- Multi clusters management;
- Resource specifications thru a number of parameters, e.g. number of nodes, processes, threads, amount of memory;
- Static characteristics can be assigned to a node and specified as a resource constraint;
- Generic (consumable) resources can be defined and associated to a node; in that case the amount of resources allocated to a given job is no more available to other jobs;
- License tokens management;
- Job isolation thru CGroups; at launching time a CGroup with appropriate level of resources is automatically created on the target system;
- Accessibility to the resources managed thru definition of associations specifying cluster, user, project (bank account) and partition. Dedicating a set of resources to a group of users would for example be implemented by creating a specific partition;
- Resources usage limitation: one can limit the amount of resources used on the cluster, e.g. number of nodes, number of cores, number of jobs submitted or concurrently executing;
- Resource limitations can be specified per user or project (bank account);
- Possibility to exclude a node from the candidate execution platforms (DRAINING state, keeps jobs already executing), e.g. for planning maintenance operations;
- Resources can be reserved for future use by specific users and/or bank accounts. Resources can also be reserved for scheduled maintenance.
- Power management. The resource and job manager is the central point to optimize power consumption. To achieve that goal Slurm has specific features such as desired CPU frequency setting at job submission, accurate accounting power consumption per job, management of idle resources by power-down or hibernation until resources get allocated, etc. In addition, an integrated energy-aware monitoring infrastructure allows adding external sensors plugins for out-of-band monitoring. This allows capturing energy usage and temperature of various components (real wattmeter, out-of-band IPMI capture, etc.).
- Resizable jobs. Resource allocations to a job can grow and shrink on demand, with specification of size and time limit ranges at submission.

Allocation

- Node allocation can be exclusive to a job, or shared between jobs;
- Node allocation can consider node definition that includes layout of boards, sockets, core, threads and processors;
- Slurm is aware of all nodes' architecture such as NUMA, sockets, cores and also hyper-threads.
- Individual items (consumable resources) within a node may be allocated to a job e.g. CPU, Socket, Core;
- Usage of GPUs or MIC co-processors is managed through generic resource declaration. Resource allocation takes benefit of this to manage nodes offering requested GPUs or MICs. Locality can also be set up in resource declaration to bind appropriate CPU to appropriate GPU or MIC (NUMA proximity).
- Slurm can be configured to support topology-aware resource allocation to optimize job performance. Slurm can allocate resources to jobs on a hierarchical network to minimize network contention. The basic algorithm is to identify the lowest level switch in the hierarchy that can satisfy a job request and then allocate resources on its underlying leaf switches using a best-fit algorithm. Use of this logic requires manual input of the topology;
- Slurm offers a complete control over task binding to resources through standard Linux Control Groups (Cgroups) mechanisms or thru Singularity containers;

Binding

- Slurm provides binding functionality, also referred to as task affinity or task/CPU binding. This functionality allows distributing and binds each task to a specified subset of the allocated CPUs on the node to which the task was distributed. Different tasks distributed to the same node may be bound to the same subset of CPUs or to different subsets.

Launching

- Slurm is able to manage the launch and communication initialization of MPI processes based on OpenMPI library;
- Support of prolog and epilog programs (root & user);
- Pre & post execution scripts on compute nodes;

Monitoring

- Display of the job state;
- Display of the resources consumed by a job during and after execution;
- Job cancellation;
- Job suspend / resume thru SIGSTOP/ SIGCONT;
- Slurm provides checkpoint control thru SCONTROL command for applications that make use of BCLR library;
- Queues are persistent across Slurm system restart;

Accounting

- The accounting system records more than 80 types of data / metrics related to the job execution; this information can be used to build charge-back functionality; related information includes usage duration, per project, per user, per queue;
- The accounting system can be extended by plugins;
- Non-intrusive profiling mechanism enabling detailed analysis or resource utilization. For each task, information regarding CPU, memory, power, network and I/O usage are stored in HDF5 files. Multiple sensors per node can be handled through a profiling framework which gathers all accounting information that will be used for analysis. Architecture for profiling is highly optimized to ensure reduced data size and scalable centralization of data between all nodes used for a job.

Architecture

Scalability
Slurm design offers a very high scalability through multi-threaded daemons and commands providing separated read/write locks on data structures. Largest computers in the world today use Slurm and higher scalability has been proved using simulation through virtualization1
In addition, the hierarchical design of the Bull management suite offers a mechanism to spread the load of epilog and periodic health-check to all management units (ISMA) to guarantee scalability.

Performance
Slurm has the capability to handle about 1000 job submissions and 500 job executions per second, meaning that in one day you can have millions of jobs submitted and/or run on your supercomputer environment.

Flexibility
Slurm is designed as core and can be extended via optional plugins. About 100 plugins are freely available to support various environments (hardware, MPI versions, etc.). On-site customization is also possible by developing specific plugins using Shell, C, Python or Lua scripting.

Power management
The resource and job manager is the central point to optimize power consumption. To achieve that goal Slurm has specific features such as desired CPU frequency setting at job submission, accurate accounting power consumption per job, management of idle resources by power-down or hibernation until resources get allocated, etc.
In addition, an integrated energy-aware monitoring infrastructure allows adding external sensors plugins for out-of-band monitoring. This allows capturing energy usage and temperature of various components (real wattmeter, out-of-band IPMI capture, etc.).

Fault tolerant
High availability of such critical component is key for any HPC environment. To ensure no point of failure for all Slurm components the Bull management suite is based on pacemaker coming from Red Hat. Slurm daemons also support automatic failover but do not handle related databases. With pacemaker approach, all services are under survey and managed in a centralized way to ease administrative operations.
In addition, applications can optionally continue to run after failure of compute node(s) and can request replacement node(s).

Network topology optimized allocations
Resource allocation is optimized considering the interconnect network topology to minimize communication latency.
Within Bull management suite, the topology is known through the global reference database providing an accurate configuration for Slurm to optimize resource allocation. It is also possible to specify, per job, the maximum number of leaf switches desired and how long to wait for such an allocation.

Security
Slurm integrates MUNGE, an authentication service for creating and validating credentials, designed to be highly scalable for use in an HPC cluster environment. It allows a process to authenticate the UID and GID of another local or remote process within a group of hosts having common users and groups. These hosts form a security realm that is defined by a shared cryptographic key. Clients within this security realm can create and validate credentials without the use of root privileges, reserved ports, or platform-specific methods.

6.10.1.5 SWR5 [SW] Container support

Bull-Slurm provides means to activate containers easily in your HPC jobs.
These mechanisms operate at job launching & completion steps, without interfering with Slurm scheduling and resource management features.

In HPC environment, containers bring the following capabilities:

- Portability: can run on nearly any Linux distribution (even old ones such as RHEL5), and container content may be also nearly any Linux distribution or part of distribution or just a binary;
- Reproducibility: the container may include all the run-time environment (libs, etc.) and also data/files required. The configuration may also be part of the container itself;
- BYOE (Bring Your Own Environment): encapsulation inside container makes it portable between systems so allows users to bring their application environment. It provides a « cloud-mode » for execution;
- Confined execution: limit resource allocation and lock users inside a controlled environment;

- Static environment: provides easy way to archive specific software environment, allow keeping track on what was done and can be re-used in the future just by reloading the container;
- Custom environment: very specific execution environment which requires many additional packages or settings without requiring to modify the base OS nor port the application to a new OS (distrib, version, etc.).

Introduction to Singularity

Singularity was designed for scientific HPC and is released under a modified BSD license. Thus, it enables the use of the low-level drivers & features of HPC network interconnects and GPUs.

One of the architecturally defined features in Singularity is that it can execute containers like they are native programs or scripts on a host computer. As a result, integration with schedulers is simple and runs exactly as you would expect.

As other containers technology, it is file-based and requires no specific services or workflow. Users only have to call a binary with the requested container and are not required to have root privileges.

Singularity maintains the coherency of user identity (coherent inside/outside container), and user context is maintained at the container launch. Similarly, all I/O is passed through the containers and directories can be shared from the host (according to defined authorizations).

Singularity is compatible with existing Docker images with no or little modifications.

Slurm Integration

Bull R&D integrated Singularity by relying on:

- Spank plugin (Slurm plug-in architecture for Node and job control)
- Environment scripts

Internally, Bull integration performs the following tasks:

- After job submission, on the submit host
    - Verify that <container> exists in the SWRepo repository
    - Synchronize <container> to all the nodes designated in swrepo-server.conf
- Before the computation begins, on each compute node selected for the job
    - Verify that <env_script> exists and is accessible

- Create a backup copy of <env_script>
- Append <executable> to the end of <env_script>
- Copy the modified <env_script> into <container>
- Create script "run_script" containing the command "singularity exec <container> <env_script>
- Replace <executable> on the srun command line with "run_script"

– When the computation ends, on each compute node

- Restore <env_script> from the backup copy

Users invoke containers jobs by means of the following options:

```
srun --singularity-image=<container>
[--singularity-env=<env_script>]
[<other srun options>]
<executable>

        <executable> to be run inside <container> image
        using <env_script> to set up the necessary environment
```

6.10.1.6 SWR6 [SW] Open-source preference for system software requirements

Atos is heavily involved in collaboration with various scientific and software communities.

Smart Management Center is deeply rooted in the Open Source ecosystem, where Atos contributes in the context of its software engineering activity. This includes:

– Integrating, validating open source software within the Smart Management Center components;
– Contributing to the community's base code in the frame of advanced research projects, as part of collaboration project with major scientific users, or to contribute to shared implementation of key market requirements

The following Open Source components are part of the Smart Management Center software:

- Operating system: Red Hat Enterprise Linux
- BlueBanquise and Ansible for installation and configuration management
- Yum for package management
- ConMan for console management
- pdsh for parallel shell
- pacemaker/corosync for High Availability (needs an additional RedHat subscription)
- Prometheus server and dashboards for cluster monitoring (planned in 2020Q2)

6.10.1.7 SWR7[SW] User management with LDAP

The user management can be managed through LDAP protocol (based on 389 Directory Server within RHEL8) with an integration of an external LDAP-based directory. The configuration will be set up under high availability to ensure no single point of failures.

6.10.1.8 SWR8 [SW] High-performance MPI implementation

Our proposal includes Intel MPI library. The Intel MPI library implements the MPI 3.1 standard on multiple fabrics. Intel MPI Library uses OFI (Open Fabrics Interface) to handle all communications, enabling a more streamlined path that starts at the application code and ends with data communications. Whether you need to run Transmission Control Protocol (TCP) sockets, shared memory, or one of many interconnects based on Remote Direct Memory Access (RDMA)—including Ethernet and InfiniBand*—Intel MPI Library covers all configurations by providing an accelerated, universal, multifabric layer for fast interconnects via OFI.

6.10.2 System management software

6.10.2.1 SMS1[SW] System management software – cluster management software

Smart Management Center supports many kinds of cluster configurations, but most common are Single Module and Multi Modules. The single module scenario is the standard.

Typical cluster deployment architectures include:

- Service nodes dedicated to services functions: Management nodes, Login nodes, Gateway nodes (IP, Lustre, IPOIB), Batch scheduler Server, Licence Server, Subnet manager Server, Security Server.
- Storage Service nodes providing file services;
- Compute nodes dedicated to application workloads

Network is typically set-up as follows:

- Low latency interconnect network supporting inter-nodes communications and I/O operations
- Dedicated administration network;
    - VLAN, subnets are supported
    - Communications with front-end network thru a specialized gateway service node

Smart Management Center can be deployed using 2 types of topologies: single module and multi-modules.

Single Module topology

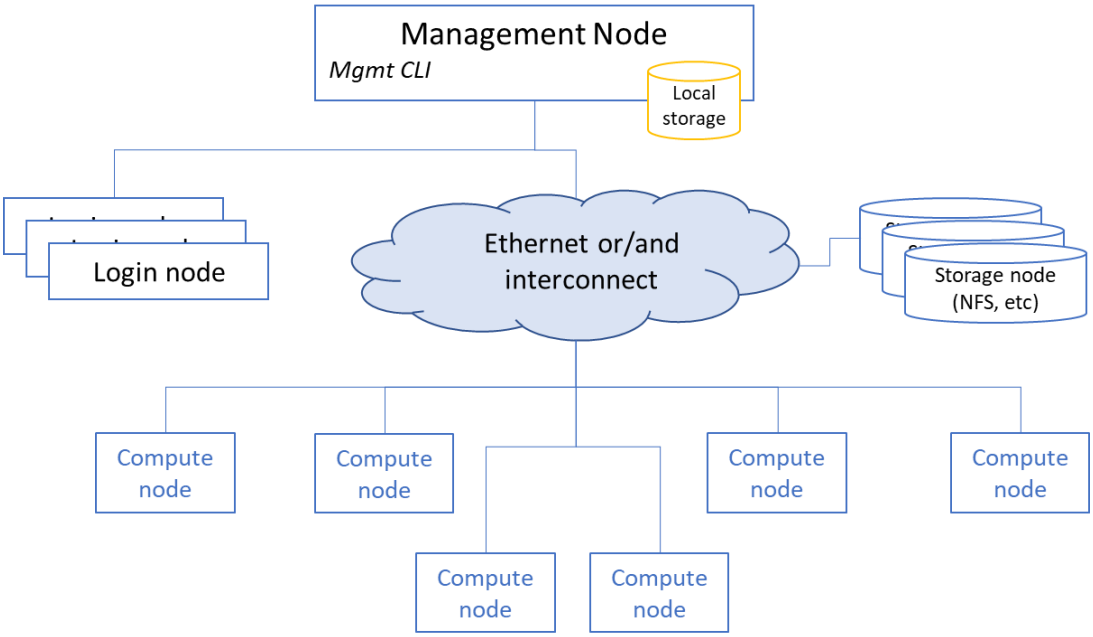


Figure 61 – Single module topology in non-HA mode

In many cases, the need for High Availability of the management nodes will be needed, in those cases, a simple single module in HA mode topology will be used:

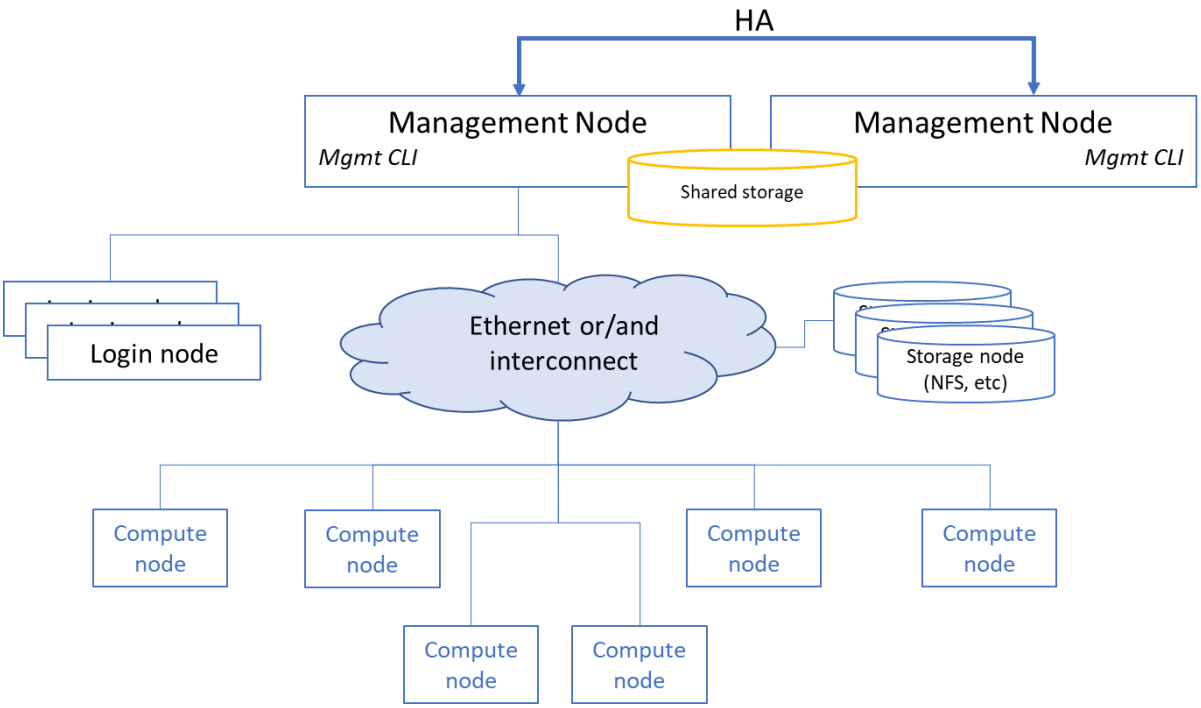


Figure 62 - Smart Management Center in HA mode

Multi-modules topology

In other cases, the multi-module topology (HA and non-HA modes) will be required to be able to scale with the required performances:

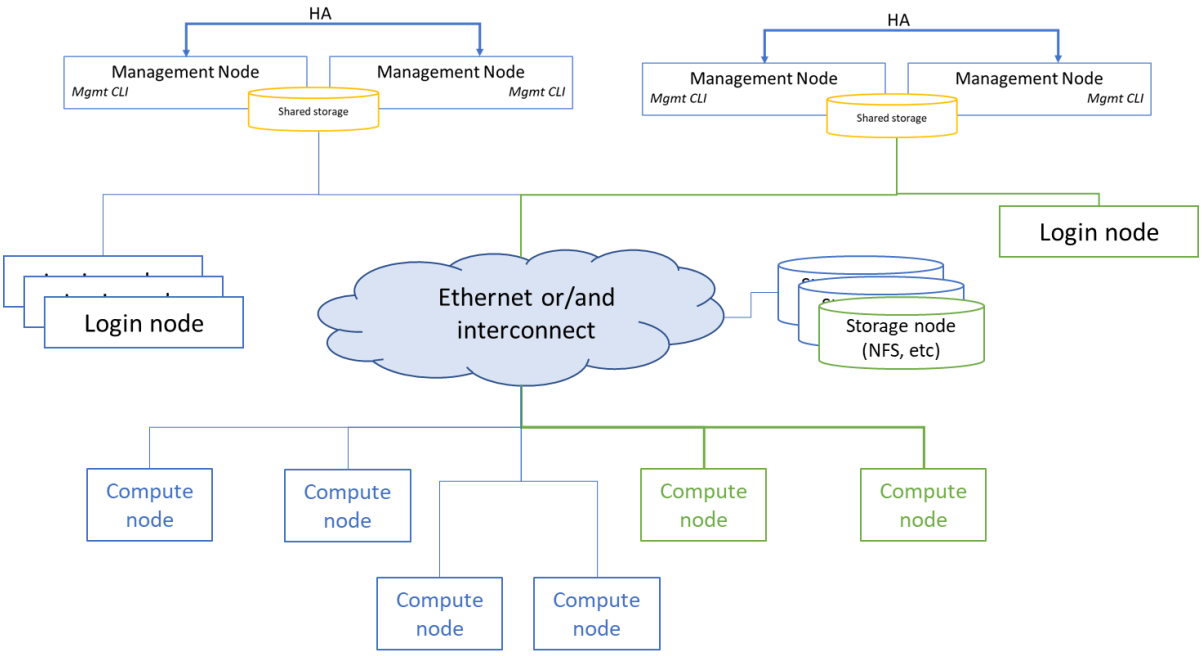


Figure 63 - Smart Management cluster - with multi-modules (blue/green) and HA

6.10.2.2 SMS2 [SW] System management software – node provisioning

Regarding node provisioning, **Smart Management Center** supports on-disk and disk-less OS setup. By default, the service nodes are on-disk systems and compute nodes are disk-less. It is also possible to have system with local disk but still use disk-less OS boot and use local drives as scratch.

The node provisioning is done by image-based installation with standardized format for diskfull and diskless nodes: with this feature we can provide an image-based deployment solution that can fit node installation on-disk or diskless. The image concept is like a virtual machine or a container in which we will find an installation of the operating system and of the additional software components in generic way. Image can be created for different type of usage such as compute, GPU nodes, service node, etc. to fit software requirement for each type of requirement. The image will have version control and a tool to allocate one image to one or a set of nodes (bootset). The image modification can be performed directly inside the image itself but will then require the node to reboot to take it into account in a permanent way. For testing, or in case the administrator does not want to reboot nodes (apply a kick fix for example), a live update can also be performed directly on the nodes (but will not modify the reference image). This mode is the default one with the **SMC** proposition.

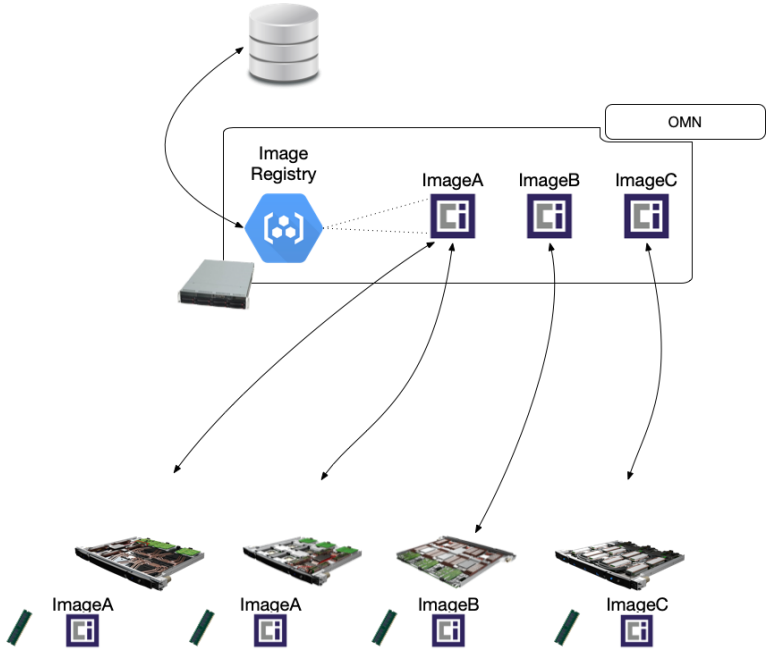


Figure 64 Image-based deployment principle

6.10.2.3 SMS3 [SW] System management software – basic hardware monitoring

The monitoring component of Atos Smart Management Center is Prometheus (server and dashboards).

Prometheus is an open source software created by SoundCloud, aimed at monitoring events and creating threshold alert management, by collecting metrics in software developed in-house. It is a monitoring that collects data in real time using an http extraction model, and flexible requests.

Prometheus joined the Cloud Native Computing Foundation in 2016 as the second hosted project, after Kubernetes.

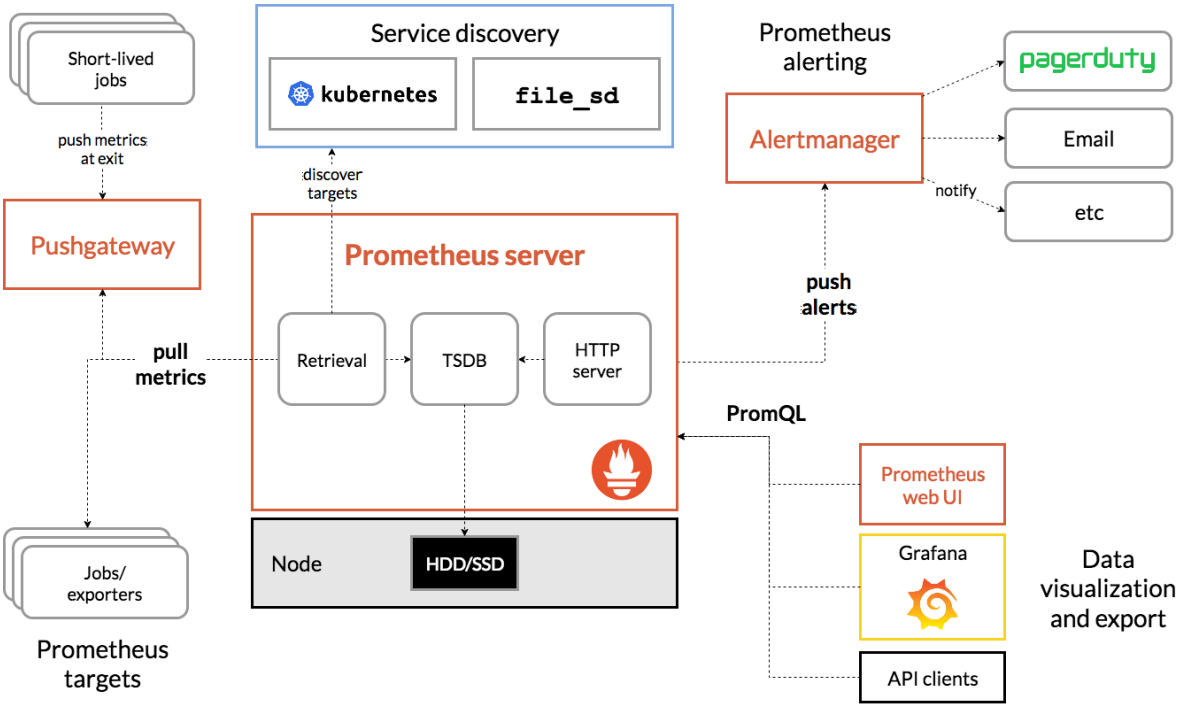
Prometheus's **main features** are:

- a multi-dimensional data model with time series data identified by metric name and key/value pairs
- PromQL, a flexible query language to leverage this dimensionality
- no reliance on distributed storage; single server nodes are autonomous
- time series collection happens via a pull model over HTTP
- pushing time series is supported via an intermediary gateway
- targets are discovered via service discovery or static configuration
- multiple modes of graphing and dashboarding support

The Prometheus **ecosystem consists of multiple components**, many of which are optional:

- the main Prometheus server which scrapes and stores time series data
- client libraries for instrumenting application code
- a push gateway for supporting short-lived jobs
- special-purpose exporters for services like HAProxy, StatsD, Graphite, etc.
- an alert manager to handle alerts
- various support tools

This diagram illustrates the **architecture of Prometheus** and some of its ecosystem components:



Source : https://prometheus.io/docs/introduction/overview/

Alertmanager

The Alertmanager handles alerts sent by client applications such as the Prometheus server. It takes care of deduplicating, grouping, and routing them to the correct receiver integration such as email, PagerDuty, or OpsGenie. It also takes care of silencing and inhibition of alerts.

6.10.3 System and application monitoring requirements

6.10.3.1 SAM1 [SW] Lightweight continuous performance profiling

Performance profiling for application is provided in Atos proposition by using the Lightweight profiler described on chapter 5.8 The system related monitoring tools are part of SMC and managed by Prometheus.

6.10.3.2 SAM2 [SW] Performance report generation

The generation of human readable reports about application and infrastructure performance is key to better understand the behaviour of the running jobs. To enable this Atos has developed a set of tools in the Smart Performance Suite solution, which will be provided with the system. In this suite you can find a dedicated tool which aims at providing insights on job execution: **Modular End of Job Report (MEJR).**

MEJR is providing modules to cover different runtime components: I/O, MPI, system, energy, etc. The base set of modules is provided by Atos, but they can be extended (new metrics, etc.) and/or new modules can be added.

The outcome is a report structured as a file including metadata information about the job and the runtime environment: batch scheduler submission line, nodes used, partition, job execution, etc. Each enabled module (I/O, system, runtime, etc.) will then contribute to the final report. Then detailed, per topic (i.e. per module) information will be provided according to what was enabled in the MEJR configuration file. This configuration file allows to define which information must be collected and then displayed in the report.

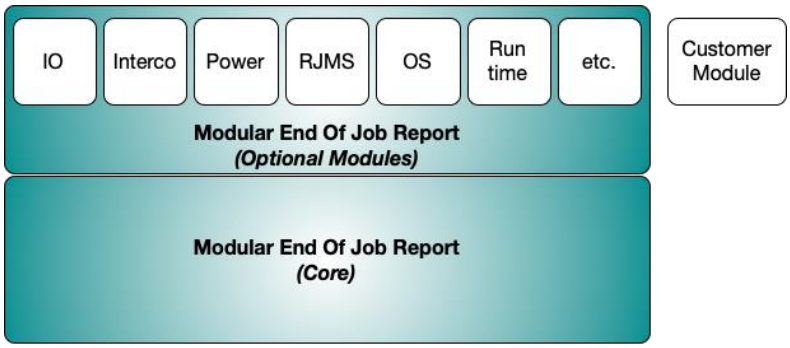


Figure 65: Modular End of Job Report components design

The metadata and mandatory information are (at minimum):

- Job submission parameters,
- Resources used during job execution,

- System configuration (kernel version, runtime version, key environment variables, interconnect information, etc.)

The standard modules Atos is targeting to provide are:

- I/O module provides aggregated information about I/O during the job execution (volume, path, etc.),
- Interconnect module provides aggregated information about interconnect configuration (type of HCA, firmware version, driver version, etc.),
- Power module provides core information about energy consumption during the job execution based on standard hardware counters,
- OS module provides key information about the system environment on which the job is running (list of key packages with version, environment information such as memory, CPU usage, etc.),
- Runtime module provides key information about the runtime used during the job execution such as the MPI or OpenMP stack (version, environment settings, etc.),
- Runtime performance module provides key information about runtime calls during the job execution (profiling),
- Resource Manager module provides key information about the Slurm environment and events.

Administrators can set the modules they want to use (or not). As the modules can be modified (forked), the content of the output can be tuned to better match a customer's expectations. It is hence possible to extend the scope of a module or reduce the information that it provides. In future versions there are plans to provide configuration file directives that will let the user select which output from which module are to be included in the final report. Fort the time being the structured nature of the report file can already be leveraged to remove unwanted information or enhance the existing items.

MEJR can be configured by default in the resource manager to enforce report generation. It is possible to generate a specific for a group leader or for administrators to provide key information about the user's runs. Last on-demand report generation by the end-user for a specific job execution is available. This is configured by the resource manager administrator or by the user himself (for on-demand). The report can be sent to the requester through a post-execution script (will required specific system configuration).

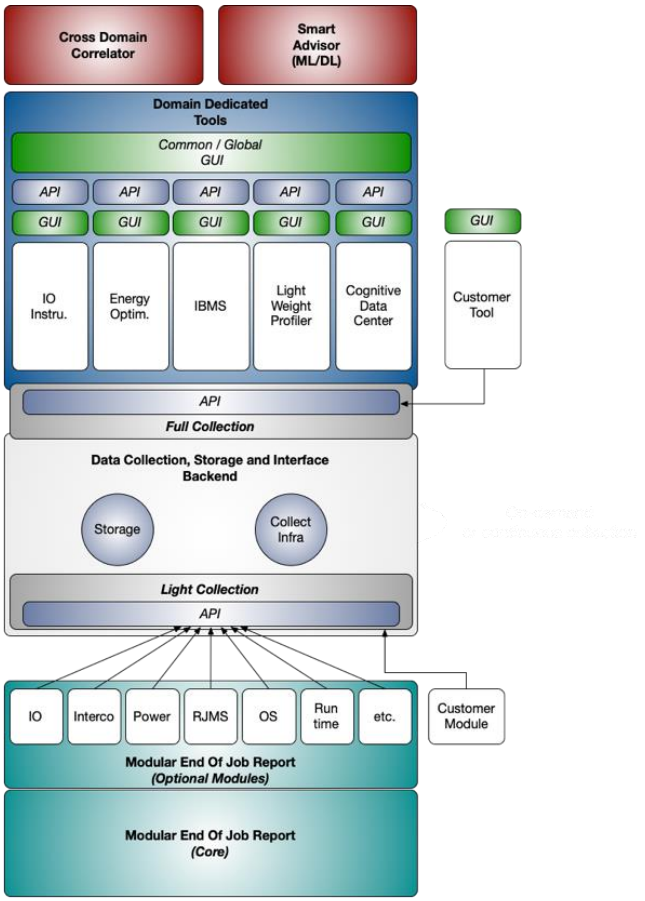


Figure 66: Performance Studio design overview including MEJR

The MEJR is part of a larger performance analysis toolset called 'Performance Studio' which can cover deeper analysis of the application itself targeting specific domain (I/O, MPI, etc.). The purpose of Performance Studio is to provide an integrated framework for data collection as well as data analysis and presentation. The MEJR is part of the solution but can also be used as a self-contained component.

The forthcoming Performance Studio will provide a centralized data repository as well as a unified data collection mechanism. In this context MEJR can rely on a specific API to gather/access the information to generate reports. This API can also be used directly by third-party tools (or modules) to generate other types of report or outputs. This API is only provided with the 'Data Collection, Storage and Interface Backend' which is part of the Performance Studio solution.

6.10.3.3 SAM3 [SW] Mechanisms for correlation

Atos Smart Management Center (SMC) includes a Message Bus mechanism designed for large scale environment, based on Fluentd and Kafka that are becoming de facto standards.

All metrics collected from the different monitoring system are logged on the local storage through Fluentd Daemons maintained on the OMNs (**O**perational **M**anagement **N**odes: Atos SMC Management Nodes).

The metrics producers may be sytemd, rsyslog, Prometheus exporter, collectd and specific monitoring systems. All these producers communicate with Fluentd daemon that will write on the local storage maintained on the OMNs. The metrics will also be streamed to Kafka so that each data may be used by any Consumers as Prometheus, Graphana or any other Data analysis tool on a near real-time timescale.

7 Installation of Procured System

7.1 Installation Time Schedule and Project Management

7.1.1 System Installation

7.1.1.1 SIN1 [SERV] Project Management

Atos will provide comprehensive project management services for the system installation.

A project manager ("The Atos Project Manager") will be assigned from our High-Performance Computing Global Delivery team. He will manage the delivery of the data center adaptations and the supercomputer, from the date of coming into force of the contract up to the final acceptance and handover to the maintenance team.

He will manage all aspects:

- Project governance, reporting, escalation and problem solving
- Project planning
- Resources allocations
- Project risks, and
- Acceptances

For the VEGA project, Atos has selected as a preferred profile an Atos Project Manager based in Slovenia, speaking both Slovenian and English language.

His mission is the following:

At inception of the project the Atos Project Manager will propose the governance to be implemented including committees, reporting and escalation processes.

Accordingly, he will be the prime interface to the Contracting Authorities during this period and will maintain a regular dialogue with the Contracting Authorities Project Manager, through both verbal and official report updates.

He will plan all actions according to the jointly agreed planning considering on-site operational constraints. This will include planning for:

- data center adaptations and more specifically cooling and power infrastructure modifications specific to the proposed Atos supercomputer;
- supercomputer installation and integration
- testing and acceptance; and
- training course (duration and course program)

He will identify and manage risks in a transparent manner and will oversee the escalation process when necessary.

He will coordinate the project globally, manage the deliveries of all Atos & Atos' partners and the engineers involved in the project (data center experts, delivery software and hardware experts, support engineers, benchmarkers and local partners…).

During the acceptances, he will organize the acceptance process with the Contracting Authorities project manager,

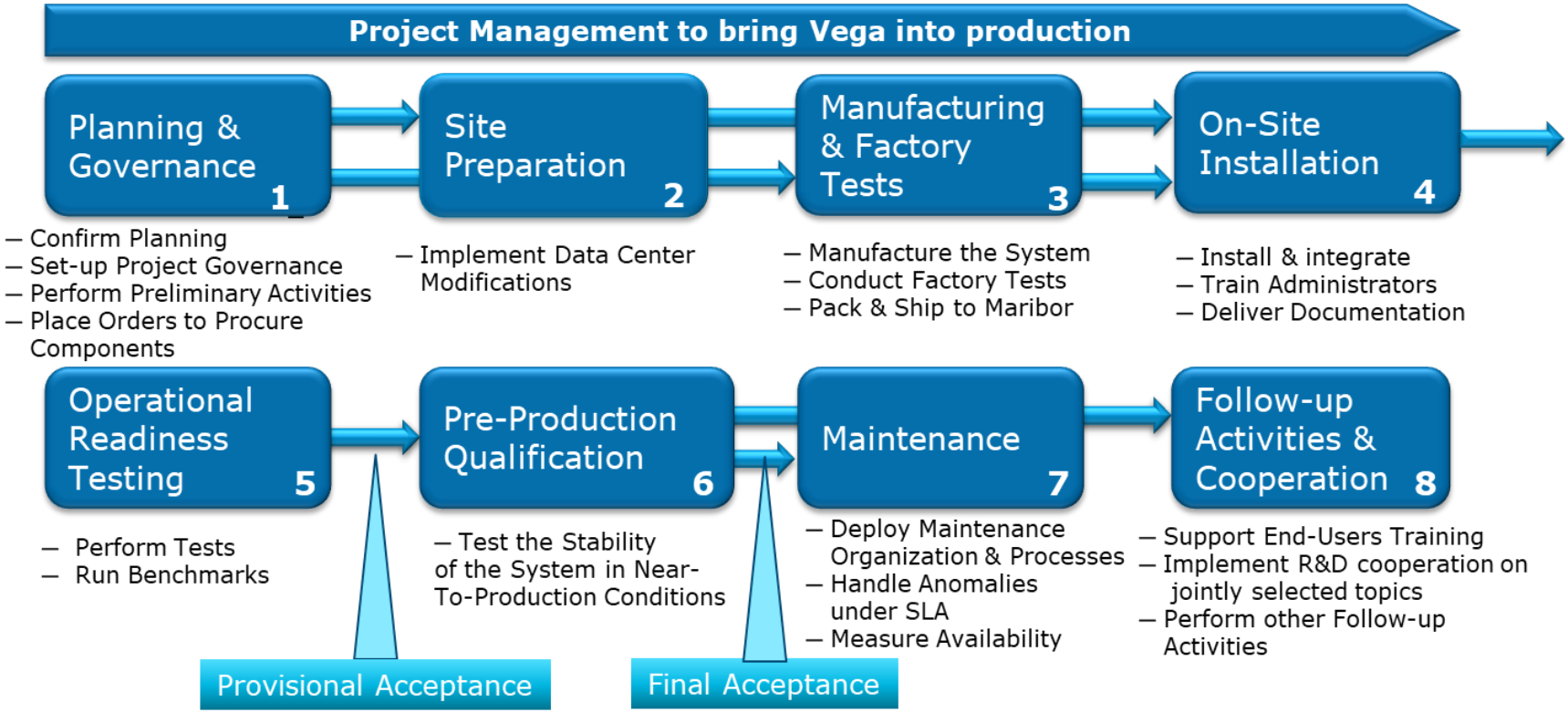
- For Provisional Acceptance, to ensure that all the tests described in the Chapters 11.3.1 to 11.3.6 of the "Annex_Technical_Specifications_23-04-2020(final)" are successfully completed in accordance with the acceptance procedure.
- For Final Acceptance, to ensure that all the test described in the Chapter 11.3.7 of the "Annex_Technical_Specifications_23-04-2020(final)" is successfully completed in accordance with the acceptance procedure

He will also organize the system administrators training in accordance with concerned trainers and Contracting Authorities resources.

For this project, this include in particular:

- Reviewing and validating all contractual obligations with IZUM
- Planning and coordinating the data center modifications works
- Preparing the datacenter plan with complete floorplan (racking, cabling, rack layout...)
- Monitoring the ordering and manufacturing processes of the supercomputer
- Coordinating the installation and configuration at Atos factory in Angers, France.
- Organizing the on-site deliveries with IZUM
- Monitoring the physical delivery of the goods on site
- Ensuring security in the process of site delivery
- Coordinating the installation, configuration and integration works
- Producing and delivering the required documents and reports, including the system configuration, setup and interconnection schemes;
- Managing IZUM participation to the software implementation
- Organizing the acceptances (Provisional & Final) in coordination with IZUM
- Ensuring the maintenance processes are in place from the Final Acceptance
- for the handover of the supercomputer to Atos maintenance team
- Ensuring planning is set for the Follow-Up Activities & Cooperation stage, and more specifically the support for end-users training.

The Atos HPC delivery methodology builds on many years of experience implementing large scale HPC systems in Europe including weather and climate centers, industry clients, national research facilities, Universities and Government agencies.



Further details on each of these stages are provided in answer to the questions below.

7.1.1.2 SIN2 [SERV] Benchmarking Support

Atos has included in this proposition the resources for running all benchmarks as described in the Chapters 11.3.4 to 11.3.6 of the "Annex_Technical_Specifications_23-04-2020(final)"

7.1.2 Time Schedule

7.1.2.1 TSC1 [SERV] Installation Time Frame

To meet the delivery requirement stated in the "ePRO_Contract_Template" document, Atos assumes the starting date for the implementation would correspond to the Contract coming into force (contract signed, and bond issued).

The starting date, "T0" is estimated in this section, as of September 1st, 2020.

Indeed, Atos understands from the Q&A session that, as the due date for remittance of propositions by the bidders has been postponed to 13.07.2020 until 12:00, a likely date for starting the execution of the contract could be beginning of September.

Should however the T0 date be too close to the date communicated by IZUM as being the latest one for contract signature during the Q&A session (question answered by 19.05.2020 at 09:49), i.e. 31.10.2020, the final acceptance date of the schedule presented here may become very difficult, or even impossible to meet.

At the time of writing this proposition, the Covid-19 pandemic is still impacting international logistics and manufacturing, as well as the ability to travel in many European countries.

The presented tentative schedule assumes however the pandemic will be over at the time the project implementation will start, or that the conditions for components procurement and manufacturing, as well as working conditions will be the same as before the outbreak, or not have a significant impact.

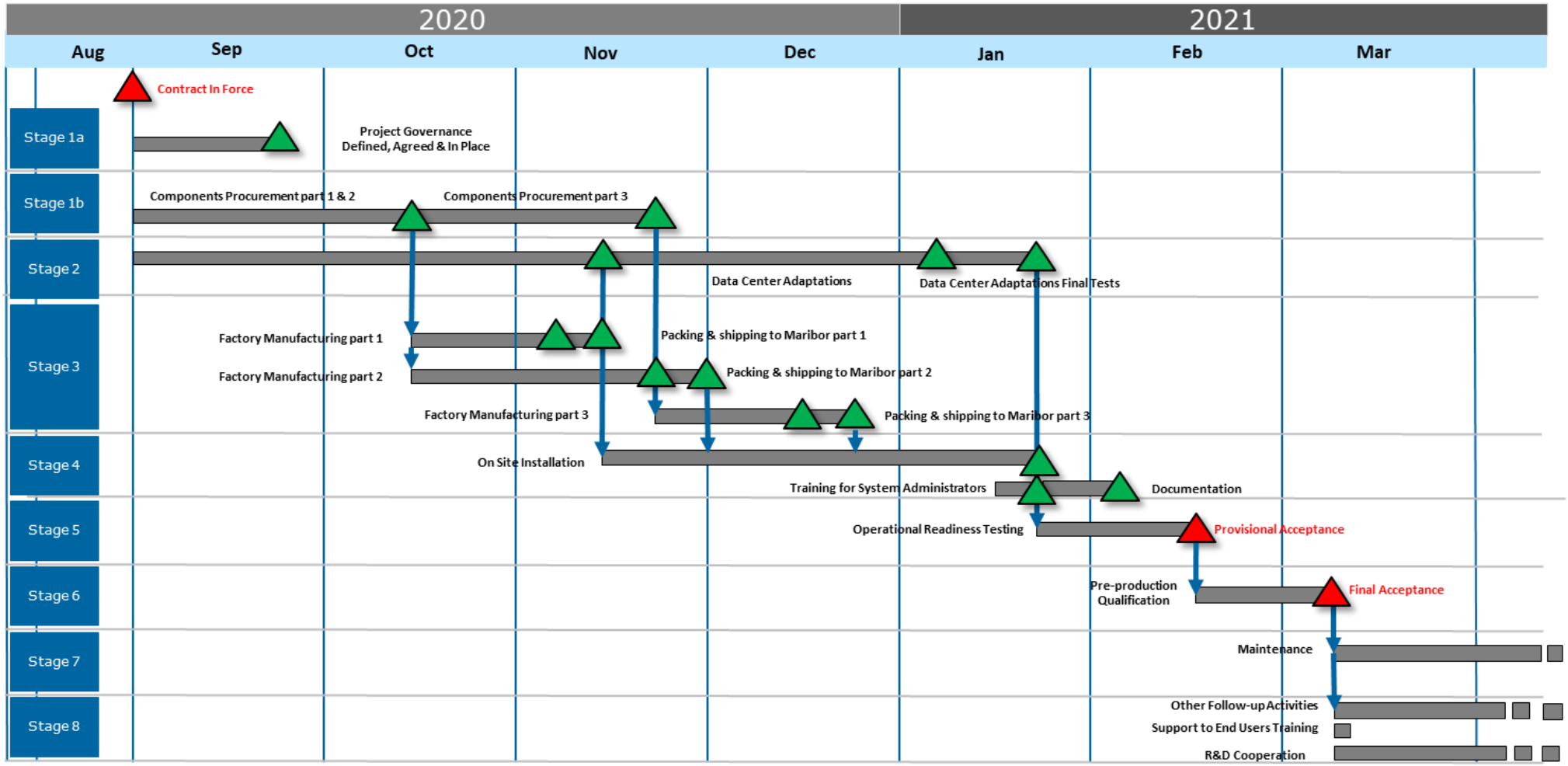
Should this not be the case, the delivery of the project may be subject to changes for procurement and manufacturing timeframes, as logistic and working restrictions may be imposed by authorities in the countries where the components (including different countries in Asia and the USA) or the system are manufactured (France) or in the country of destination and project implementation (Slovenia). This may include potential travel ban or restriction.
IZUM and Atos may have to adapt this schedule having in mind the VEGA project completion date (final acceptance date) set in the "ePRO_Contract_Template" document.

The diagram below provides a view of the project implementation, with a tentative schedule. The detail of each stage is indicated in the next section.

The tentative schedule allows parallelism between tasks to mitigate potential slippages:

- The procurement of components can be done in several batches with the production at factory to start accordingly
- This would allow several on-site deliveries (3 in the presented diagram below), to start the on-site installation at the soonest, for example with services nodes
- While certain components for data center modification can be subject to long lead-times, provision is made for the site preparation to be done in several steps (whenever necessary) so as not to delay the first batch(es) of on-site installation

More about risks and mitigation actions can be found in the corresponding "project risks" chapter.



7.1.2.2 TSC2 [SERV] Time Schedule

Description of the Activities for each Stage

Stage 1a: Planning & Governance and Preliminary Activities

Planning & Governance activities will include:

- Establish governance including
  - project progress meetings, committees, escalation processes,
  - responsibilities and roles as indicated in the TSC4 [SERV] section,
  - Content of project progress reports
- Review the main parts of the project: configuration, performance, commitments Confirm planning for:
  - data center adaptations and more specifically cooling and power infrastructure modifications;
  - delivery, installation and integration of the various parts of the system;
  - testing and validation (acceptances); and
  - training course (duration and course program)
- Confirm documentations to be delivered, as indicated in TRA [SERV] System Documentation Accessibility
- Review the risk register ("VEGA Project – Risk Register document")

A project kick-off meeting will be organized, during which the above will be presented. A corresponding "VEGA Project - Planning & Governance" document will be signed by both parties.

The following preliminary activities will also be performed:

- Check the configuration:
  - Review architecture, racks (number and type water/air/position), Ethernet and IB networks organization, servers function (service nodes, management nodes node, compute nodes, GPU nodes, other nodes), switches (number and type), storage configuration (number of disks, Raid implementation, Disk array …), PDU and power outlets…
  - Review parameters: Raid, MDS node, OSS node, Lustre disk array, high- availability, batch scheduler, other software, user authentication, Root password providing, security functions …
- Establish with IZUM the site parameters: equipment host-naming conventions, network color cabling, network access points, cabling requirements, IP address and external access to network. This will lead to the completion of the "VEGA Project – Site Parameter List"

- Establish delivery details; access for vehicle, loading bay, lift restrictions, walk delivery route, access for engineers and parking, health and safety conformance, building access times, contacts and process for delivery, local traffic conditions and any restrictions to access route
- Outline, discuss and agree installation plans, equipment floor plan.

Stage 1b: Manufacturing: Ordering of Components

Provision is made in the tentative schedule for the possibility to have 2 main batches of components delivery and then related manufacturing and factory testing, depending on the specific lead times for a few critical ones.

This Stage correspond to an internal Atos process.

Atos will:

- Handle a configuration review during which Atos will review:
  - All the hardware items (MI – Marketing Identifier) of the configuration: racks, platforms, processors, memory, disks, add-on cards, cables, power supply…. in order to launch the ordering process
  - All the software licenses and references (MI) to be ordered: management suite, compilers, etc.…
  - The physical and logical organization of the equipment in the racks and their location in the data center. This will result in the production of a list ("The Netlist") including a summary of the characteristics of the racks (power consumption, weight, size, height/width), all cables (power, fiber, Eth or IB copper cable whenever appropriate) and their length and tag.
  - The components ordering documents will be defined at the end of this review
- Order the corresponding components as well as spare parts
- Issue the corresponding manufacturing entry file (XML files) to schedule the manufacturing

Stage 2 Site Preparation

- Preliminary site inspection
  - Comprehensive technical review of all installed equipment and in-place services
  - Fluid flow review
  - Coordination with IZUM regarding acceptable time slots for execution of modification work considering ongoing business processes at site and project milestones imperatives
- Plan of execution of the Data Center preparation/modification works
  - Floor plan layout for RIVR1, RIVR2 and utility spaces

  - Raised floor static reinforcement calculation
  - Cooling plan
  - Power supply plan
  - Cabling plan
  - Phase execution and time plan
- RIVR 1 site preparation
  - removal and decommissioning of doors and part of separation walls in RIVR 1, necessary for decommission of CRACK units and installation of new HPC racks,
  - installation of temporary protective walls and other protection to current server room of IZUM,
  - decommissioning of raised floor, CRACK units, existing racks with electrical installation and cold aisle, electrical distribution panels, cooling pipes
  - execution of new wall penetrations with fire protection sealing, according to modification plans,
  - installation of a new metal under-floor support system for the delivery and installation of BullSequana XH2000 racks, installation and detailing of the raised floor according to layout plan,
  - reinstallation of fire protective partition walls and all doors (after final installation),
  - new CRACK installation with new piping system,
  - new cable support system installation.
- RIVR 2 site preparation is not required as Atos will accommodate the site characteristics and reuse existing equipment.
- Electrical supply system adaptation for Supercomputer, increasing available power
  - removal of all unnecessary switchboards and other electrical equipment,
  - new cable pathways preparation from the transformer station to all electrical users,
  - new UPS room cleaning and preparation,
  - adaptation of the transformer station with extended power capabilities,
  - installation of necessary new switchboards according to Atos power needs,
  - installation of the new UPS for increased power consumption,
  - installation of all new necessary cables for new equipment connections, including cables from the transformer to switchboards, UPS devices and end racks.
- Cooling system adaptation for Supercomputer with increased available cooling power and installation of water-cooled systems
  - removal of existing cooling equipment and installations from facility,
  - emptying of cooling system,
  - removal of external cooling units on roof,

  - removal of unnecessary cooling pipes, pumps and other devices,
  - disassembly of cooling station in garage,
  - closing of unnecessary wall penetrations,
  - preparation of new pipe pathways from the cooling station in the garage to RIVR1,
  - installation of reinforcements for external units on the roof,
  - preparation of new pipe pathway from the fourth floor to the roof for adiabatic water,
  - installation of new external adiabatic dry-cooler units for heat rejection,
  - installation of new pumps and heat exchangers, including redundant pump for on-site stock,
  - installation of a new hydraulic system for water temperature regulation,
  - installation of new pipes with necessary vents and other equipment,
  - installation of new local PLC for regulation control and process automation,
  - installation of new distribution pipes under the raised floor with connection to the BullSequana XH2000 racks in RIVR1,
  - installation of new CRACK units for the additional UPS room,
  - installation of new pipes for the additional UPS room.
- Management system adaptation
  - installation of new management processors for RIVR1, new UPS and new cooling station,
  - installation of new sensors and other end elements,
  - installation of management cabling and connections,
  - software integration of new infrastructure devices into the existing DCIM,
  - integration of HPC devices (processors, etc.) to the existing DCIM,
  - development and integration of regulation processes for the cooling system,
  - development and integration of the energy measurement system,
  - development and integration of power capping mechanisms,
  - integration of genset utilization,
  - integration of availability measurement and reporting.
- Site inspection and validation
  - recurring inspections and consulting during buildup of site
  - recurring tests and activation of the new completed systems
  - management of system integration and functional tests
  - final handover

Stage 3 Manufacturing: System Manufacturing and Factory Tests

This Stage correspond to an internal Atos process.

- Manufacturing corresponds to the node assembly and testing, racking, integration of network equipment and storage devices in order to build the supercomputer.
  Final racking and networking schematics are completed.
  Atos has multi-level integration and testing steps before assembly of the supercomputer. Tests are conducted on component level (boards, memory, disks, CPU...), node level (CPU node, GPU, node, service node) and cluster level (supercomputer cluster). All steps are part of internal quality process before delivery to customer site.
- Software installation may begin when the configuration checklist is done and when the preload is generated. It can be done at the factory, or on IZUM Site, either remotely form Atos' premises or on-site depending on what is the most appropriate to meet delivery constraints. (Please note that Atos factory is a secure site, therefore remote connection is a subject to limitations).
- Factory tests will then be run on a part of (or the entire) supercomputer at the factory in (almost) the same conditions as on the IZUM site
  This will consist in running unitary-tests or multi-tests to burn in the cluster's components (mainly, CPU, memories, disks), removing the weak components (low perf or error) and reviewing the equipment's and their operational reliability
- The supercomputer will then be de-assembled and packed for transportation, along with the corresponding initial spare part kit. A shipping list/delivery note is issued to the freight forwarder.
- Atos will then transport and insure the goods from its factory in Angers, France to IZUM Site in Maribor, Slovenia.

At the end of this stage, the risk register is updated.

While Atos will make its best efforts to procure the components to manufacture the system and deliver it on IZUM Site in one batch, uncertainties on lead-times lead us to make in the planning, the possibility of manufacturing and delivering in 3 batches, in order to meet the delivery of a portion of the goods on IZUM Site by 31.1.2021.

Stage 4 On-Site Installation & Administrator Training

This stage will include, **for on-site installation**:

- Coordinating hardware and software resources on-site and remote support resources
- Check the delivered goods against the shipping list & conduct a visual inspection of the packages.

- Unpack and verify the status. In case of any missing, damaged or non-compliant item, raise the corresponding ticket as per the Atos procedure (Supply Chain Post Delivery Claim)
- Dispose all surplus packing
- Move and align all racks following the implementation of the layout diagram
- Route the external cables (Fiber / Copper) in the cable path
- Cable the equipment's following the netlist and the tags of each cables
- Connect the power supply
- Connect the water pipes
- Check the power line and water connection
- Open the water taps for cooling
- Install software of the management & service node if not done at factory
- Check, test and burn in network links
- Deploy nodes if it is not already done during at factory
- Power-on the supercomputer
- Tune and run global tests

As Atos BullSequana XH2000 racks cannot be leaned on its back due to weight restriction. Atos will provide a special tool for movement of this racks inside the datacenter during delivery.

In case of component breakdown during the On-Site Installation stage, the DOA (Dead On Arrival) process is used to change the dead components at the factory and to follow it during the process, as per the Atos procedure (Supply Chain Post Delivery Claim).
DOA ticket is analyzed by factory team and when approved, the new equipment is sent back from the factory with an RMA form which enables to return the defective equipment at the factory
New equipment may be delivered directly on site or can be taken from the initial spare parts kit, which will then be replenished.

During this stage, Atos will produce the Document of Technical Architecture (also called DAT). It will include the physical layout for each of the rack, along with the cables that will be connected to each rack, as well as physical maps of all the networks of the solution indicating what is connected in each of the ports of the networks. This document will be named "VEGA Project – Document of Technical Architecture"

When this stage is completed, the "VEGA Project - On-Site Installation Acceptance" document is issued, acting as an acceptance of the stage.

At the end of this stage, the risk register is updated.

This stage will also include the **administrators training**:

As a standard part of the installation process, Atos will provide an on-site, tailored, workshop training for the system administration team. The materials provided to support training are left with the support team and can be used for further internal training.

IZUM staff will be welcome to attend appropriate parts of the On-Site Installation where practical hands-on experience and instruction can be received

The documentation for this training will also cover the operational procedures that need to be performed to keep the infrastructure working optimally.

The content and duration of the training are detailed in the answer to requirement TRA2 [SERV].

Stage 5 Operational Readiness Testing (Provisional Acceptance)

This stage corresponds to the performance of all tests described in sections 11.3.1 to 11.3.6 of the "Annex_Technical_Specifications".

This includes the following tasks:

- Agree with the Contracting Authorities on detailed planning and resources
- Agree with the Contracting Authorities about the details regarding the execution of the functional tests described in section 11.3.3 depending on the system design.
- Set-up the proper environment for the tests to be run in accordance with the provisions of section 11.2
- Perform with IZUM:
    a. The hardware checklist as per section 11.3.1
    b. The software checklist described as per section 11.3.2
    c. The functional tests as per section 11.3.3
    d. The synthetic tests and benchmarks as per section 11.3.4
    e. The application benchmarks as per section 11.3.5

  The rules described in section 11.3.6 will apply for the IO500, HPCG and HPL benchmarks

To conduct these tests:

- Maintenance periods may be requested for implementation of corrections and component failures may be addressed provided that the rules defined in section 11.2 are not compromised.
- At each successive test stage any test failures encountered will be documented, agreed and corrective actions defined with timescales.
- Throughout the period Atos software and hardware personnel will be on hand, either on-site or remotely, to assist with the conduct of the tests,

investigate and help resolve any perceived difficulties and determine the cause of any failures.

Each test result will be signed-off by the parties and reported in a folder to constitute the Provisional Acceptance document ("VEGA Project – Operational Readiness Testing – Provisional Acceptance")

At the end of this stage, the risk register is updated.

Stage 6 Pre-Production Qualification (Final Acceptance)

This stage corresponds to the performance of the test described in section 11.3.7 of the "Annex_Technical_Specifications".

To conduct this test:

- Maintenance periods may be requested for implementation of corrections and component failures may be addressed provided that the rules defined in section 11.3.7 are not compromised.
- At each successive test stage any test failures encountered will be documented, agreed and corrective actions defined with timescales.
- Throughout the period Atos software and hardware personnel will be on hand, either on-site or remotely, to assist with the conduct of the tests, investigate and help resolve any perceived difficulties and determine the cause of any failures.

The test result will be signed-off by the parties and reported in a folder to constitute the Final Acceptance document ("VEGA Project – Pre-Production Qualification – Final Acceptance)

At the end of this stage, the risk register is updated.

Stage 7 Maintenance

This stage will validate the organization in place for the Maintenance period as more specifically described in the appropriate sections of the Atos proposition.

A workshop will be held during which Atos will detail the corresponding organization, processes, roles and responsibilities.

Stage 8 Follow-Up Activities & Cooperation

The stage will start upon final acceptance. It covers the following activities:

- The end-user training workshop: as required in TRA3 [SERV], Atos will support, during the first year, the introductory course organized by IZUM, with a talk about system-specific application tuning
- Regular meetings (one per semester) in a place to be agreed upon the parties to cover the presentation of new products and services, and discussions about cooperation. The frequency of such meetings could be different upon agreement between the parties.
- Invitations of IZUM representatives to attend events either organized by Atos or for which Atos is a sponsor and/or an exhibitor.
- The BUX Membership. As an owner of an Atos system, IZUM can become a BUX Member. The BUX (Bull User group for eXtreme computing) is an independent world-wide group of users that will cooperate to increase the capabilities of large-scale, parallel scientific and technical computing supplied by Atos, to promote the exchange of information and understanding of these systems, and to provide guidance to Atos on the essential development and support issues for large-scale technical systems.
- BUX was founded on June 20th, 2011 at ISC'11 in Hamburg, as a result of discussions that took place in Paris first half of 2011 between users of Atos extreme computing solutions. The objectives of the BUX are:
- Sharing of experience between Members and with Atos
- Providing Atos with inputs and helping set priorities for Atos (technology, service and support …)
- Networking with Atos HPC experts during BUX events
- Highlighting emerging technologies and helping foster long-term developments
- (more about BUX can be found here: https://atos.net/en/solutions/high-performance-computing-hpc/bux)
- Cooperation as more specifically detailed in 8.4.1.1 COL1 [SERV] Cooperation Topics

Time Schedule of the Activities

Below is the calendar duration of each of the above listed stages (in calendar weeks):

Starting Date	Completion Date	Duration (calendar weeks)	Stage
T0	T0+3	3	Stage 1a "Planning & Governance and Preliminary Activities" (excluding the procurement of components)
T0	T0+6 & T0+11	6 & 11	Stage 1b "Planning & Governance and Preliminary Activities" for the procurement of components. Atos made provision for 2 main deliveries of components at the Atos factory (part 1 & 2 and then part 3)
T0	T0+19	19	Stage 2 "Site Preparation". Site preparation will be completed in several steps due to long lead time for a few components. The objective is to start the on-site installation earlier without waiting for the final completion.
T0+6	T0+10 & T0+12 & T0+15	Respect. 4, 6 and 9	Stage 3 "Manufacturing & Factory Testing". The planning makes provision for the possibility to manufacture, test and then deliver on IZUM site in 3 batches to start the on-site installation at the soonest.
T0+10	T0+19	9	Stage 4 "On-Site Installation & Administrator Training" will be completed in several steps according to physical deliveries. Training and documentation will last another 2 weeks, leading to a completion of the stage at T0+21
T0+19	T0+22	3	Stage 5 "Operational Readiness Testing". It is concluded by the issuance of the Provisional Acceptance"
T0+22	T0+25	3	Stage 6 "Pre-Production Qualification". It is concluded by the issuance of the Final Acceptance

7.1.2.3 TSC2 |SERV] Project Risks

Atos places great emphasis on the reduction of risk and mitigation strategies. Several risks are characterized into multiple categories within our risk register, as described below:

Risk #	Risk description	Pre–mitigation Probability (H/M/L)	Pre-mitigation Impact (H/M/L)	Post-mitigation Probability (H/M/L)	Post-mitigation Impact (H/M/L)	Mitigations
D1	Design (overall): too many unproven main technological choices	H	H	L	L	This risk has been dealt with during the design phase by selecting as main technologies, proven ones having already been deployed in an integrated way for large scale projects (Most core components exist today so there are no dependencies on R&D schedules): a) BullSequana XH2000 as DLC cabinets b) BullSequana X2410 blades running AMD Epyc 7H12 CPUs c) Mellanox interconnect integrated in Atos BullSequana XH2000 (codesigned between Atos & Mellanox) d) the Lustre parallel file system running on a DDN high performance storage subsystem The main new component will be the new BullSequana X2425 blade with the new Ampere GPU for which mitigation actions are described below

D2	Design (CPU specific): use of new CPU from AMD	H	H	L	L	Considering the VEGA project timeframe our information indicate AMD Milan will only be available in volume for a limited number of projects for which they have already been allocated by its manufacturer. We have therefore decided to choose AMD Rome, widely available, which also benefits from a better price to performance ratio.
D3	Design (GPU specific): use of newly released Nvidia Ampere GPU	H	H	L	L	We have selected the newly released Ampere GPU from Nvidia. To mitigate the risk using a new component, we have considered the following elements: a) As indicated above, the new BullSequana X2415 blade with the new Ampere GPU will be delivered to other projects in Europe (already awarded to Atos) before the delivery of the VEGA project will start. b) To mitigate the risk of availability Atos has included in its joint forecast with Nvidia the VEGA project to ensure their allocation. The current VEGA delivery timeframe is compatible with the general availability and shipment conditions of Nvidia

D4	Design (high-speed storage specific): inability to deliver the required performances	M	H	L	H	Atos has selected long term partner DDN high-end storage and jointly sized the equipment with DDN engineers to deliver our committed performance.
I1	Implementation: delays in components sourcing and thus in manufacturing and on site-delivery	M to H	H	L	H	For main components (BullSequana XH2000 cabinets and blades, Mellanox interconnect, AMD CPUs and Nvidia GPUs, Atos can elect to pre-order long lead item parts at its risk to minimize potential delays, as these components can be shared between several projects
I2	Implementation: delays in site preparation (data center modifications)	M	M	L	L	Atos has selected NTR as its partner for the data center adaptations. NTR long experience with IZUM and Atos early engagement with NTR allowed both parties to work in anticipation on the design of the modifications.
I3	Implementation: power and cooling requirements at delivery different than expected	L	M	L	L	Atos has precise tools to estimate power consumption Atos supplied to NTR all requirements for Atos BullSequana XH2000 and other components (storage, networking…). Power and cooling requirements have been jointly rightsized.

I4	Implementation: the system does not provide the expected performance. The impact is that the benchmarks could not be reproduced at the committed numbers	M	M	L	L	Atos uses well proven benchmarks methodology for result extrapolation, and has acquired significant experience with requested benchmarks and selected technologies During implementation, Atos can decide to add additional nodes/storage to or provide optimization services as required to make up the performance shortfall.
I5	Implementation: High rate of 'early life' failure. The impact would be that components would need to be replaced meaning that Atos may fail the availability tests initially	H	H	M	L	Atos undertakes robust factory testing during stage 3 to shake out early life failures in the factory. Atos will provide well sized on-site spare parts kit to quickly replace failing components Atos added contingency time in the current schedule and doesn't assume that everything will run smoothly to plan.
I6	Implementation: High numbers of bugs found in pre-Production testing. The impact would be system instability or loss of functionality	L	M	L	L	Besides Atos and NTR teams on site, Atos factory and R&D are working closely with our suppliers to act on time at factory or IZUM location if need be

M1	Maintenance: Inability to guarantee the SLA	M	M	L	M	Atos has set-up an organization combining its NTR partner resources for call handling and level 1, with its own resources for level 2 and 3. Standard Atos processes and tools will be implemented. NTR personnel will be trained and certified by Atos. The initial spare part stock to be held on IZUM Site is designed based on Atos accumulated experience of similar projects.

7.1.2.4 TSC4 [SERV] Responsibilities and Roles

Below is the description of the roles and responsibilities of all parties involved during system installation and early operation in the form of a RACI (Responsible, Accountable, Consulted, Informed) model.

This is made for the 3 parties involved in the project:

- IZUM as the VEGA project owner
- Atos as the supplier and prime contractor
- Atos' partner NTR as Atos subcontractor for:
  - Data center modifications design and implementation, as well as maintenance of equipment supplied in this frame
  - Provider of manpower to be included in the Atos project delivery team and managed by the Atos Project Manager

Main Tasks/Stages	Atos	Atos' partner NTR	IZUM	Comment
Define Project Governance	A & R	I	C (and approve)	Atos will design the Project Governance and propose it to IZM for review, adjustments and validation
Project Management	A & R	I	C	Atos will regularly report to IZUM project progress. Necessary planning adaptation will be discussed with IZUM.
Preliminary Activity	A & R	I	C	IZUM will be asked to actively participate to specific tasks such as the definition of on-site parameters
Manufacturing: Ordering of Components	A & R		I	IZUM will be kept informed regularly about the progress
Manufacturing & Factory Tests	A & R		I	IZUM will be kept informed regularly about the progress
Site Preparation	A	R	C	NTR will be responsible for the execution of these works while Atos will remain accountable IZUM will be responsible to facilitate access to Atos and Atos' partner personnel
On-Site Installation	A & R		C	Atos will propose IZUM personnel to participate to certain tasks to get in-depth knowledge about the system

Operational Readiness Testing	A & R		C	IZUM to participate to the acceptance process and issue the corresponding acceptance certificate
Pre-Production Qualification	A & R		C	IZUM to participate to the acceptance process and issue the corresponding acceptance certificate
Maintenance	A & R	R	I	NTR will be responsible for the execution of call handling and level 1 support for all products to be delivered in this proposition. NTR will also deliver level 2 and 3 and spare parts for infrastructure products delivered for data center adaptations. Atos will be responsible for level 2 and 3 and spare parts for all other products to be delivered in this proposition. Atos will remain accountable for the maintenance and support. IZUM will be kept informed of all issues and will participate to the regular joint meetings and approve Atos performance.

8 Maintenance, Support, Training and Collaboration

8.1 Maintenance and Support

8.1.1.1 MSR1 [SERV] Maintenance and Support Duration

Atos has included in this this proposition Maintenance and Support for all Infrastructure Products, Hardware Products and Software Products(1) which are part of this proposition for a period of 5 (five) years starting from the Final Acceptance.

The cost for one additional year of maintenance (year 6) is 612,505€ (six hundred and twelve thousand five hundred and five euros) excluding VAT.

(1): as defined in section 9.2 of the document Infrastructure Products, Hardware Products and Software Products mean respectively the equipment for data center modification (especially power supply and cooling), the hardware (especially networking, compute and storage elements), the software with their Documentation proposed by Atos and supplied to IZUM, to be delivered and installed by Atos.

These definitions are used more specifically in this section 8.1 and in section 9.2 related to Maintenance and Support

8.1.1.2 MSR2 [SERV] Maintenance and Support Coverage

Atos has included in this proposition Maintenance and Support for all Infrastructure Products, Hardware Products and Software Products which are part of this proposition.

Details of service level are provided in sections 9.2.1 SLA and Support Services and 9.2.7 Call Center and SLA, of this document.

Details of the operating mode for handling incident is provided in section 9.2.8 Incident Management of this document

8.1.1.3 MSR3 [SERV] Special Software Support Coverage

All software components (the Software Products) which are part of this proposition are covered by the support. For specific open source software (Ceph storage) the support is provided by the community.

Hardware dependent software components which are part of the Atos Supercomputer Management Software (SMC) are fully supported.

8.1.1.4 MSR4 [SERV] Reaction Times

Reaction time, intervention time and recovery times are more specifically indicated in sections 9.2.1 SLA and Support Services and 9.2.7 Call Center and SLA, of this document.

8.1.1.5 MSR5 [SERV] On-Site Stock

To ensure the require response and resolution time, Atos will create a specific local spare parts stock to be located on IZUM Site, which will remain the property of Atos, and that IZUM shall keep securely.

Atos has a logistics organization located around a central warehouse (in France) and a few regional logistic bases (in different countries), covering the entire world.

The central store is in the Paris region near Orly (Wissous) and provides operational support by ensuring a standard replenishment of the regional logistics bases in J + 1 and managing the various partners in the supply chain 24/7.

Logistics bases are 24/7 proximity stores dedicated to providing and transporting the spare parts and subassemblies necessary for the rehabilitation of customer systems. For the VEGA project, located in Europe, the central store in France will also play the role of regional logistic base.

All defective parts are returned to a centralized shop in Angers in France for repair when possible. The parts used during troubleshooting are new or reconditioned by an Atos' authorized repairer.

8.1.1.6 MSR6 [SERV] Preventive Maintenance Actions

Preventive maintenance may be required in situation like:

- a high count of MCE errors occurs for memory DIMMs or CPUs
- many retired pages are reported for GPUs.
- IB cables with a high reported failure rate.
- discs with too many defective sectors will be replaced.

In such circumstances, Atos can decide to replace such components.

In the frame of preventive maintenance, Atos can also apply any firmware update recommended by the hardware vendor in respond to bug correction for hardware anomalies according with the vendor's procedures.

However, to be compliant with the provisions made in section 9.2.6 Preventive Maintenance and section 9.2.4 Planned Maintenance of the RFP Annex Technical

Specifications, operations described above that may affect the availability of the system will be performed during the planned maintenance agreed slots.

8.1.1.7 MSR7 [SERV] Data Destruction

To ensure compliance with this requirement, Atos will elect

- Either to apply the "Non-Disk Return option" as part of the Maintenance and Support to ensure that there is no danger of any data being taken off site (In this case, the defective disks will remain on-site and the property of IZUM), or
- to delete all IZUM data before disks are taken off-site. In case a reliable data deletion is not possible (by means of a proper deletion software or deletion process), the device will be rendered unusable (physically destroyed).

8.1.1.8 MSR8 [SERV] Serviceability Constraints

During the Maintenance and Support period, the following should be taken into consideration when hardware intervention is required.

Cooling:

Regarding to cooling system for delivered infrastructure there are no constraints from serviceability point of view as all cooling systems are redundant.

CPU nodes:

Each BullSequana X2410 CPU blade includes 3 x compute nodes. As one compute node is embedded with two other nodes, for any corrective action on a compute node at level of memory DIMMs, CPU, Motherboard, the blade (and therefore the 3 nodes) must be stopped.

GPU nodes:

Each BullSequana X2415 GPU blade includes 2 CPUs and 4 GPUs. Therefore, for any corrective action on Memory DIMMs, CPU, GPU, Mother board, the blade (CPU + GPU) must be stopped.

Service Nodes:

Atos proposal includes a redundant configuration for service nodes.
For any corrective and preventive action which requires to stop the node, there will be no impact on functionality, while only the performance could be deteriorated.

High-Speed Storage:

The hardware configuration for DDN Storage solution is fully redundant.

For any corrective and preventive hardware intervention there are no service constraints.

Large-Capacity Storage:

The large capacity storage is based on several Object Storage Nodes and Object Storage Gateways and their interconnection.
All nodes are interconnected on two different type on networks:

- Network for management
- Network for exchanging data

The internal network is based on 8 x Mellanox SN 2010 switches (18x 25GbE + 4x 100GbE ports per switch).
For the network management switch, any preventive and corrective action which requires to stop the switch will stop the management access for all 22 nodes attached to the switch. That means there is not impact on the performance of system but during the intervention time we lose managing the cluster
For the network exchanging data switch, any preventive and corrective action which requires to stop the switch will stop exchanging data for all 22 storage nodes connected to switch.

Switches IB:

For Level 1 switch, any preventive and corrective action which requires to stop the switch will stop the access to the 24 nodes attached to the switch.
For level 2 switch, any preventive and corrective action which requires to stop the switch will not impact the communication but could deteriorate the performance.

Ethernet network:

The proposal includes different kind of ethernet networks:

- The compute management network. This network is based on 2x Mellanox SN2410 switches.
- The management network for the rest of the nodes. This network will be based on Mellanox 4610 switches.

In both cases the switches equipment will be fully redundant from serviceability perspective.
All Mellanox switches products have 1+1 hot-swappable power supplies and N+1 hot-swappable fans.
For any corrective & preventive action there will be not impact on functionality. The only impact could be on performance.

Connectivity to other data centers. Top-level Ethernet network.

The solution proposed by Atos for external connectivity is based on 2 x CISCO Nexus N3K-C3636C-R (routers) and CISCO Nexus N3K-C3408-S switches (top-level) on redundant configuration.

Cisco equipment are hardware designed with dual-redundant power supplies. CISCO Nexus N3K switches come with Cisco NX-OS software that helps ensure continuous availability
For any corrective & preventive action there will be not impact on functionality.
The only impact could be on performance.

Firmware in Hardware Products:

Atos will apply firmware updates recommended by hardware vendor in response to bug correction for hardware anomalies according to the vendor's procedures. These procedures could be a disruptive or nondisruptive process. For each future upgrade of firmware Atos will inform IZUM.

During the Maintenance and Support period, the followings should be taken into consideration when software intervention is required:

- If the process of installing update & upgrade requires the reboot of the equipment for a short time, then the equipment will be out of service during this time.
- If does not, then the impact will be on performance only.

Atos and IZUM will agree on the most suitable time frame and procedure to minimize the impact on availability of the system.

8.1.1.9 MSR9 [SERV] Escalation Management Process

Technical escalation process between level 1, 2 and 3 is presented in section 9.2.8. Incident Management

Management escalation process is as follows:

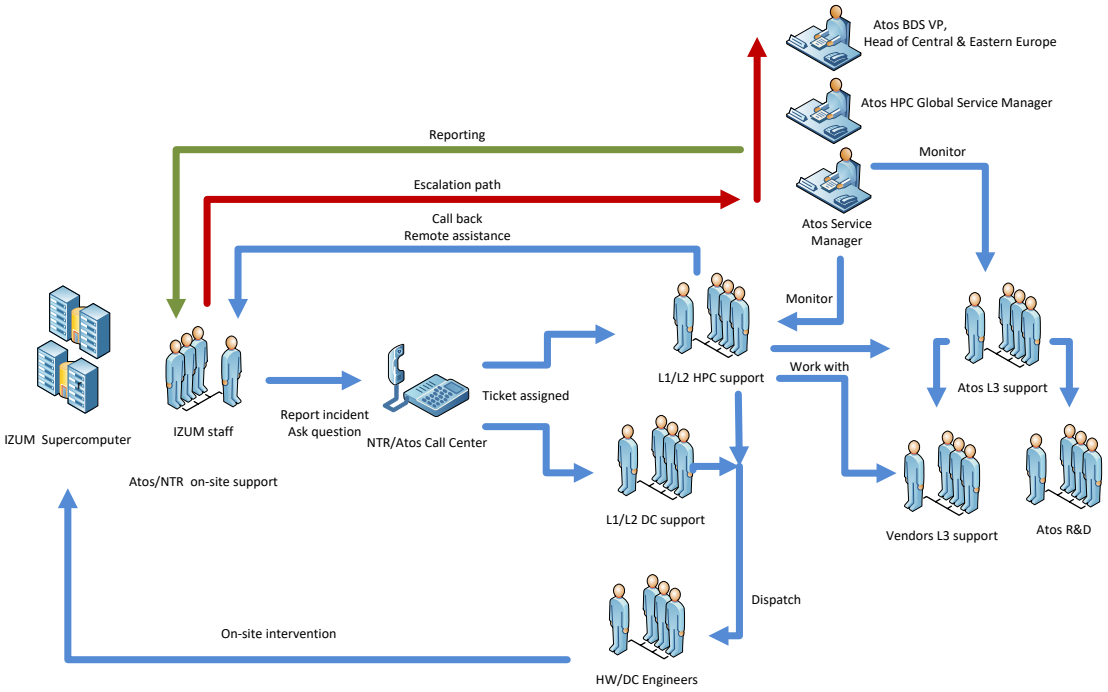
Any default in the above-mentioned technical escalation process can be primarily reported to the Atos Service Manager, who will be appointed during the project implementation.

The second level of escalation will be the Atos HPC Global Service Manager

The third level of escalation will be the Atos BDS VP, Head of Central & Eastern Europe.

Name of the Atos representatives will be communicated during the Maintenance Workshop.

This is summarized in the picture below:



8.1.1.10 MSR10 [SERV] Regular Maintenance and Support Meetings

Atos Service Manager will attend the regular review meetings covering Maintenance and Support (maximum four on-site meetings per year).

Statistics from such activities will be collected, analyzed with the support teams in preparation for the review meetings.

All necessary costs for travel for Atos personnel to attend the on-site meeting will be covered by Atos.

8.1.1.11 MSR11 [SERV] Pre-Production Qualification Test

The system will be able to pass the acceptance tests as proposed in section 11.3.7 of the Annex Technical Specification. Atos maintenance and support team will be on duty to deliver any Maintenance and Support service that may be required during this stage.

8.1.1.12 MSR12 [SERV] Responsibilities and Roles Documentation

In complement to the RACI presented in section 7.1.2.4 of this document, more details are provided here for the Maintenance & Support stage.

As indicated in section 7.1.2.4, and based on definitions presented in section 9.2 and incident management presented in section 9.2.8:

- Atos will remain accountable for the maintenance and support performance offered in this proposition, for all aspects and the required reporting.
- Atos' partner (and subcontractor) NTR will be responsible for the execution of call handling and level 1 support for all products to be delivered in this proposition.
- NTR will also deliver level 2 and 3 and spare parts for infrastructure products delivered for data center adaptations.
- Atos will be responsible for level 2 and 3 and spare parts for all other products to be delivered in this proposition.
- IZUM will be kept informed of all issues and will participate to the regular joint meetings and approve Atos performance.

This is indicated in the table below:

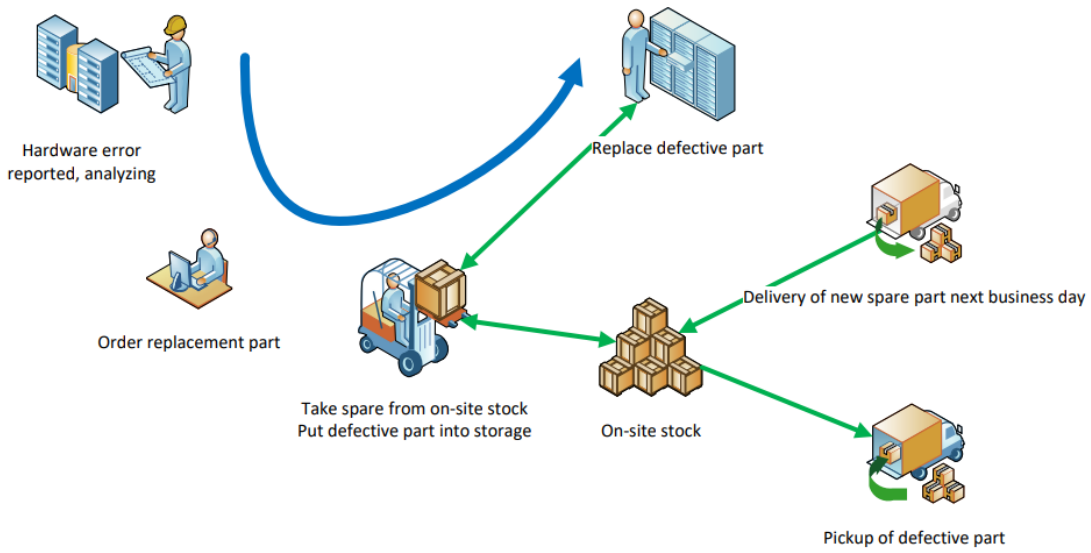
Main Tasks	Atos	Atos' partner NTR	IZUM	Comment
Call handling	A	R	I	
Level 1 for Products(1)	A	R	I	
Levels 2 & 3 and spare parts for Infrastructure Products(2)	A	R	I	
Levels 2 & 3 and spare parts for Hardware & Software Products(3)	A & R	I & C	I	
Reporting	A & R	C & R	I	Atos will be responsible for consolidating all information to build the reporting. Atos' partner NTR will be responsible for data collection and reporting about Infrastructure Products to be consolidated by Atos.
Regular Maintenance & Support Meetings	A & R	C	C	

8.1.1.13 MSR13 [SERV] Maintenance and Support Documentation

Details about the support workflow are presented in section 9.2.8 Incident Management and in section 8.1.1.9 Escalation Management Process, above.

In case a part will have to be replaced, it will primarily be sourced from the local spare part stock, which will then be replenished as indicated in section 8.1.1.5 On-Site Stock, above.

The following picture illustrates the process:



8.2 Licenses

Atos has included in this proposition, the licenses and support (for the duration of the project, i.e. (installation and maintenance periods), corresponding to the following software products.

These software products may include third party developed software and may include open source software.

The license to use Atos software products is governed by Atos terms and conditions.

The license to use such third-party software, even if provided by Atos, shall be subject to the terms and conditions of the third-party software owner(s) which shall in all respects be binding upon the Contracting Authorities.

Product	License tokens for 5 years
Intel Parallel Studio XE Cluster Edition Linux	5 Floating Seats
Rogue Wave TotalView	256 processes
DDN Insight Monitoring Software & ExaScaler Software	Based on included components
Atos Smart Management Center	per node - 5Y subscription

Atos BEO (incl. BDPO)	per node - 5Y subscription
Atos Performance Toolkit	per node - 5Y subscription
Atos IO Instrumentation	per node - 5Y subscription
Singularity Support (Sylabs.io)	per cluster - 5Y subscription
Monitoring, Clary & Inventory software (for Maintenance & Support – Monitoring)	per node – 5Y subscription

According to the Q&A session and more specifically answer dated 28/5/2020 at 14.59, the open source version of Ceph will be used as well as the community driven support.

8.3 Training and Knowledge Transfer

8.3.1.1 TRA1 [SERV] System Documentation Accessibility

Atos will provide system specific documentation describing the configuration, procedures and support workflow as part of the system handover. More details about documentation delivered at each stage can also be found in the description of the stages in answer to requirement "TSC2 [SERV] Time Schedule".

In addition, a set of online generic user-level and system-level hardware and software documents will be supplied.

The documentation will continually be reviewed and updated to reflect any changes in the system configuration during the maintenance and support time frame.

Below is the list of specific documents to be delivered, organized by stage:

Document reference number	Name of the document	Stage
DOC-S1	VEGA Project - Planning & Governance	Created in Stage 1a (updated by joint agreement if required)
DOC-S2	VEGA Project – Risk Register	Created in Stage 1a and updated thereafter until Final Acceptance
DOC-S3	VEGA Project – Document of Technical Architecture	Stage 4
DOC-S4	VEGA Project – Site Parameter List	Stage 4
DOC-S5	VEGA Project - Site Preparation Adaptations	Stage 2
DOC-S6	VEGA Project - On-Site Installation Acceptance	Stage 4
DOC-S7	VEGA Project - Operational Readiness Testing – Provisional	Stage 5

	Acceptance (This will include the Benchmarking Acceptance Report)	
DOC-S8	VEGA Project – Pre-Production Qualification – Final Acceptance	Stage 6
DOC-S9	VEGA Project – Presentation	Stage 8. The document corresponds to the printed material supporting the talks delivered by Atos when supporting the IZUM-organized end-user training
DOC-S10	VEGA Project – Monthly Availability Report VEGA Project – Monthly Preventive and Corrective Maintenance & Support Report	During Maintenance & Support

8.3.1.2 TRA2 [SERV] Training for System Administrators

Atos will organize a 3-day training on-site at the end of stage 4 "On-Site Installation".

This training will cover:

- Global system architecture
  - Hardware components
    - Atos servers
    - Network components
    - Rack layout
  - Network topology
- System installation and configuration
  - How to check system availability
  - Start/stop procedures
  - Nodes update and deployment procedures
- System usage from a user prospective:
  - How to compile an application
  - How to launch an MPI application
  - How to use the batch system

The training will be conducted by an Atos HPC expert. A maximum of eight attendees is recommended to ensure the best experience.

Besides, the maintenance workshop will cover the Atos support procedures:
How to raise an issue/a bug
Which information to collect when reporting an issue.

8.3.1.3 TRA3 [SERV] Training for Users

Atos will participate to the introductory course organized by IZUM for its user community the first year.

Atos will deliver a talk about system-specific application tuning. This could cover basic programming environment and application porting, as well as addressing advanced application performance optimization and tuning.

This will take place on IZUM Site, for a maximum of 2 consecutive days, once a year, during the maintenance and support time frame (5 talks in total). The parties shall agree on the timing of the sessions at least 2 months in advance.

This can address, at the choice of IZUM, either:

- the basic programming environment and application porting
- advanced application performance optimization and tuning for the system

The talk will be delivered by an expert from the Atos Center of Excellence for Parallel Programming (CEPP).

CEPP is made up of a pool of application experts whose mission is to optimize the efficiency of our customers' production. It thus offers services ranging from the handling of a computer to the provision of porting, optimization and code parallelization services.

Atos CEPP also benefits from the close cooperation of our technology partners such as Intel, AMD, Nvidia, Mellanox, or DDN. This allows the team to be constantly trained in the latest technologies (software, computing technologies, networks and storage) and to offer early phase access to our customers.

CEPP experts can advise and assist in analysing, optimizing and carrying specific codes. This could include:

- proof of concept to demonstrate performance gains;
- workshops to learn from experts before porting, optimizing and accelerating customer’s simulations;
- application and solution benchmarks;
- porting, optimizing and configuring customer’s applications.
- tailor-made training courses.

8.4 Cooperation

8.4.1.1 COL1 [SERV] Cooperation Topics

The Atos proposal aims to provide a set of technology innovation and activities with our subject matter experts and technology partners that complements the skills of the IZUM team and focuses on areas of joint interest, taking into account the broad list of potential areas of collaboration presented in the introduction to the RFP.

Atos has successfully implemented this model with several customers around the world, and we believe that it is very well suited to the goals and aspirations of IZUM.

Atos is proposing the following collaborative activities. Atos & IZUM shall agree about which activity they will select and define the corresponding implementation detailed plan:

1. PhD Studentship in Energy Efficient Computing, or
2. Next Generation Genomics Architectures, or
3. Accelerated and AI-led Medical Imaging

Besides, Atos summarizes additional activities, which, for some of them, are already proposed in reply to other requirements of this RFP.

PhD Studentship on Energy Efficient Computing

To work on the projects described below Atos will provide support for a PhD student at one of the IZUM partner universities. However, rather than providing just funding, we propose to enhance the student's experience by actively working with them. Subject to agreement with the host university, we will co-supervise the PhD student to work on a project on energy efficient computing and deep learning related to HPC applications. Specifically, we wish to explore the integration of AI techniques within applications to dynamically make use of energy profiling information so that over time applications can 'learn' how to run in the most energy efficient way.
In addition, we will provide:

- Secondments to Atos for periods of time to work with the HPC and Big Data practice, providing relevant industrial experience.
- Linking existing and new PhD students with Atos customers – e.g. CFD work with engineering companies and Nuclear physics work with our customers in nuclear industries.

Next Generation Genomic Architecture

Though the Atos HPC, AI and Quantum Life Sciences Center of Excellence we offer the following added value project.

Atos has developed a reference architecture for Genomics applications which is now widely used within hospitals and research labs. Elements of this architecture are constantly being evolved and we are currently looking at different techniques for accelerating genomics and any other omics technique. NVIDIA is one of the partners of our Center of Excellence.

With NVIDIA we propose to work with IZUM to deploy:

- Nvidia Toolkit for Bio-informatics – utilizing Nvidia's optimized applications for long read sequencing enable acceleration of DNA analysis
- Nvidia Parabricks – accelerate GATK4 application by 30X over CPU, utilizing standardized BROAD pipeline for short read sequencing.

These development toolkits would be supplemented with appropriate training to ensure that they can be effectively deployed. Specifically, we will run workshops to support the acceleration of healthcare and life science applications.

Deep Learning accelerated Medical Imaging

Though the Atos HPC, AI and Quantum Life Sciences Center of Excellence we offer the following added value project.

Deep Learning is now established as a cognitive computing technique based on deep neural network technology. It has broad applicability from automated voice recognition to medical imaging where it can be used, for example, to predict the likelihood of developing breast cancer from mammography scans. The Deep Learning process is described in Figure 1 below.

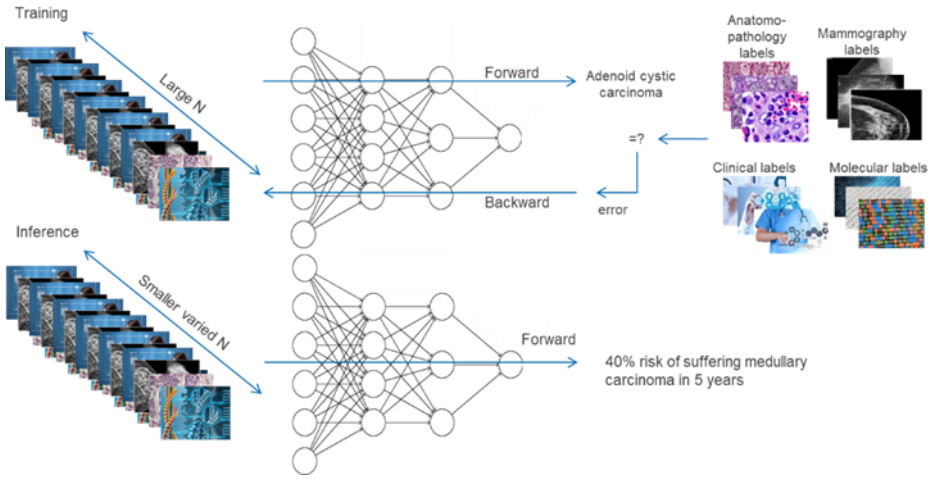


Figure 1 - Deep Learning Process applied to the breast cancer prediction (fictional use case)

The Development Stage, which is where the system is trained, tested and refined using test data to produce a reference data set is a very compute intensive task, requiring access to the HPC computing resources. However, the Production Stage, where use cases can be tested against the reference data set to spot anomalies can be carried out in real time on edge devices, making it ideal for a clinical environment.

Jointly with NVIDIA we will provide IZUM with Nvidia Clara for Medical Imaging – accelerating AI for medical imaging by deploying the most advanced toolkit for AI in medical imaging. Clara has three main aspects:

- Clara Train: A domain optimized developer application framework that includes APIs for AI-Assisted Annotation, making any medical viewer AI capable and a TensorFlow based training framework with pre-trained models to kick start AI development with techniques like transfer learning, federated learning and AutoML
- Clara Deploy: An extensible reference development framework that facilitates turning AI models into AI-powered clinical workflows with built-in support for DICOM communication and the ability to interface with existing hospital infrastructures.
- Embedded Devices: To enable the creation of new devices, Clara AGX enables medical devices that need the ability to perform real-time AI and advanced image, video, and signal processing.

Atos propose to work with IZUM to look at how AI can be incorporated into medical imaging models using Edge devices based on Atos BullSequana Edge systems.

We will provide access to BullSequana Edge devices and with Nvidia will provide training on the use of Clara and on broader Deep Learning topics. This will be done by Atos teams and the Nvidia Deep Learning Institute.

- The Deep Learning Institute (DLI) will provide education and training program to support the development of an AI Ambassador for IZUM, this education program can focus on specific areas from CUDA coding to Medial Imaging algorithm development. Once certified this data science Ambassador will then be qualified to teach DLI training to others, broadening the ecosystem.
- IZUM will build on its skills in the emerging world of cognitive computing establishing itself as the leading site for Deep Learning and the adoption of broader Artificial Intelligence techniques to solve industry problems. As Deep Learning techniques are particularly strong for image processing (spotting anomalies in a series of images such as breast scans) then we see this as a key mechanism for driving collaborations in life sciences and with health care providers who would like to benefit from the technology but do not have the relevant skills to deploy it currently.

Other Activities

Besides R&D cooperation topics, other forms of cooperation will take place.

The ones corresponding to IZUM requirements in other sections of the RFP:

- During the 5-year Maintenance & Support period, Atos CEPP experts will deliver up to 10 man-days per year for helping IZUM improve the production efficiency.
- For the first year after Final Acceptance, as described in section 8.3.1.3 TRA3 [SERV] Training for Users, Atos CEPP experts will deliver a talk about system-specific application tuning, which could cover basic programming environment and application porting, as well as addressing advanced application performance optimization and tuning.
- IZUM and Atos will organize regular meetings (one per semester) in a place to be agreed upon the parties to cover the presentation of new products and services, and discussions about cooperation. The frequency of such meetings could be different upon agreement between the parties.

Additional ones in complement:

- Atos will propose to IZUM representatives to attend events either organized by Atos or for which Atos is a sponsor and/or an exhibitor.
- As the owner of an Atos system, IZUM can become a BUX Member. The BUX (Bull User group for eXtreme computing) is an independent world-wide group of users that will cooperate to increase the capabilities of large-scale, parallel scientific and technical computing supplied by Atos, to promote the exchange of information and understanding of these systems, and to provide guidance to Atos on the essential development and support issues for large-scale technical systems.
  BUX was founded on June 20th, 2011 at ISC'11 in Hamburg, as a result of discussions that took place in Paris first half of 2011 between users of Atos extreme computing solutions. The objectives of the BUX are:
  - Sharing of experience between Members and with Atos
  - Providing Atos with inputs and helping set priorities for Atos (technology, service and support …)
  - Networking with Atos HPC experts during BUX events
  - Highlighting emerging technologies and helping foster long-term developments
  - (more about BUX can be found here: https://atos.net/en/solutions/high-performance-computing-hpc/bux)

8.4.1.2 COL2 [SERV] Cooperation Resources

The resources for the implementation of the activities in COL1 [SERV] Cooperation Topics are included in this proposition.

9 Services

9.1 Installation (Including project plan) of the supercomputer

9.1.1 Installation Plan

The plan is more specifically described in section 7.1.2.2 of this document. It includes the training sessions, which are more specifically described in section 8.3.1.2 and 8.3.1.3.

There is no additional hardware or software components to be supplied by IZUM for the supercomputer to operate, except of course any application software, which are outside the scope of this RFP.

9.1.2 Hardware and Software Installation

As presented in the section 7.1.2.2, site adaptations and hardware and software delivery are included in the plan. Provision is made for parallelism between tasks and deliveries in batches (or stages) if required. Such batches could concern services nodes, compute partition, graphic partition, high-speed storage or high-capacity storage.

9.1.3 Provisional Acceptance

Description of this stage is described in section 7.1.2.2 (Stage 5 Operational Readiness Testing (Provisional Acceptance)).

9.1.4 Final Acceptance

Description of this stage is described in section 7.1.2.2 (Stage 6 Pre-Production Qualification (Final Acceptance)).

9.2 Maintenance and Support of The Supercomputer

Atos describes below the organization and processes to deliver the SLA required by IZUM.

In this section, the following definitions will apply:

- Infrastructure Products, Hardware Products and Software Products mean respectively the equipment for data center modification (especially power supply and cooling), the hardware (especially networking, compute and storage elements), the software with their Documentation proposed by Atos and supplied to IZUM, to be delivered and installed by Atos.

- Product(s) means Infrastructure Product(s), Hardware Product(s) and/or Software Product(s).
- Firmware (or microcode or machine code) means a series of instructions or data designed for operating at or accessing a level below the external user interface. Firmware is incorporated in the components of specific Hardware Product.
- Documentation mean the document including detailed description of the Product and other technical, user and training documentation intended to facilitate the installation, testing, use, maintenance and support of the Product including any modifications, revisions or new releases.

For IZUM, Maintenance of the Infrastructure Product(s) and Hardware Product(s) and Support of Software Product(s) will, during the requested Maintenance and Support Period (5 years from the Final Acceptance date), includes the Warranty as described below, as well as additional services, to be provided under a defined service level and according to conditions as defined below.

Warranty

Warranty of Infrastructure and Hardware Products

Atos warrants that, for the Maintenance and Support Period each Hardware Product shall be free from defects in material and workmanship and shall conform to its Documentation.

Atos's entire obligation and IZUM's exclusive remedy under the aforesaid warranty (at Atos's reasonable option) are to repair any defective Infrastructure or Hardware Product or defective component of the Infrastructure or Hardware Product or to replace them with new or an equivalent.

When Atos elects to replace the component found to be defective, which corresponds to its mainstream standard process, Atos will use only new or refurbished but guaranteed parts of equivalent performance. The replaced component will become the property of Atos, while conversely, the replacing spare parts will become the property of IZUM as from the date of signature of the field intervention report.

The above warranty does not apply to defects (i) caused by normal use and wear of the Infrastructure or Hardware Product (ii) caused by the use or operation of the Infrastructure or Hardware Product (or any part thereof) other than as intended or recommended by Atos in the Documentation (including improper installation by IZUM should IZUM decides to be in charge of the installation), (iii) caused by modifications or alterations made to the Infrastructure or Hardware Product (or any part thereof) by anyone other than Atos or its subcontractor, (iv) which are the result of the Infrastructure or Hardware Product (or any part thereof) being subjected to unusual physical or electrical stress, flood, fire, neglect, abuse, accident, misuse, (v) caused by

other IZUM's systems connected with the Infrastructure and Hardware Product and shall not apply (vi) if the serial number(s) has(have) been removed or altered.

Consumable parts, such as batteries etc., are excluded from the aforesaid warranty.

Warranty of Software Products

Atos warrants that the Software Product will substantially conform to its Documentation at the time of delivery from Atos.

This warranty does not cover defects arising:

- from any modification by IZUM or a third party of the Software Product (unless made with Atos' written consent);
- from operator error or misuse of the Software Product or from any hardware product or any fault in any hardware product or any software used in conjunction with the Software Product;
- use of a Software Product release older than the release level currently supported unless otherwise specified in writing.

Atos represents that it has authority to grant IZUM the license or sub-license for the Software Product delivered to IZUM.

Atos hereby expressly states that it does not give any other warranty, whether express or implied, with regards to the Software Products including any warranty of merchantability or fitness for a particular purpose. Atos does not warrant that the operation of the Software Products shall be error free or uninterrupted. Atos further gives no warranty whatsoever about third-party software (including but not limited to open source software) if they are provided "as is" and without warranty.

Maintenance & Software Support

Atos shall provide Maintenance for the Infrastructure and Hardware Products and Software Support to the Software Products supplied by Atos to IZUM with the Level of Service defined hereinafter.

Description of Maintenance & Software Support

Under the Maintenance, Atos shall provide corrective maintenance and preventive maintenance of Infrastructure and Hardware Product, as well as Firmware updates, according to Atos' procedures.

Under the Software Support, Atos shall provide corrective and preventive maintenance through applying a patch or through an update of the current version of Software Product.

Conditions for the performance of the Maintenance and Software Support

To allow Atos to properly perform the Maintenance and Software Support, IZUM will:

- provide access to its Site to Atos personnel and Atos' subcontractors as may be reasonably necessary to maintain or support the Infrastructure, Hardware or Software Products
- ensure protection of its data and software as may be necessary and undertake, but not limited to, all necessary data and software backup before Atos' intervention;
- ensure that it uses up to date virus detection and eradication tools, and observes good practice to prevent viruses from contaminating its IT environment
- allow Atos to perform updates and releases on Products when requested
- describe in detail any errors and problems detected with the Product(s) and provide any information requested by Atos on such errors and problems and comply with all instructions given by Atos;
- authorize and facilitate remote access to the Products and make them available for the time needed to complete the Maintenance and Software Support.

Atos will have to perform its obligations in respect of:

- Slovenian labor laws
- IZUM rules and regulations
- Slovenian public holidays

Limitations/Exclusions

The Maintenance and Software Support shall not include maintenance and support of the Product required:

- in case of defects that are not covered by the Warranty;
- due to any causes external to the Product and the use of the Product with software and/or equipment other than those mentioned in the Documentation or recommended by Atos;
- due to any modification of the law or regulations.

The Maintenance and Software Support shall not cover:

- restoration of any data or files damaged or lost by IZUM or a third-party or due to any external event or factor;
- Any modification of the Product due to any change of the law or regulation except for mandatory modifications of the Product which are required for safety reasons.

The scope of Maintenance and Software Support is limited to the Products and therefore excludes in particular:

- products or any other items (hardware or software) which are not included in the contract
- troubleshooting and repair of the communication links and/or equipment provided by third parties (this will remain the responsibility of IZUM or his communications provider).

Assistance such as assistance for configuration and tuning

In accordance with the clarification made during the RFP question & answer session (dated 16.06.2020 at 10:22), Atos has included in its proposition a provision for 10 man-days per year during the 5-year Maintenance & Support period for such services.

In order to ensure the availability of the most appropriate expert, IZUM and Atos will agree at least 2 weeks in advance about the type of service required and the date of its execution.

9.2.1 SLA of Maintenance and Support Services

To satisfy IZUM's requirements, Atos has adapted its standard service level, as more specifically detailed in the sections below.

Support Severity Definitions:

The list below describes the severity levels to define the impact and severity when registering an incident:

- Severity 1: System blocked: severe impact to business operation, system is unavailable
- Severity 2: System degraded: degraded performance or interruption in system availability
- Severity 3: Minor impact or question: No or minor impact on business operation, system available

Intervention & Resolution of Incidents	
Diagnostic assistance hotline period	9am-6pm Mon-Fri (CET)**(1)**
On-Site Service period	9am-6pm Mon-Fri (CET)**(1)**
Intervention Time: on-site or remote service time for a Severity 1 incident (blocked system)	4 working hours from incident registration(2)
Intervention Time: on-site or remote service time for a Severity 2 or 3 incident	4 working hours from incident registration(3)

Recovery Time: using on-site or remote service for a Severity 1 incident (blocked system)	Less than 1 working day from incident registration for at least 90% of the reported incidents(4)
Recovery Time: using on-site or remote service for a Severity 2 or 3 incident	less than 2 working days from incident registration for at least 90% of the reported incidents(5)
Spare part on-site delivery	Next business day(6)

(1) from 9am to 6pm CET on working days from Monday to Friday excluding public holidays and non-working days, calls notified after 4pm being considered as opened on the next working day.
(2) and **(3)** as required in section 9.2.8 of Annex Technical Specifications of the RFP. The indicated recovery time correspond to the time required to repair or bypass the incident.
(3(4) as required in section 9.2.1 of Annex Technical Specifications of the RFP for "blocking failure", i.e. Severity 1 incidents
(5) as required in section 9.2.1 of Annex Technical Specifications of the RFP for the "non-blocking failures", i.e. Severity 2 and 3 incidents
(6) Atos will create a specific local spare parts stock to be located on IZUM Site, as indicated in section 8.1.1.5 of this document

We understand from the question answered on 16 June at 10:37 that SLA refers to all IT equipment and to those parts of the supply infrastructure that are or will be in a highly available configuration. It does not apply to the parts of the supply infrastructure where components are not duplicated. In this case, maintainer of the supply infrastructure must eliminate the failure according to the principles of best effort within a maximum of 14 days from the time of failure.

9.2.2 Availability

Referring to the following sections of the RFP Annex-Technical Specifications

- 9.2 Availability
- 9.2.2.1 Rules to Establish the Overall availability of the Configuration
- 9.2.3 Availability Progress Reports
- 9.2.5 Corrective Maintenance (for the part related to alarms)
- 9.2.8.1 Technical Resolution Engagement
- 9.2.10 Operational Service Quality

Atos understands IZUM willigness to measure the availability of the Supercomputer.

Atos understands that the ability to reach the NWH availability target may be impacted by the unability of Atos engineers to acces IZUM site to fix anomalies.

Atos understands that incidents impacting the availability of the supercomputer, due to failure of components not delivered by Atos shall be excluded from the measurement of the availability.

9.2.3 Availability Progress Report

Atos will install the appropriate software tools to monitor the Supercomputer to measure the requested indicators: software for inventory, monitoring and alert, and for supercomputer availability calculation.

9.2.4 Planned Maintenance

Atos and IZUM will agree on schedule to perform, whenever necessary and upon Atos' request a planned maintenance considering:

- Half a day per month
- No more than two global power maintenances of the computing center each year to carry out interventions on the procured system environments on non-working days

9.2.5 Corrective Maintenance

Atos will set up an alarm system.

9.2.6 Preventive Maintenance

Details are provided in section 8.1.1.6 Preventive Maintenance, of this document.

9.2.7 Call Center and its SLA

Call handling and qualification	
Call Center availability	9am-6pm Mon-Fri (CET)**(1)**
Call Center response time (answering/accepting the incoming call)	15 minutes for 90% of the requests(2)
Specialist call-back time	4 working hours from incident registration(2)

(1) from 9am to 6pm CET on working days from Monday to Friday excluding public holidays and non-working days, calls notified after 4pm being considered as opened on the next working day.
(2) as requested in section 9.2.7 of Annex Technical Specifications of the RFP

9.2.8 Incident Management

Definition of Support Levels

Support Level Definitions:

- Level 0 (L0): Receiving Service Requests (and Recording Incidents)
- Level 1 (L1): Incident Management (and Ownership)
- Level 2 (L2): Problem Management (and Major Incident resolution)
- Level 3 (L3): Change Management (and Release Management)
- Field Service, Dispatching and Spare Part Logistics (FS)

Receiving and Managing Service Requests

The standard process is as follows. It is based on the following steps: validation, registration and interview, checking for existing solution, defining an action plan, implementing the plan (solution and repair), closing and reporting.

Receiving Service Requests and Recording Incidents
Receive Service Requests from customer to the call center and collect information for record in the incident management system: 1. Record and verify equipment serial number, service, customer name, address, area code, telephone number, technical contact name and other relevant information. 2. Verify end user service entitlement and escalate to Atos Service Manager any issues in service verification. 3. Record incident subject and description or add updates on existing tickets.
Record impact/severity and create an incident ticket in the appropriate queue of the ticketing system.
Incident Management
At level 1:
Deliver Level 1 remote technical support service (See description here under).
Analyze incident logs and evidence and provide a resolution to customer using standard problem determination procedures
When applicable, access to Atos HPC known issues documentation and problem determination procedures on Atos service web site to check if a solution is already available.
If repair part(s) are required, list those parts (part posting) and contact local logistic team to request a delivery from local or central warehouses.
If no solution found, transfer incident to the support level 2 with required information from level 1 (problem description, problem source Identification, trace, log, etc...)
Manage communication with customer and report customer escalations to local Atos service manager

At level 2:
Receiving escalated Level 1 Incidents in appropriate L2 queue and record problems in incident management system
Deliver Level 2 remote technical support service
If no solution found, transfer incident to support level 3 with required information from level 2 remote technical support
At Level 3
Deliver level 3 remote technical support service.
Verify and adjust the list of required parts for corrective actions, when applicable
Manage on-site interventions with field service. On site field service will cover: replacement of defective parts, running diagnostic to verify problem resolution and collecting logs if necessary, updating firmware as a corrective action or when provided with instruction to do so by support, communicating technical updates on incident resolution to support
Manage the on-site activity and raise an alert to local Atos Service Manager if there is a risk on Service Level fulfillment.
Provide Level 2 on-site assistance to field service if required in an action plan validated by Level 2 support expert.
Close call in incident management systems.

Each intervention will be recorded in an intervention form, signed by Atos and IZUM, which shall specify: the location, date and time of the notification, date and time of the intervention, type of equipment (name, manufacturer, model, serial number), the damage ascertained, operations performed for troubleshooting, part(s) that have been replaced.

9.2.8.1 Technical Resolution Engagement

This applies to incidents having an impact on the availability of the system. Not to fall beyond the targeted availability, Atos can promptly act on the system to solve identified incidents. Incidents can be identified by the monitoring tools or alarms or reported by IZUM.

Atos however understands that the SLA described in 9.2.1 will prevail over the one described in this section of the Annex Technical Specifications as indicated in the question answered on 16 June 2020 at 10:21: "In case of discrepancies with other SLA requirements for Maintenance and support, the requirements of section 9.2.1 prevail"

9.2.9 Relation with the Supplier

Atos will appoint for the VEGA project:

- An Atos IZUM Account Executive. He will be the primary interface to IZUM for all contractual and commercial matters. He will be named at contract signature.
- An Atos Project Manager, who will be responsible for the technical delivery of the project from contract signature to the hand over to the maintenance and support

team (final acceptance date). His role and mission is more particularly detailed in section 7.1.1.1 Project Management.
- An Atos Service Manager, responsible for the Maintenance & Support stage of the project. He will be responsible for Atos commitments for this stage.

9.2.10 Operational Service Quality

Atos understands that, for incidents that do not affect availability (for example redundant equipment or equipment with high availability (HA) capabilities, Atos must repair or bypass the incident within 2 working days.

Atos also understand that IZUM may decide to run benchmarks when the system will be configured "in production" mode, and to review the results with Atos.

9.2.11 Technical and Administrative Accountancy

The Atos Service Manager will prepare the:
- VEGA Project – Monthly Availability Report
- VEGA Project – Monthly Preventive and Corrective Maintenance & Support Report

9.3 Training and Knowledge Transfer

Details of the trainings are provided in sections 8.3.1.2 and 8.3.1.3 of this document.

Details about the documentation are provided in section 8.3.1.1 of this document.

9.3.1 Documentation

Details of the documentation are provided in section 8.3.1.1 of this document

9.4 Risk Management

Risk management is presented in section 7.1.2.3 of this document

A description of roles and responsibilities in terms of "RACI" is presented in section 7.1.2.4 of this document

9.5 Dismantling of the Supercomputer

Atos will provide on IZUM's request the procedure to dismantle the supercomputer.

10 Benchmarks and Performance

Atos has included in this proposition a Benchmarking Report, which contains committed results with clear description on formulas and process reaching the results.

Atos will run the defined benchmarks to reproduce the committed performance values on the installed supercomputer at IZUM, as part of the Stage 5 Operational Readiness Phase to reach the Provisional Acceptance.

A full explanation of the analysis which leads to the committed benchmark performance is included in the benchmark report. The extrapolations rely on the measures undertaken on Atos benchmark machines and those of our partners which are like, though not identical with, the systems proposed in their processor, memory performance, and system interconnect.

Atos will run the benchmark in a presence of IZUM through a management console confirming start of run and final result. All runs will be documented with log files and stored on the system. A proper documentation will be built with those logs presented on a media as part of the Benchmarking Acceptance Report. Atos will not erase the logs on the system without IZUM approval.

Atos usually proposes a method whereby the Atos Project Manager is present on site and the Atos benchmark team remotely connected to the supercomputer. Other setups are possible and subject to prior agreement between IZUM and Atos.

10.1 Introduction

This report describes the technical analysis, experiments and the commitments made for the Authority's benchmark suite of applications, performed by application experts from the Atos Applications and Performance Group.

The Applications and Performance Group is headquartered in Grenoble, France, with offices in the UK, Germany and India. The team consists of approximately twenty engineers with backgrounds in different scientific fields including physics, chemistry, and applied mathematics.

The team provides application performance expertise in the contexts of performance focused tenders, porting and optimization of key applications and benchmarks, and undertaking commercial performance focused software engineering projects. The team is also responsible for running and publishing results of standard benchmark suites to promote Atos solutions.

10.2 Configuration of benchmark supercomputer

10.2.1 Benchmark system configuration

The benchmarks have been run by applicative experts of the "Applications & Performance Team" on bullx supercomputers hosted on our premise. The specifics of the supercomputers used are described in the following tables:

- AMD Rome 7742 (BULL/Spartan)

Compute nodes	AMD Rome / Spartan
Processor	AMD Rome 7742
CPU Nominal Frequency	2.25 GHz
CPU TDP	225W
# cores per socket	64
# sockets per node	2
# cores per node	128
# Flops per cycle	16 (AVX2)
Memory size	256 GB
Memory type	DDR4@ 3200 MT/sec
# memory canals per socket	8
Parallel file system	Lustre (DDN Exascaler 7990)
Interconnect	HDR100 – Infiniband
Operating System	BULL SCS AE 5 (RHEL 7.6)
# Nodes	240

- Volta V100 - Skylake 6148 (BULL/Spartan)

Compute nodes	Volta V100 - Skylake 6148 / Spartan
Processor	Intel Skylake 6148
CPU Nominal Frequency	2.4 GHz
CPU TDP	150W
# cores per socket	20
# sockets per node	2
# cores per node	40
# Flops per cycle	32 (AVX512)
Memory size	384 GB
Memory type	DDR4@ 2666 MT/sec
# memory canals per socket	6
GPU Type	Nvidia Tesla Volta V100 32GB SXM2
# GPU per node	4

Parallel file system	Lustre (DDN Exascaler 7990)
Interconnect	HDR100 – Infiniband
Operating System	BULL SCS AE 5 (RHEL 7.6)
# Nodes	16

- Ampere A100 - Rome 7742 (Nvidia/Selene)

Compute nodes	Ampere A100 – Rome 7742 / Selene
Processor	AMD ROME 7742
CPU Nominal Frequency	2.25 GHz
CPU TDP	225W
# cores per socket	64
# sockets per node	2
# cores per node	128
# Flops per cycle	16 (AVX2)
Memory size	2048 GB
Memory type	DDR4@ 3200 MT/sec
# memory canals per socket	8
GPU Type	Nvidia Tesla Ampere A100 40GB SXM4
# GPU per node	8
Interconnect	8x HDR200 – Infiniband
# Nodes	240

10.2.2 Proposed solution and extrapolation method

The scope of this section is not to give a detailed description of the solution, rather we will give a simple overview of our proposed solution. It is composed of two partitions:

- A standard compute node partition composed of 960 AMD Rome 7H12 dual socket nodes. The AMD Rome 7H12 is a 64 cores processor running at a nominal frequency of 2.6 GHz.
- An accelerator partition composed of 60 GPU nodes. Each node has two AMD Rome 7H12 CPUs and 4 Nvidia Ampere GPUs, the GPUs are connected to each other via Nvidia SXM3 links and connected to the CPUs via PCI Gen 4 links.

Since the Compute node partition is essentially a larger version of our benchmark cluster (processors varying by their frequency), we do not have to correct benchmark performance considering memory bandwidth. Ours benchmark performance extrapolations are simply based on frequency projections.

As for our Accelerator node projections, we have when possible used different techniques.

To estimate the behavior when we increase the number of GPUs, we have used multi-node executions on our Nvidia Volta GPUs.

To estimate the single node performance, we have used two techniques:

- if possible, we have executed the code on our single AMD + Nvidia Ampere test node to get the performance for a single node.
- If not possible, we have executed a GPU frequency dependency analysis of the benchmark on our Nvidia Volta nodes and used this information to project on Nvidia Ampere. Considering that the GPU frequency sensitive part of the GPU execution time would be corrected by:
  - The ratio of frequency between Nvidia Volta and Nvidia Ampere GPUs,
  - The ratio of streaming multi-processors between Nvidia Volta and Ampere GPUs, this assumption considers that the frequency sensitive part is fully independent of the GPU memory bandwidth is a completely scalable.
  - Finally, the non-GPU frequency sensitive part of the GPU execution time would be considered to be related to the GPU memory bandwidth, and thus corrected by the ratio between the Nvidia Volta and Ampere memory bandwidths.

10.2.3 Summary of our projections

In this section we present a summary of our performance projections for the proposed solution.

We will use all compilers, libraries and tools available for the acceptance, to reach our commitments respecting all your constraints described in the RFP.

We would like to point out that the performance linked to the benchmark commitments will be reproduced with a tolerance of 5% on average for all benchmark/code test cases.

Benchmark details					Tenderer fulfilment	
Req. No.	**Specification**	**Priority**	**Better high/low**	**Weight (b_i)**	**Value**	**Unit**
HPL1	**HPL CPU benchmark extrapolated R_{max} [PFLOPS]**	High (normalization)	H	4	**3.801**	**PFLOPS**
HPL2	**HPL GPU benchmark extrapolated R_{max} [PFLOPS]**	High (normalization)	H	4	**3.000**	**PFLOPS**
HPL3	**Theoretical R_{peak} for compute and GPU nodes [PFLOPS]**	Doc (validation)	N/A	N/A	**N/A**	**N/A**
HPG1	**HPCG CPU benchmark extrapolated R_{max} [PFLOPS]**	High (normalization)	H	4	**0.0476**	**PFLOPS**
HPG2	**HPCG GPU benchmark extrapolated R_{max} [PFLOPS]**	High (normalization)	H	4	**0.0689**	**PFLOPS**

STR1	**STREAM single processor benchmark for Compute partition [MB/s] {COPY}**	High (normalization)	H	0,25	**176 522**	**MB/s**
STR1	**STREAM single processor benchmark for Compute partition [MB/s] {SCALE}**	High (normalization)	H	0,25	**177 181**	**MB/s**
STR1	**STREAM single processor benchmark for Compute partition [MB/s] {ADD}**	High (normalization)	H	0,25	**176 868**	**MB/s**
STR1	**STREAM single processor benchmark for Compute partition [MB/s] {TRIAD}**	High (normalization)	H	0,25	**176 970**	**MB/s**
STR2	**STREAM single processor benchmark for GPU partition [MB/s] {COPY}**	High (normalization)	H	0,25	**176 522**	**MB/s**

STR2	**STREAM single processor benchmark for GPU partition [MB/s] {SCALE}**	High (normalization)	H	0,25	**177 181**	**MB/s**
STR2	**STREAM single processor benchmark for GPU partition [MB/s] {ADD}**	High (normalization)	H	0,25	**176 868**	**MB/s**
STR2	**STREAM single processor benchmark for GPU partition [MB/s] {TRIAD}**	High (normalization)	H	0,25	**176 970**	**MB/s**
IOR1	**IOR benchmark for LCST partition [GB/s]**	High (normalization)	H	1	**207.6**	**GB/s**
IOR1	**IOR benchmark for HPST partition [GB/s]**	High (normalization)	H	1	**400**	**GB/s**
GRO2	**GROMACS Water GMX50 benchmark for CPUs [ns/day]**	Very high (normalization)	H	10	**11 658.60**	**ns/day**
GRO4	**GROMACS Water GMX50 benchmark for GPUs [ns/day]**	Very high (normalization)	H	10	**817**	**ns/day**

QES1	**Quantum ESPRESSO benchmark for CPUs [s]**	Very high (normalization)	L	10	**18.77**	**s**
QES2	**Quantum ESPRESSO benchmark for GPUs [s]**	Very high (normalization)	L	10	**7.43**	**s**
HEP1	**HEPSpec06 benchmark for CPUs [total score]**	Very high (normalization)	H	20	**3 162 000**	**total score**
FOA1	**OpenFOAM benchmark for CPUs [s]**	Very high (normalization)	L	10	**3.948**	**s**
TEN1	**TensorFlow Resnet-50 for GPUs [images/s]**	Very high (normalization)	H	10	**294 710**	**images/s**

10.3 Synthetic Benchmarks

10.3.1 HPL

High-Performance Computing LINPACK Benchmark (HPL) is a software package that solves a (random) dense linear system in double precision (64 bits) arithmetic on distributed-memory computers.

For getting an efficient HPL we run with optimized binaries shared by our partners. For the CPU HPL, it is possible to run an HPL using CPUs from Compute and GPU partitions. Thus, we commit with CPUs from both partitions.

Name	Number of nodes	Performance [Pflop/s]
HPL1	1020	**3.801**
HPL2	60	**3.000**

HPL: Commitments

As requested, we give the theoretical R_{peak} CPU performance for Compute and GPU partitions as well as R_{peak} for graphic subsystems double precision (FP64) performance for GPU nodes.

Name	Number of nodes	Performance [Pflop/s]
HPL3 - CPU	1020	**5.431**
HPL3 - GPU	60	**4.680**

HPL: R_{peak}

10.3.2 HPCG

The High-Performance Conjugate Gradients (HPCG) Benchmark project is an effort to create a new metric for ranking HPC systems. HPCG is designed to exercise computational and data access patterns that more closely match a different and broad set of important applications, and to give incentive to computer system designers to invest in capabilities that will have impact on the collective performance of these applications.

HPCG is a complete, stand-alone code that measures the performance of basic operations in a unified code with sparse matrix-vector multiplication, vector updates, global dot products, local symmetric Gauss-Seidel smoother and sparse triangular solve (as part of the Gauss-Seidel smoother).

The HPCG benchmark is limited by the memory bandwidth of a system and therefore the reached FP performance is typically rather low. For the target CPUs a value of 55 GFlops is expected for a single compute node. The GPU performance at scale is expected at 287 GFlops.

Name	Number of nodes	HPCG [Pflop/s]
HPG1	1020	**0.0476**
HPG2	60	**0.0689**

HPCG: commitments

10.3.3 STREAM

STREAM is a synthetic benchmark program, written in C or Fortran, that measures sustainable memory bandwidth in MB/s and the corresponding computation rate for simple vector kernel.

We used the AOCC compiler to get our best performance. As CPUs from Compute and GPU partitions are the same, commitments are equal.

Name	COPY [MB/s]	SCALE [MB/s]	ADD [MB/s]	TRIAD [MB/s]
STR1	**176522**	**177181**	**176868**	**176970**
STR2	**176522**	**177181**	**176868**	**176970**

STREAM: commitments

10.3.4 IOR

IOR is a parallel IO benchmark that can be used to test the performance of parallel storage systems using various interfaces and access patterns.

10.3.4.1 LCST

All units of throughput depicted here are based on powers of 10 (1 MB/s = 1 000 000 bytes/s; 1GB/s = 1000 MB/s).

The proposed system can sustain a throughput of **207 GB/s**.

The median throughput performance of a single OSD drive is taken as the basis of this calculation.

The median performance of a single drive is considered the median value between its maximum throughput (obtained when accessing the outer most sectors of the drives) and its minimum throughput (when accessing the inner most sectors, which is usually half of the maximum throughput). The maximum throughput for the ST16000NM008G being 260MB/s, the minimum is 130MB/s and median performance is (260+130)/2 = 195MB/s. The raw throughput provided by the system is then 61 servers * 24 drives * 195 MB/s resulting in 285.48GB/s. However, considering the erasure coding configuration of 8+3, the useful throughput is actually 285.48 * 8 (effective data) / 11 (real written data) = 207.6 GB/s.

10.3.4.2 HPST

10.3.4.2..1 Hardware & software configurations for benchmarking

- ES400NVX
  - o 8 x InfiniBand HDR100 Uplinks
  - 4 x vOSS/vMDS with EXAScaler 5.1
  - 20 x 3.84TB NVMe – in 2 x DCR-Pool 10/0 with RAID6 8+2p
  - four MDTs
  - four OSTs
  - chunksize of 128kiB (FullStripeSize of 1024 kiB)

- (up to) 40 x Client Compute Node
  - 2 x Xeon 14-Core CPU @2.0 GHz (E5-2660 v4)
  - 128GB RAM
  - InfiniBand HDR100 Uplink

All tests were run on EXAScaler 5.1 and SFA-OS version 11.8.0. These were the latest stable versions available.

10.3.4.2..2 Offered configuration

- 10 x ES400NVX each with:
  - 8 x InfiniBand HDR100 Uplinks
  - 4 x vMDS/vOSS with EXAScaler 5.1
  - 21 x 7.68TB NVMe – in 2 x DCR-Pool 10/0 with RAID6 8+2p + 1 HotSpare
  - four MDTs
  - four OSTs
  - chunksize of 128kiB (FullStripeSize of 1024kiB)

The benchmarked configuration is matching a single building block of the offered configuration with the following differences:

- The benchmarked configuration is using 3.84TB NVMe devices while the proposition contains 7.68TB devices. The larger devices are at least as fast as the smaller devices. This difference is not relevant for any extrapolation of performance.

- The benchmarked system did not contain a Hot Spare device – which is not relevant for the achieved performance. This difference can be ignored.

10.3.4.2..3 Results and extrapolations

The following sections are describing the benchmarks that were done as well as the extrapolations done for the offered configurations.

Unless stated differently in the individual sections all performance extrapolations are made with expecting the following pre-requisites hold true:

- The number of client nodes and processes per client are set to show optimal system performance.
- The number of client nodes and processes per client are set individually for each of the separate tests to achieve best possible performance.
- All files might be pre-created prior to any performance test to achieve optimal data placement on all targets and show optimal system performance.
- Performance tests are run as "file per process" with each client process writing to and reading from a separate file unless explicitly stated otherwise.
- Filesystem tuneable (such as stripe setting and chunksizes) are adjusted individually per benchmark to achieve best possible performance.
- Network checksums between clients and servers are disabled
- Flash based file systems are "trimmed" before each benchmark run to achieve best performance.
- Benchmark results are obtained in a "best of five" mode.
- Since performance of filesystems depend on the ageing process of the filesystem benchmarks are done on a newly delivered and empty filesystem.
- All results were achieved with InfiniBand HDR100 links between servers and clients. For all extrapolations the expectation is that a fully tuned and error-free InfiniBand fabric with at least the same speed is available.
- For all extrapolations the expectation is that the only workload running on the fabric is the benchmark itself – other than that the fabric must be empty.

The IOR benchmark was run on the previously described benchmark configuration with the results that follows.

The run-rules specifically asked for a single MPI rank per used compute node. This imposes a challenge on the filesystem to achieve the high performance with a relatively low number of MPI ranks (expecting in the range of 750 compute nodes) while maximum performance usually is achieved with a far higher number of MPI ranks accessing the storage concurrently.

A second challenge is the relatively small transfer size of 64kiB. Although all transfers are sequential and get aggregated into full RPCs on the client side before submitting it to the server nodes this imposes an overhead.

To get a good understanding for the performance we did run several different iterations. The below table shows the Write and Read results for a scaling number of MPI ranks (nodes):

IOR xfer=64K, FPP, sequential		
NN=NP	Write (MB/s)	Read (MB/s)
1	1419.24	3575.85
2	2756.59	7030.68
4	5481.13	14541.91
8	10958.47	28965.76
16	21717.29	39536.34
24	32386.06	46135.72
32	40427.44	47784.73
40	41070.16	48077.53

We observe the following things:

- A single stream does yield a performance of about 1400 MB/s writes and 3500 MB/s reads. We will focus on the write figures in the following since they seem to be the limiting factor
- Performance scales (practically) linear within the number of MPI ranks (nodes) until 24 – 32386 MB/s aggregated results in 1349 MB/s for each rank
- With 32 and more ranks the performance per task goes down as the storage building block reaches its maximum performance
- The highest performance of the building block is achieved with all 40 available client nodes. The achieved 41 GB/s in writes and 48 GB/s in reads match the maximum performance that the hardware allows exactly

Here are the details of the full run for 40 MPI ranks:

```
IOR-3.1.0: MPI Coordinated Test of Parallel I/O
------------------------------------------------------------------------
WARNING: There was an error initializing an OpenFabrics device.
Local host: ec38
Local device: mlx5_0
------------------------------------------------------------------------
IOR-3.3.0+dev: MPI Coordinated Test of Parallel I/O
Began : Thu Jun 25 19:51:32 2020
Command line : /work/tools/bin/ior -f ior.txt
Machine : Linux ec01
Start time skew across all tasks: 2495358.01 sec
TestID : 0
StartTime : Thu Jun 25 19:51:32 2020
Path : /ai400
```

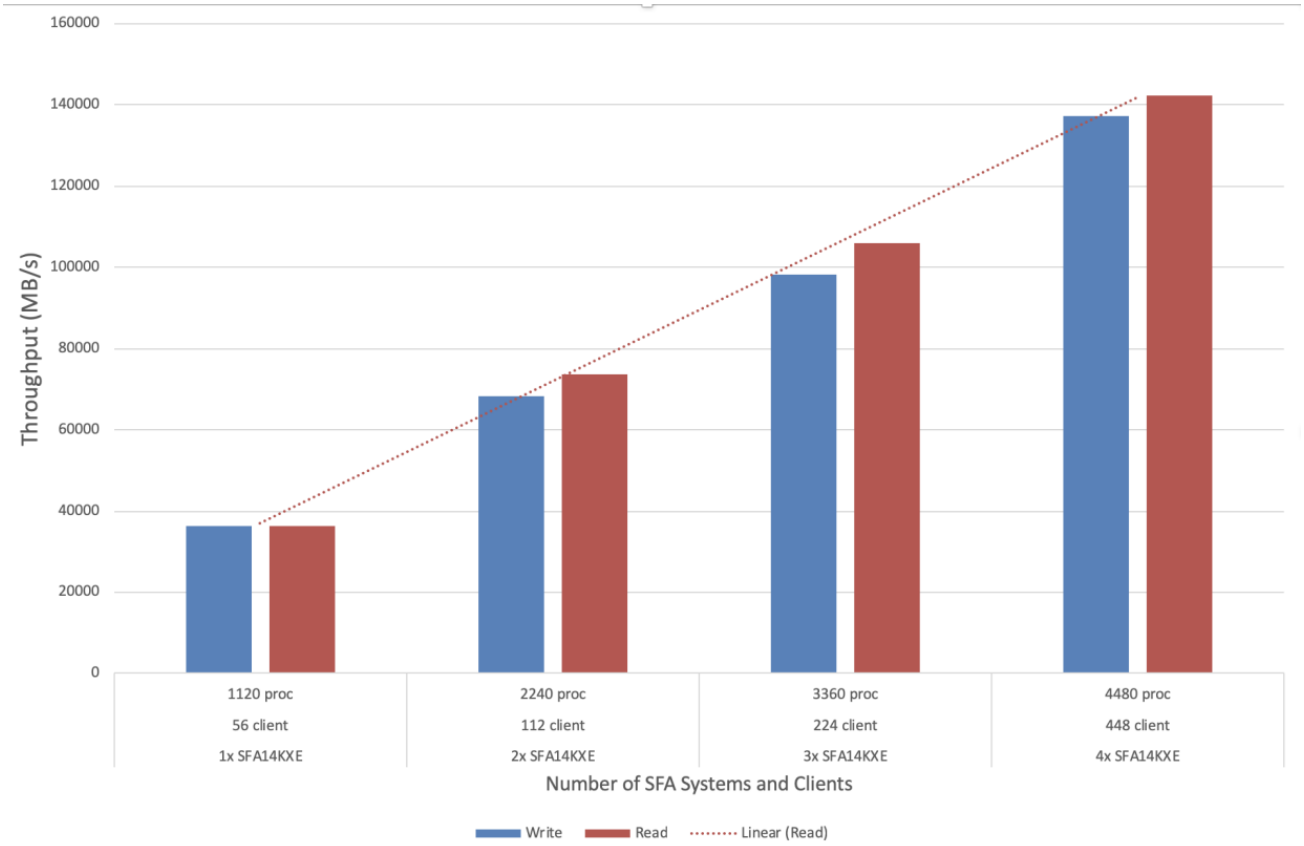
```
FS : 15.8 TiB Used FS: 0.0% Inodes: 128.0 Mi Used Inodes: 0.0%
Participating tasks: 40
Options:
api : POSIX
apiVersion :
test filename : /ai400/scratch
access : file-per-process
type : independent
segments : 1
ordering in a file : sequential
ordering inter file : no tasks offsets
tasks : 40
clients per node : 1
repetitions : 1
xfersize : 65536 bytes
blocksize : 192 GiB
aggregate filesize : 7.50 TiB
Results:
access bw(MiB/s) block(KiB) xfer(KiB) open(s) wr/rd(s) close(s) total(s)
iter
------ --------- ---------- --------- -------- -------- -------- --------
----
delaying 10 seconds . . .
Commencing write performance test: Thu Jun 25 19:51:42 2020
write 39168 201326592 64.00 0.009452 200.78 2.62 200.79 0
delaying 10 seconds . . .
Commencing read performance test: Thu Jun 25 19:55:13 2020
read 45850 201326592 64.00 0.001041 171.52 55.97 171.52 0
remove - - - - - - 9.62 0
Max Write: 39167.56 MiB/sec (41070.16 MB/sec)
Max Read: 45850.30 MiB/sec (48077.53 MB/sec)
Summary of all tests:
Operation Max(MiB) Min(MiB) Mean(MiB) StdDev Max(OPs) Min(OPs) Mean(OPs)
StdDev Mean(s) Stonewall(s) Stonewall(MiB) Test# #Tasks tPN reps fPP reord
reordoff reordrand seed segcnt blksiz xsize aggs(MiB) API RefNum
write 39167.56 39167.56 39167.56 0.00 626680.92 626680.92 626680.92 0.00
200.78658 NA 0 40 1 1 1 0 1 0 0 1 206158430208 65536 7864320.0 POSIX 0
read 45850.30 45850.30 45850.30 0.00 733604.86 733604.86 733604.86 0.00
171.52166 NA 0 40 1 1 1 0 1 0 0 1 206158430208 65536 7864320.0 POSIX 0
Finished : Thu Jun 25 19:58:14 2020
```

In a different set of measurements for a different project scaling behavior had been tested with the following result:

This measurement compares sequential performance stepping from 1 x ES400NV to 2 x ES400NV. Comparing the write results from 1 to 2 building blocks with the appropriate transfer sizes we scale with a factor of 1.98 to 2.1. The super-linear scaling for the transfer size of 4kiB could not be explained and is expected to be an outlier in the measurement.

Excluding the 4kiB value all scaling is within 2.01 and 1.98.

Adding another data point we can illustrate the ability of the filesystem to scale linear further:



Within a 2% margin the above graph shows that with a scaling number of building blocks and scaling number of clients a practically linear scaling can be expected.

Using these data points and the initial measurement we come to the following performance expectation for the proposed system:

- A single building block achieves 41 GB/s sequential writes and 48 GB/s sequential reads
- These values are achieved with 40 MPI ranks on 40 physical nodes
- To achieve the requested 400 GB/s for writes a ~97.5% efficiency is required – which is justified from the measurements above
- For sequential reads the achievable value will exceed the requested 400GB/s by more than 10%
- To achieve these performance levels, it is expected to need at least 400 MPI ranks/client nodes

The exact settings such as "number of nodes", "number of MPI ranks", "EXAScaler stripe- and chunksize" etc. will be set during the running of the acceptance benchmarks to show the best possible performance.

Sequential read/write with the requested IOR will be **400 GB/s** with a sufficiently high number of client nodes/MPI ranks.

10.3.4.3 Commitments

Name	Partition	Minimum value of max. read and max. write [MB/s]
IOR1	LCST	**207.6**
IOR2	HPST	**400**

IOR: commitments

10.4 Application Benchmarks

In this section, we explain the compilation, execution and runtime commitments for the applications benchmarks.

10.4.1 GROMACS

10.4.1.1 Benchmark description

GROMACS is a versatile package to perform molecular dynamics, i.e. simulate the Newtonian equations of motion for systems with hundreds to millions of particles. It is primarily designed for biochemical molecules like proteins, lipids and nucleic acids that have a lot of complicated bonded interactions.

10.4.1.2 Compilation

For CPU runs, GROMACS version 2020.1 was used for this test and for compilation GCC-5.5.0 and IntelMKL 2018 used. We use the following cmake command line to generate the makefile:

```
export CFLAGS=$FLAGS;
export CXXFLAGS=$FLAGS;
export MPICC=mpiicc
export MPICXX=mpiicpc
export MPIF90=mpiifort
export CC=gcc; export FC=gfortran; export CXX=g++; export F77=gfortran

cmake .. -DGMX_FFT_LIBRARY=mkl -DGMX_SIMD=AVX2_256 -DGMX_MPI=on -
DGMX_OPENMP=ON -DGMX_HWLOC=OFF -
DMKL_LIBRARIES='/opt/intel/compilers_and_libraries_2018.5.274/linux/mkl/li
b/intel64/libmkl_scalapack_lp64.a;-Wl,--start-
group;/opt/intel/compilers_and_libraries_2018.5.274/linux/mkl/lib/intel64/
```

```
libmkl_cdft_core.a;/opt/intel/compilers_and_libraries_2018.5.274/linux/mkl
/lib/intel64/libmkl_intel_lp64.a;/opt/intel/compilers_and_libraries_2018.5
.274/linux/mkl/lib/intel64/libmkl_sequential.a;/opt/intel/compilers_and_li
braries_2018.5.274/linux/mkl/lib/intel64/libmkl_core.a;/opt/intel/compiler
s_and_libraries_2018.5.274/linux/mkl/lib/intel64/libmkl_blacs_intelmpi_lp6
4.a;-Wl,--end-group;-lpthread;-lm;-ldl' -
DMKL_INCLUDE_DIR=/opt/intel/compilers_and_libraries_2018.5.274/linux/mkl/i
nclude

make
```

For GPU, GROMACS was built with

- GCC 5.5.0
- CUDA 10.2
- Intel MKL 2018

We use the following cmake command line to generate the makefile:

```
export CFLAGS=$FLAGS;
export CXXFLAGS=$FLAGS;
export MPICC=mpiicc
export MPICXX=mpiicpc
export MPIF90=mpiifort
export CC=gcc; export FC=gfortran; export CXX=g++; export F77=gfortran

cmake .. -DGMX_FFT_LIBRARY=mkl -DGMX_SIMD=AVX_512 -DGMX_GPU=ON -
DGMX_MPI=on -DGMX_OPENMP=ON -DGMX_HWLOC=OFF -
DMKL_LIBRARIES='/opt/intel/compilers_and_libraries_2018.5.274/linux/mkl/li
b/intel64/libmkl_scalapack_lp64.a;-Wl,--start-
group;/opt/intel/compilers_and_libraries_2018.5.274/linux/mkl/lib/intel64/
libmkl_cdft_core.a;/opt/intel/compilers_and_libraries_2018.5.274/linux/mkl
/lib/intel64/libmkl_intel_lp64.a;/opt/intel/compilers_and_libraries_2018.5
.274/linux/mkl/lib/intel64/libmkl_sequential.a;/opt/intel/compilers_and_li
braries_2018.5.274/linux/mkl/lib/intel64/libmkl_core.a;/opt/intel/compiler
s_and_libraries_2018.5.274/linux/mkl/lib/intel64/libmkl_blacs_intelmpi_lp6
4.a;-Wl,--end-group;-lpthread;-lm;-ldl' -
DMKL_INCLUDE_DIR=/opt/intel/compilers_and_libraries_2018.5.274/linux/mkl/i
nclude

make
```

10.4.1.3 Execution

As stated in Q&A 18.05.2020 11:06 "The GROMACS GluCL benchmark is no longer required" as well as in Q&A 08.06.2020 14:34 "(GRO1 and GRO 3 are no longer valid)", we did not run the benchmark GluCL.

According to RFP we used Water benchmark with 1536k atoms, changed constraints to h-bonds and run the benchmark with 5000 steps.

For CPU run, the code is launched through this command:

```
time srun  --mpi=pmi2 --
cpu_bind=map_cpu:0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,2
2,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,
47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,65,66,67,68,69,70,71
,72,73,74,75,76,77,78,79,80,81,82,83,84,85,86,87,88,89,90,91,92,93,94,95,9
6,97,98,99,100,101,102,103,104,105,106,107,108,109,110,111,112,113,114,115
,116,117,118,119,120,121,122,123,124,125,126,127 gmx_mpi mdrun -s
topol_pme.tpr -nsteps 5000 -ntomp $OMP_NUM_THREADS -noconfout
```

For GPU the code is launched through this command:

```
time mpirun -np $NP -ppn $PPN gmx_mpi mdrun -dlb auto -nb gpu -pin on -s
topol_pme.tpr -nsteps 5000 -ntomp $OMP_NUM_THREADS -tunepme -noconfout
```

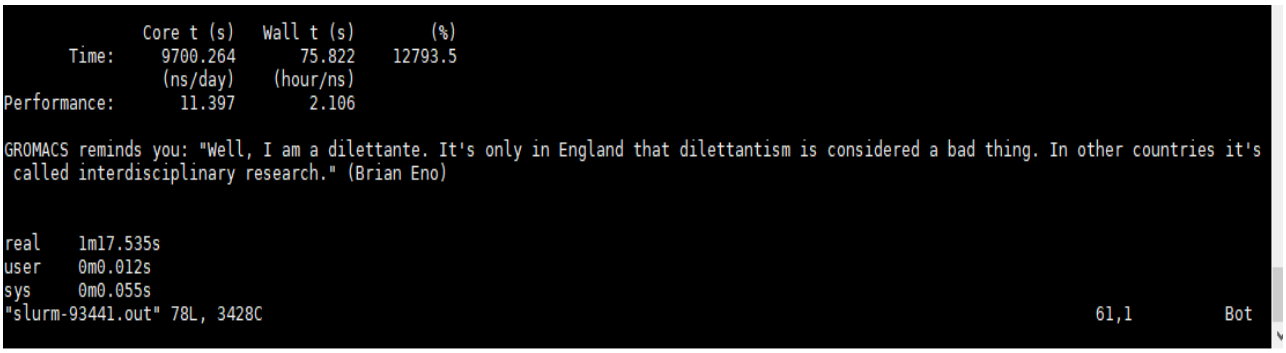
10.4.1.4 Measurements and results

The timings are retrieved at the end of the code standard output, for both the iteration and wall clock timings:

```
               Core t (s)   Wall t (s)        (%)
       Time:     9700.264       75.822    12793.5
                 (ns/day)    (hour/ns)
Performance:       11.397        2.106
```

Here is an example of output screenshot:

Single node Gromacs:- 1m17.5 sec



10.4.1.4..1 CPU

Scalability

The following table is showing our results on our supercomputer with SKU AMD 7742:

#nodes	#MPI	Turbo	Wall time	ns/day
1	128	ON	1m17.535s	11.397
2	256	ON	0m36.785s	24.966
4	512	ON	0m22.792s	43.181
16	2048	ON	0m14.193s	115.601
32	4096	ON	0m12.748s	121.986

GROMACS: Scalability study on AMD ROME 7742.

Frequency dependency

Frequency (GHz)	#nodes	#MPI	Wall time	ns/day
2.25	1	128	1m28.331s	10.014
2.00	1	128	1m35.906s	9.22
1.50	1	128	2m1.321s	7.278

GROMACS: Frequency dependency on AMD ROME 7742.

The previous table is showing the timings for different frequencies. These results are giving a frequency dependency of 27%.

Memory bandwidth

Condition	#nodes	#MPI	PPN	Wall t	ns/day
Compact	16	1024	64	0m15.863s	72.730
Scatter	16	1024	64	0m15.645s	73.315

GROMACS: Memory bandwidth dependency on AMD ROME 7742.

To study the memory bandwidth usage, we used 16 number of nodes and PPN 64. For the compact mode, we put all the processes by node on one socket. For the scatter mode, we spread the processes equally among the two sockets. The difference between those two runs is giving a memory dependency of 2%.

MPI profiling

Following is the single node MPI profile obtained with IPM tool. Communications use about 16.98% of the total time.

```
#                            [time]        [calls]       <%mpi>
<%wall>
# MPI_Alltoall               637.927       640128        33.58          5.70
# MPI_Recv                   525.896       804496        27.68          4.70
# MPI_Sereturnecv               395.323       6.37078e+06   20.81
3.53
# MPI_Waitall                253.914       820544        13.36          2.27
# MPI_Init_thread            28.898        128           1.52           0.26
# MPI_Bcast                  23.518        16672         1.24           0.21
# MPI_Allreduce              17.3004       50080         0.91           0.15
# MPI_Irecv                  4.61665       500416        0.24           0.04
# MPI_Comm_split             4.28168       4512          0.23           0.04
# MPI_Isend                  2.96092       1.30467e+06   0.16           0.03
# MPI_Scatterv               2.76149       96            0.15           0.02
# MPI_Start                  2.03547       288           0.11           0.02
# MPI_Barrier                0.431225      256           0.02           0.00
# MPI_Gather                 0.0107331     672           0.00           0.00
# MPI_Comm_free              0.00121784    384           0.00           0.00
# MPI_Type_commit            0.000485659   32            0.00           0.00
# MPI_Send                   0.000247478   240           0.00           0.00
# MPI_Type_contiguous        0.000152588   32            0.00           0.00
# MPI_Comm_rank              0.000132561   1697          0.00           0.00
# MPI_Comm_size              0.000116587   1440          0.00           0.00
# MPI_Finalize               7.22408e-05   128           0.00           0.00
# MPI_Get_processor_name     6.38962e-05   256           0.00           0.00
# MPI_TOTAL                  1899.88       1.0518e+07    100.00
16.98
```

10.4.1.4..2 GPU

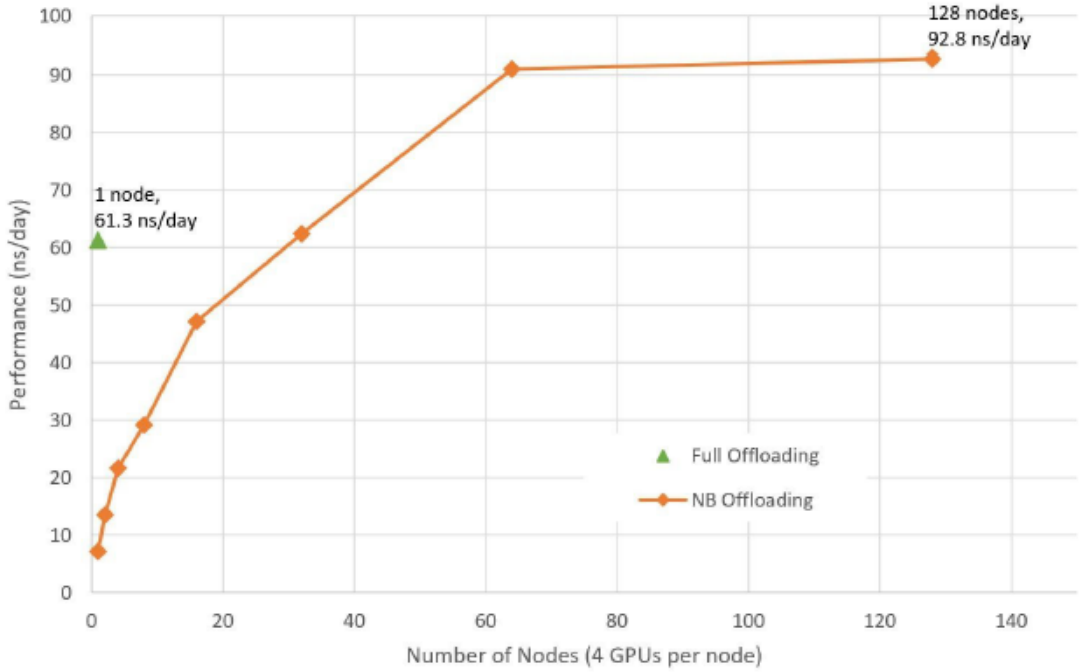
V100 scalability

First, we made tests on our V100 Spartan cluster. For these tests the number of MPI tasks used per node was 20 and the number of threads was set 2 for all experiments. The following table shows the scaling behavior of this setup with up to 4 nodes and 16 nvidia V100 GPU cards.

#nodes	#MPI	OMP	Turbo	Walltime	ns/day
1	20	2	ON	1m17.518s	11.808
2	40	2	ON	0m43.104s	22.561
3	60	2	ON	0m33.367s	30.354
4	80	2	ON	0m28.369s	36.554

GROMACS: V100 scalability results.

A100 scalability



GROMACS: A100 scalability - Performance of the test case on varying node counts, each with 4 GPUs per node.

In the figure above, we show results using two different software configurations. The orange diamonds denote the results where GPU offloading is restricted to Non-bonded (NB) forces. Scaling is observed up to 64 nodes, at which point the performance plateaus at just over 90 ns/day. However, with only partial offloading, these runs are limited by the CPU and network, and hence are an extremely inefficient use of resources.

Recent developments in GROMACS allow much more efficient utilization of GPU resources through full offloading as described in the paper Heterogeneous Parallelization and Acceleration of Molecular Dynamics Simulations in GROMACS and the blog article Creating Faster Molecular Dynamics Simulations with GROMACS 2020. The green diamond shows a result using full offloading on a single node: the performance is just over 60 ns/day, which is comparable with the 32-node partially-offloaded result, in other words around a factor of 32 more efficient.

It is not currently possible to run the fully-offloaded configuration above a single node using the release version of GROMACS. This is due to a restriction in the software, for which full offload is currently only possible with the use of the internal thread-MPI library which is restricted to a single node. We have a patched version which relaxes this restriction to allow use with OpenMPI, but even with this the best result is still on a single node since the limiting factor is the PME force calculation which cannot be decomposed across multiple GPUs.

The most efficient configuration with current software is therefore to run the "Full offloading" version on a single node.

Software configuration

To obtain the previous results, we use the standard GROMACS 2020 release version with the runtime options given in the IZUM document, except we add -nstlist 200 which reduces the frequency of the neighbor search step which improves performance without any loss of accuracy.

For the "NB Offloading" results we compile using CUDA 11.0.167, G++ 8.4.0 and OpenMPI 4.0.4rc.

For the "Full Offloading" results compile as above except using GROMACS internal thread-MPI,
and we achieve such full offloading by adding

```
-pme gpu -npme 1
```

and setting the following environment variables:

```
export GMX_GPU_DD_COMMS=1
export GMX_GPU_PME_PP_COMMS=1
export GMX_FORCE_UPDATE_DEFAULT_GPU=1
```

10.4.1.5 Projection and commitments

We are projecting the performance measured in ROME 7742 to 7H12 for CPU commitments:

#SKU	#nodes	PPN	#MPI	Turbo	ns/day
7H12	1	128	128	ON	**11.43**

GROMACS: CPU commitments

Regarding the run on GPU, by far the most efficient configuration with current software is to run with full GPU offloading on a single compute node. The requirement to run the benchmark "on at least 4 GPU nodes" forces 3 of the nodes to be idle.

We must adjust the measured result given in the previous section to account for the fact that intra-node GPU-GPU communications on DGX-A100 have 3X the bandwidth of those on Redstone (since the NVSwitch configuration allows more NVLinks). We use the profiler to discover that 175us per step are being spent in such communications, and therefore a further 350us per step, or 1.75 seconds across the whole 5,000 step simulation, must be added to the baseline time (which is 14.108s). This adjusts the projected performance to 54.5 ns/day.

We therefore provide two projections using this projected performance as follows:

Configuration	#nodes	ns/day
Fully offloaded version, using 1 simulation per node	1	**54.5**
Fully offloaded version, using 1 simulation per 4 nodes, running on a single node with other 3 nodes idle	4	**54.5**

GROMACS: GPU commitments

From which we compute the values committed in the excel sheet:

Req. No.	Value
GRO2	**11658.60**
GRO4	**817**

GROMACS: Value for evaluation

following the formulae $Value_{GRO2}$ = R_A*N_A+R_B*N_B = 11.43*960+11.43*60 and $Value_{GRO4}$ = R_A*N_A/4 = 54.5*60/4.

As explained in the RFP and following the updates in the Q&A 12.05.2020 11:11 "The sum of the extrapolated value for compute and GPU partitions should be provided." and in Q&A 16.06.2020 10:24 "The Tenderer should obtain the performance results when running benchmark on the test system and then compute the sum of the performance of enough smaller runs to use all the nodes in Compute and GPU partition. The sum in ns/day should be provided for CPUs on both Compute and GPU partitions.", the formula for the value of GRO2 has been updated as above.

As explained in the RFP and following the updates in the Q&A 16.06.2020 10:24 "The Tenderer should obtain the performance results when running benchmark on at least 4 GPU nodes on the test system and then compute the sum of the performance of enough smaller runs to use all the nodes in GPU partition. The sum in ns/day should be provided for GPU partition.", the formula for the value of GRO4 is $Value_{GRO4}$ = R_A*N_A/4 with N_A the number of GPU nodes and R_A the performance on 4 GPU nodes.

10.4.2 Quantum ESPRESSO

10.4.2.1 Benchmark description

Quantum Espresso stands for Open Source Package for Research in Electronic Structure, Simulation, and Optimization. It is an integrated suite of computer codes for electronic-

structure calculations and materials modelling, based on density-functional theory, plane waves, and pseudopotentials (norm-conserving, ultrasoft, and projector-augmented wave). It is freely available to researchers under the terms of the GNU General Public License. Quantum ESPRESSO is written mostly in Fortran90 and parallelized using MPI and OpenMP.

10.4.2.2 Compilation

Two versions have been used, one for CPU measurements and another one for GPU measurements.

For CPU measurements:
QE 6.5
Compiler Intel 2018.5.274
Intel MPI

We used the arch flag mpi_intel_linux_64, but we modified the make.inc as following

```
DFLAGS         =  -D__DFTI -D__MPI -D__SCALAPACK
CFLAGS         = -O3 -march=core-avx2 -fno-alias -ansi-alias -unroll -m64
$(DFLAGS) $(IFLAGS) -qopenmp
F90FLAGS       = -O3 -march=core-avx2 -fno-alias -ansi-alias -unroll -m64
-nomodule -qopenmp -fpp $(FDFLAGS) $(CUDA_F90FLAGS) $(IFLAGS) $(MODFLAGS)
FFLAGS         = -O3 -march=core-avx2 -unroll -m64 -assume byterecl -g -
traceback -qopenmp
LDFLAGS        = -qopenmp
BLAS_LIBS      =  -lmkl_gf_lp64 -lmkl_intel_lp64 -lmkl_intel_thread -
lmkl_core -lirc -lsvml
LAPACK_LIBS    = -lmkl_blacs_openmpi_lp64
SCALAPACK_LIBS = -lmkl_scalapack_lp64 -lmkl_blacs_openmpi_lp64
```

Then we compiled with:

```
make pwall &>${DIR_COMPIL}/make.log
make install &>${DIR_COMPIL}/make_install.log
```

For GPU measurements on V100:

- QE-gpu 6.5a1 release
- Compiler PGI 19.10
- CUDA 10.1

```
DFLAGS         =  -D__CUDA -D__PGI -D__MPI -D__GPU_MPI -D__USE_CUSOLVER -
D__DFTI
GPU_ARCH=70
CUDA_RUNTIME=10.1
CUDA_F90FLAGS=-Mcuda=cc70,cuda10.1 -I/software/compilers/pgi/linux86-64-
llvm/19.10/include $(MOD_FLAG)$(TOPDIR)/EIGENSOLVER_GPU/lib_eigsolve
```

```
CFLAGS         = -fast -Mpreprocess $(DFLAGS) $(IFLAGS)
F90FLAGS       = -fast -Mcache_align -Mpreprocess -Mlarge_arrays -mp
$(FDFLAGS) $(CUDA_F90FLAGS) $(IFLAGS) $(MODFLAGS)
FFLAGS         = -fast -mp
LDFLAGS        = -pgf90libs  -Mcuda=cc70,cuda10.1
BLAS_LIBS      = -L${MKLROOT}/lib/intel64 -lmkl_intel_lp64 -lmkl_core -
lmkl_intel_thread -lpthread -lm -ldl
FFT_LIBS       = -L${MKLROOT}/lib/intel64/ -lmkl_cdft_core -lmkl_core
FOX_FLAGS = -fast -Mcache_align -Mpreprocess -Mlarge_arrays
CUDA_LIBS=-Mcudalib=cufft,cublas,cusolver -L/usr/local/cuda-10.1/lib64 -
lcusolver $(TOPDIR)/EIGENSOLVER_GPU/lib_eigsolve/lib_eigsolve.a
```

We linked with MKL and added the -D__DFTI flag to make.inc (use MKL for CPU FFTs). We used the -D__USE_CUSOLVER flag which toggles the use of cuSOLVER instead of the custom solver integrated into QE.
We did not use the -D__SCALAPACK flag. This flag seems to improve the application scalability but made the code more than 2x slower up to 24 nodes.

For GPU measurements on A100:

- QE-gpu 6.5a1 release
- Compiler PGI dev/nightly (April 2020)
- CUDA 11.0.97 (dev)
- OpenMPI 4.1.0a1 with UCX 1.7.0-rc2
- MKL 2020.0-088

The code is installed with:

```
CC=pgcc FC=pgfortran ./configure --with-cuda=$CUDA_HOME --with-cuda-
runtime=11 --with-cuda-cc=80 --enable-openmp --with-scalapack=no
```

We linked with MKL and added the -D__DFTI flag to make.inc (use MKL for CPU FFTs). We also use the flag -D__GPU_MPI enabling Cuda-aware MPI support (labelled as experimental).

To compile with PGI 20, we made the following changes to the configuration script that does neither affect results nor performance:

```
sed "s/mkl_\\\${ompimp}/mkl_intel/g" -i install/configure
```

10.4.2.3 Execution

We run the Graphene@Ir(111)-(9x9)-3L benchmark (GRIR443). This benchmark makes a single SCF iteration (electron_maxstep = 1) with "pw.x" program. The structure is a graphene-layer on top of three-layer Ir(111)-(9x9) slab and consists of 443 atoms; the

thickness of vacuum layer is about 20 Angstroms. Number of k-points is 4, i.e., a(2x2x1) k-mesh is used.

The CPU version is executed that way:

```
APP_CMD="pw.x -nk 4 -i grir443.in -ntg 2 -ndiag 200"
RUN_CMD="srun --distribution=plane=${ppn} -n ${tasks} --
cpu_bind=verbose,mask_cpu:${RUN_MASK} ${APP_CMD} "
```

The time is extracted with the following command:

```
grep "PWSCF" slurm-${JOB_ID}.out
```

10.4.2.4 Measurements and results

10.4.2.4..1 CPU

Frequency	#nodes	#MPI	#PPN	#OMP	Time (s)	Dependency
TURBO	4	512	128	1	301.70	70%
2250	4	512	128	1	313.95	
2000	4	512	128	1	368.05	
1500	4	512	128	1	454.25	
TURBO	4	256	64	2	289.50	80%
2250	4	256	64	2	297.78	
2000	4	256	64	2	346.55	
1500	4	256	64	2	440.50	
TURBO	4	256	64	1	563.06	70%
2250	4	256	64	1	655.51	
2000	4	256	64	1	712.84	
1500	4	256	64	1	875.38	

QE: CPU frequency study on ROME 7742

The previous table shows a frequency dependency of 70% and 80%, depending on the number of MPI processes and on the number of nodes.

10.4.2.4..2 GPU

V100 results

CPU Frequency (MHz)	#nodes	#MPI	#PPN	#OMP	Time (s)	Dependency
TURBO	4	16	4	10	395.70	40%
2400	4	16	4	10	433.99	
2200	4	16	4	10	449.06	
2000	4	16	4	10	462.52	

QE: CPU frequency study on SKL 6148 with 4 V100

GPU Frequency (MHz)	#nodes	#MPI	#PPN	#OMP	Time (s)	Dependency
1530	4	16	4	10	425.16	20%
1335	4	16	4	10	435.12	
1132	4	16	4	10	446.61	
930	4	16	4	10	465.32	

QE: GPU frequency study on SKL 6148 with 4 V100

The previous table shows a CPU frequency dependency of 40% and a GPU frequency dependency of 20% on V100.

A100 – NVlink sensitivity

In this test, we change the number of used NICs through UCX_NET_DEVICES and split the GPUs across more nodes to prevent direct NVlink connectivity, in order to evaluate the impact of NVlink bandwidth.

Nodes	GPU used / node	NCI used / node	Cores used / node	Nb of pools (-npool)	Time (s)	Time FFTs (s)
16	1	1xHDR200	8	4	101.7	10.3
4	4	4xHDR200	32	4	109.9	7.4

QE: Tests with and without NVlink connectivity

The configuration with 4 GPUs per node has pools that are directly connected with NVSwitch, while the configuration with 1 GPU per node must use GPUDirect RDMA (over IB) which has much lower bandwidth but better latency. Looking at the individual QE timers the configuration with 4 GPUs per node is 3s faster in FFTs (which are dependent on NVlink bandwidth) but slower for every other timer (mainly the Eigensolver, which is likely latency-bound), resulting in an 8s difference overall. As the per-GPU NVlink bandwidth on the proposed NVIDIA HGX A100 4-GPU based system will be 3x lower than in this configuration (1 GPU per node), and the difference in FFTs is only 3s, we

conclude that the NVSwitch vs NVlink bandwidth difference will only have a negligible performance effect on this benchmark.

A100 – Network sensitivity

In this test, we change the number of used NICs through UCX_NET_DEVICES in order to model the impact of network bandwidth.

Nodes	GPU used / node	Nb of pools (-npool)	NICs used	Time (s)
4	4	4	4 per node	109.9
4	4	4	1 per node	112.8

QE: Comparing performance when using less NICs

We observe only a 2% performance impact when going from 1xHDR200 per GPU to 1xHDR200

per 4 GPUs, even though this increases latency and reduces available bandwidth by 4x.

A100 – CPU sensitivity

In this test, we change the number of cores and MKL_NUM_THREADS to estimate the effect of CPU.

Nodes	GPU used / node	Cores used / MPI rank	MKL_NUM_THREADS	Nb of pools (-npool)	Cores used / node	Time (s)
4	4	6	5	4	24	120.3
4	4	8	6	4	32	111.5
4	4	12	11	4	48	111.0
4	4	16	15	4	64	113.7

QE: CPU sensitivity – NPOOL = 4. Comparing performance with different number of cores and MKL_NUM_THREADS. Using 4 MPI ranks per node.

Nodes	GPU used / node	Cores used / MPI rank	MKL_NUM_THREADS	Nb of pools (-npool)	Cores used / node	Time (s)
4	4	6	5	1	24	199.5
4	4	8	6	1	32	191.8
4	4	12	11	1	48	190.5
4	4	16	15	1	64	195.4

QE: CPU sensitivity – NPOOL = 1. Comparing performance with different number of cores and MKL_NUM_THREADS. Using 4 MPI ranks per node.

This test shows that there is a low impact of the number of cores and that it can even actually be detrimental to performance.

A100 – Number of pools

Nodes	GPU used / node	Cores used / node	NICs used / node	Nb of pools (-npool)	Time (s)
4	4	32	2xHDR200	1	191.8
4	4	32	2xHDR200	4	111.5

QE: Number of pools. Measurements on the DGX A100 cluster. Using 4 MPI ranks per node and MKL_NUM_THREADS=5.

We observe that the -npool parameter has a large performance impact. This is expected, because having fewer pools generates more parallelism. The main drawback of using more pools is the increased memory consumption, but A100 still has enough memory to run the test case with 4 pools.

10.4.2.5 Commitments

According to the previous parts, our commitments on Quantum Espresso are:

#nodes	CPU	Partition	#GPU/nodes	Npools	Time (s)
4	7H12	Compute	0	default	**265**
4	7H12	GPU	4	4	**111.5**

QE: Commitments

From which we compute the values committed in the excel sheet:

Req. No.	Value
QES1	**18.77**
QES2	**7.43**

QE: Value for evaluation

following the formulae $Value_{QES1}$ = (R_A/N_A+R_B/N_B)*4 = (265/960+265/60)*4 and $Value_{QES2}$ = (R_A/N_A)*16 = (111.5/(60*4))*16.

As explained in the RFP and following the updates in the Q&A 12.05.2020 11:12 "... where N_A and N_B are number of nodes in Compute and GPU partitions respectively, and R_A and R_B are benchmarks results of node types A and B." and in Q&A 28.05.2020 14:57 "the total score is: R_A/N_A*16 where N_A is the number of graphic subsystems in the GPU partition, and R_A is the benchmark results of the benchmark running on 16 GPUs.", the formulae have been updated as above.

10.4.3 HEPSpec06

10.4.3.1 Benchmark description

HEPSpec06 is a high energy physics benchmark based on SPEC CPU@2006.

10.4.3.2 Compilation

We use the file linux32-gcc_cern.cfg that we have retrieved in spec2k6-2.24.tar.gz from the link: https://w3.hepix.org/benchmarking/how_to_run_hs06.html#run-rules.
The application is compiled in 32-bit mode even if we are on a 64-bit OS.
We follow the rules explained on the website.

10.4.3.3 Execution

Whereas the 64-bit version gives better performance, we use the 32-bit version as it is required to publish validated results.

The run command is:

```
time ./runspec.sh -a 32 -b all_cpp -d "Optional description of the run"
```

10.4.3.4 Measurements

#nodes	CPU	Result
1	7742	3081.45

HEPSpec06: Measureements

10.4.3.5 Commitments

According the HEPSpec06 features, our commitments on HEPSpec06 are:

#nodes	CPU	Partition	Time (s)
1	7H12	Compute/GPU	**3100**

HEPSpec06: Commitments

From which we compute the values committed in the excel sheet:

Req. No.	Total score
HEP1	**3 162 000**

HEPSpec06: Value for evaluation

following the formula $Value_{HEP1}$ = R_A*N_A+R_B*N_B = (3100*960 + 3100*60).

10.4.4 OpenFoam

10.4.4.1 Benchmark description

OpenFOAM has an extensive range of features to solve anything from complex fluid flows involving chemical reactions, turbulence and heat transfer, to acoustics, solid mechanics and electromagnetics. For the required benchmark "motorbike", we used the solver simpleFoam.

10.4.4.2 Compilation

OpenFOAM v1906 was compiled the following libraries:

- GCC 9.1
- OpenMPI 4.0.3
- HPCX 2.6.0 with UCX 1.8

OpenFOAM was compiled with these main flags:

```
-O3  -ltbb -funroll-all-loops  -march=core-avx2
```

After modifying c++Opt, cOpt and bashrc files, the binaries are obtained simply by executing

```
./Allwmake -j 128
```

10.4.4.3 Execution

OpenFOAM is submitted through slurm with the following simplified script (IntelMPI version):

```
export WM_ARCH_OPTION=64
tasks=$(( ppn*SLURM_NNODES ))
BASHRC=$INSTALL_DIR/BUILD_OpenFOAM_v1906_OPENMPI_4.0.3_HPCX_GCC_9.1_O3_AVX2_SETTINGS_SC
OTCH_O3_preciseetc/bashrc
source $ BASHRC
if [ $SLURM_NTASKS == 2048 ]
then
decomp="32 16 4" # 2048
elif [ $SLURM_NTASKS == 1024 ]
then
decomp="32 16 2" # 1024
elif [ $SLURM_NTASKS == 512 ]
then
decomp="32 8 2" # 512
elif [ $SLURM_NTASKS == 256 ]
then
decomp="16 8 2" # 256
decomp="16 2 8" # 256
elif [ $SLURM_NTASKS == 128 ]
then
#decomp="16 4 2" # 128
#decomp="8 8  2" # 128
decomp="16 2  4" # 128 #
#decomp="16 1  8" # 128
elif [ $SLURM_NTASKS == 64 ]
then
decomp="8 4 2" # 128
decomp="16 2 2" # 128
decomp="8 2 4" # 128
fi
sed -i "s;_N_;${SLURM_NTASKS};g" system/decomposeParDict_
sed -i "s;x_y_z;${decomp};g" system/decomposeParDict_
rm system/decomposeParDict
```

```
mv system/decomposeParDict_ system/decomposeParDict
# copy motorbike surface from resources directory
cp $FOAM_TUTORIALS/resources/geometry/motorBike.obj.gz constant/triSurface/
runApplication surfaceFeatureExtract
runApplication blockMesh
runApplication $decompDict decomposePar
# Using distributedTriSurfaceMesh?
if foamDictionary -entry geometry -value system/snappyHexMeshDict | \
  grep -q distributedTriSurfaceMesh
then
   echo "surfaceRedistributePar does not need to be run anymore"
   echo " - distributedTriSurfaceMesh will do on-the-fly redistribution"
fi
#runParallel $decompDict snappyHexMesh -overwrite
mpirun -np $SLURM_NTASKS  --report-bindings --rankfile ./rank_file --hostfile ./machinefile snappyHexMesh -
overwrite -parallel
#- For non-parallel running: - set the initial fields
# restore0Dir
#- For parallel running: set the initial fields
restore0Dir -processor
runParallel $decompDict patchSummary
runParallel $decompDict potentialFoam
runParallel $decompDict checkMesh -writeFields '(nonOrthoAngle)' -constant
fi

module load bpm/icc-2018.1.163/openmpi-hpcx-4.0.2rc3
#runParallel $decompDict $(getApplication)
mpirun -np $SLURM_NTASKS --report-bindings --rankfile ./rank_file --hostfile ./machinefile simpleFoam  -
parallel

runApplication reconstructParMesh -constant
runApplication reconstructPar -latestTime

mv log* $DIR_NFS
mv $SLURM_SUBMIT_DIR/slurm-$SLURM_JOBID\.out $DIR_NFS
```

Note that for each test case, the "system/decomposeParDict" file is adapted to consider the number of used MPI tasks.

Also, the mesh is adapted to the required size in the RFP by modifying the "system/blockMeshDict" as follow:

```
blocks
(
    hex (0 1 2 3 4 5 6 7) (100 40 40) simpleGrading (1 1 1)
);
```

10.4.4.4 Measurements and results

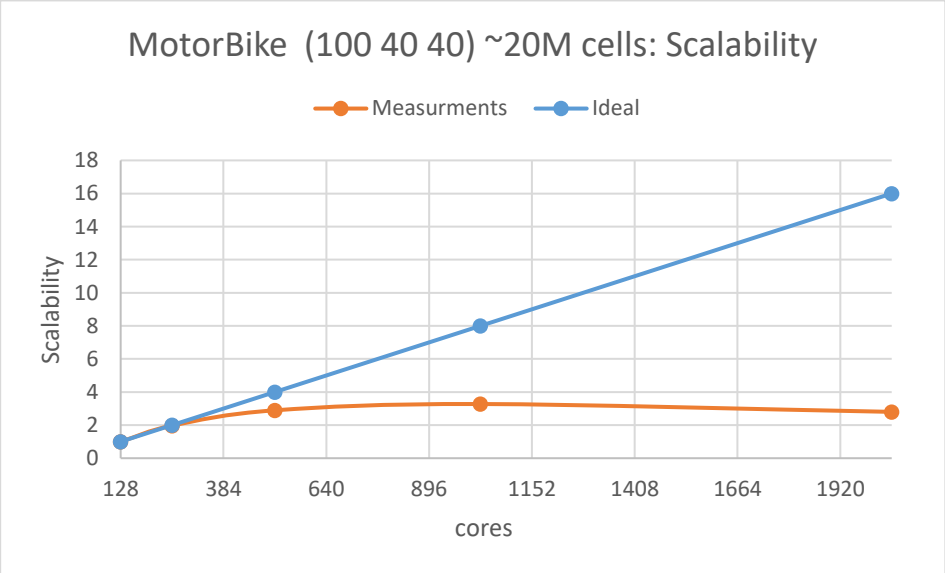
10.4.4.4..1 Execution time

The times execution of OpenFOAM on Spartan with different number of nodes and cores are summarized in the table and figure below.

Node	Cores	Execution time [s]	lock time [s]	Boost	GRID decomposition
1	128	239.16	242	ON	16 2 4
2	256	121.41	123	ON	16 8 2
4	512	82.62	85	ON	32 8 2
8	1024	73.03	76	ON	32 16 2
16	2048	85.4	91	ON	32 16 4

OpenFOAM: Measurements

One can see in the figure below the very bad scalability behavior of the test case as the number of cores increases. Indeed, using 2048 cores leads only to distribute less than 10.000 cells per core, which is very weak in terms of dedicated work per core.



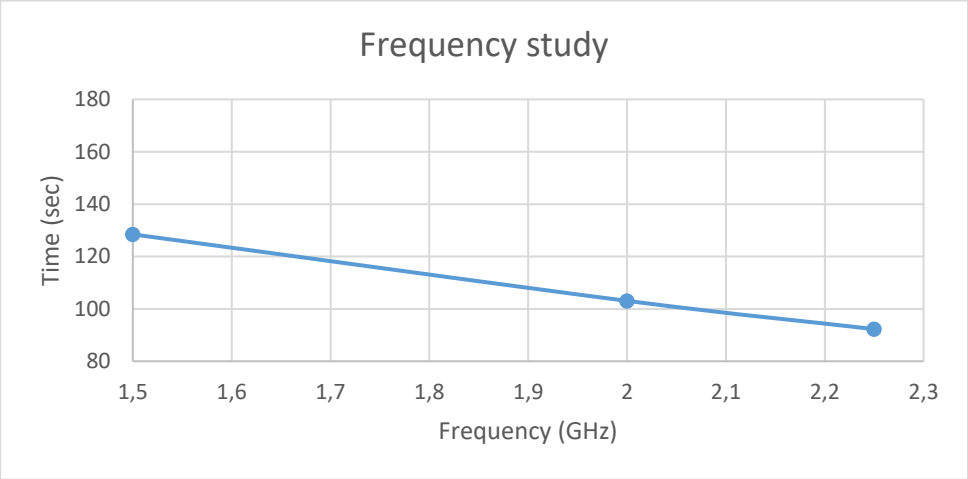
OpenFOAM: Scalability

As it was permitted by the customer to decrease the number of cores for commitment when the performance is better, we decided to take our commitments using only 128 cores instead of 2000 cores, then minimizing the product (Time x number of cores). The score formula will be modified accordingly.

10.4.4.4..2 Application profile

Application profile analysis is carried out using 1024 cores in order to show the different component of performance sensitivity.

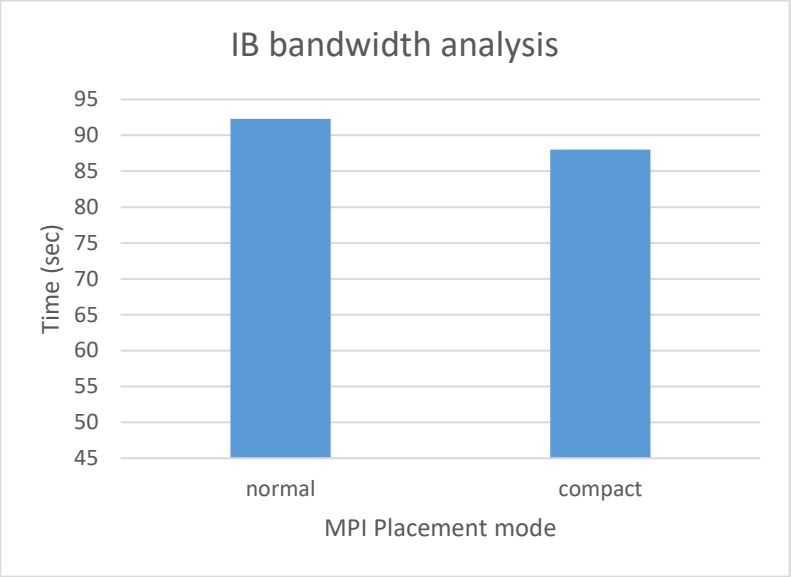
First, the CPU frequency were varied to highlight the dependency of the benchmark to it. The results are summarized on the following figure:



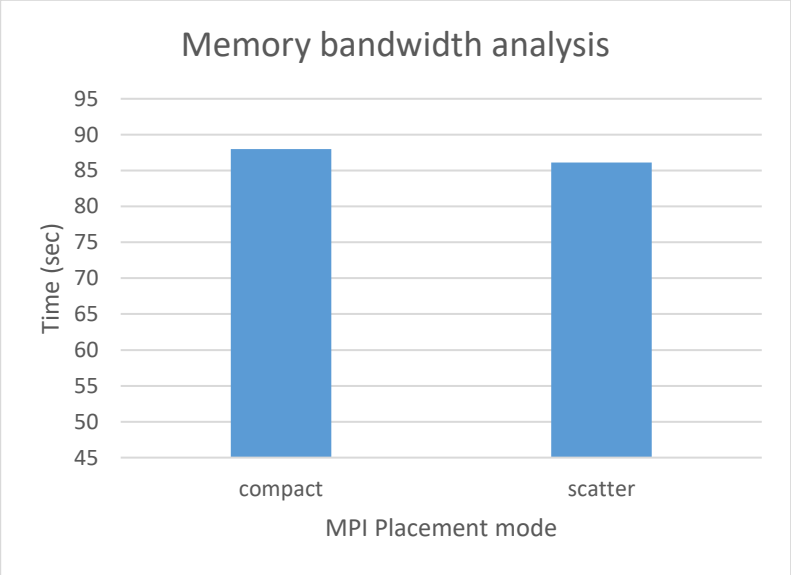
OpenFOAM: Frequency analysis

The results show that decreasing the CPU clock frequency by 50%, from 2.0 to 1.5 GHz, leads to 40 % of performance decrease. This means that this benchmark is highly sensitive to CPU frequency.

The effects of memory and interconnect bandwidth, through compact and scatter placements, were carried-out and the results are summarized in the following figures:



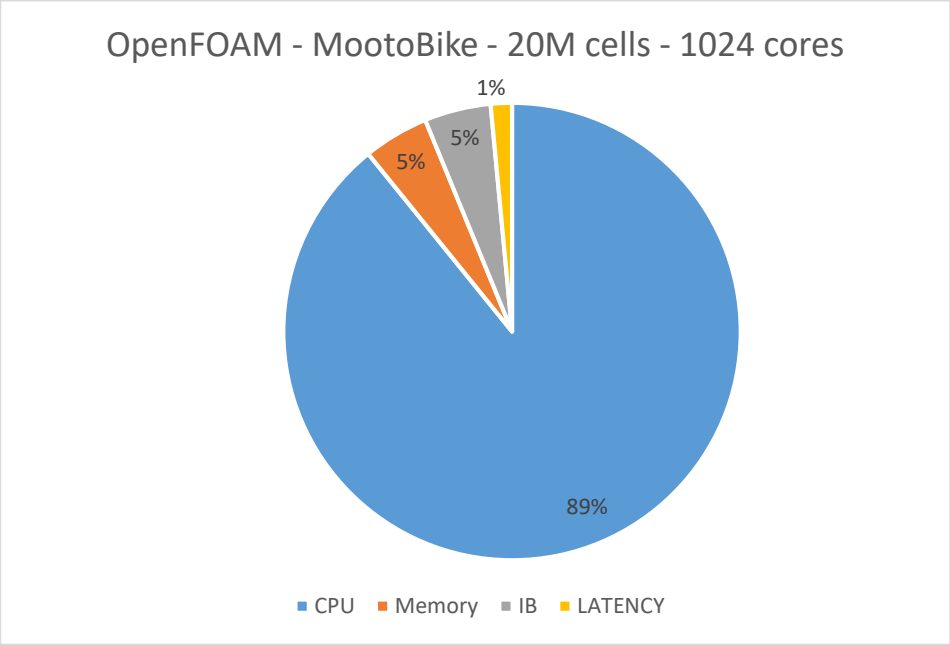
OpenFOAM: Interconnect bandwidth analysis



OpenFOAM: Memory bandwidth analysis

The results show clearly a weak dependency of the performance with respect to the memory and IB bandwidths.

These results lead to determine the benchmark profile regarding its performance sensitivity to CPU frequency, memory and IB bandwidths and latency, as follow:

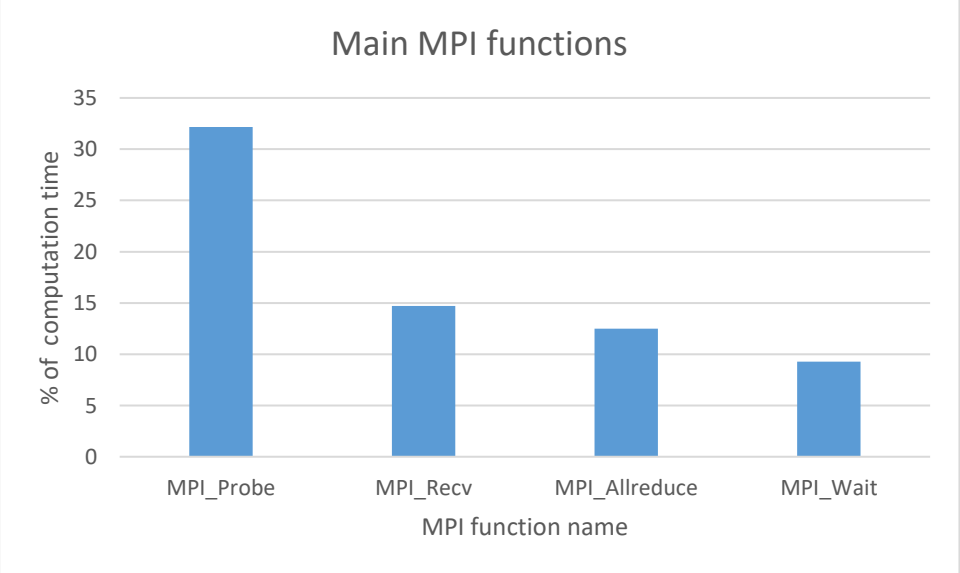


OpenFOAM: Profile

Indeed, 89% of this benchmark performance depends on the CPU clock frequency.

10.4.4.4..3 MPI profile

MPI communication represents between 26% and 79% of total parallel time, the main MPI called functions are shown in the figure bellow:



OpenFOAM: Main MPI functions calls

Note that even the MPI cost is not neglectable, however the IB speed has no impact on the present benchmark as seen in compact/scatter study. In general, communication

caused by imbalance are not sensitive to IB speed but to CPU clock frequency, which is the case for this benchmark.

10.4.4.5 Commitments

Based on the analysis and the results above, our commitments for OpenFOAM are summarized in the following table:

#nodes	CPU	GPU/node	Partition	Cores	Time (s)
1	7H12	0	Compute/GPU	128	**223**

OpenFOAM: Commitments

From which we compute the values committed in the excel sheet:

Req. No.	Total score
FOA1	**3.948**

OpenFOAM: Value for evaluation

following the formula $Value_{FOA1}$ = R_A/NC_A*128+R_B/NC_B*128 = 223/(960*128)*128+223/(60*128)*128.

As explained in the RFP and following the updates in the Q&A 12.05.2020 11:11 "Yes, the number of cores should be documented. In this case, the formula in Req. No. FOA1 shall be adjusted accordingly to account for the total computing power.", the formula for the value of FOA1 has been updated as above.

10.4.5 Tensorflow Resnet-50

10.4.5.1 Benchmark description

Resnet-50 is a convolutional neural network implemented in TensorFlow, that we used for image classification on a database from ImageNet.

10.4.5.2 Compilation

For this benchmark we have used a docker container that we build from the Nvidia Google Cloud Tensorflow version 20.02-tf1-py3. This docker container is using Cuda 10.2 and OpenMPI 3.1.4.

10.4.5.3 Execution

This docker container is launched with the software singularity (version 3.5.3-1.el7). Following commands show how we execute this benchmark in single and double precision:

```
export PYTVERSION=20.02-tf1-py3
export CONTAINER=docker://spartan10:5443/nvidia/tensorflow:${PYTVERSION}
export SINGULARITY_TMPDIR=/scratch_local/$USER
mkdir -p ${SINGULARITY_TMPDIR}

export DATA_DIR=
/scratch_na/backup_aloy/scratch_gpfs_aloy/AI/MLPERF/datasets/ImageNet_2012
/tf_records/train

export CKPT_DIR=
/scratch_na/users/bpuddua/2020_Vega/TF/FP$PREC/${nodes}N_${num_gpu}G_$SLUR
M_JOB_ID

if [ $PREC -eq 16 ]; then
CMD="tf_cnn_benchmarks.py \
--batch_size=256 \
--model=resnet50 \
 --optimizer=momentum \
 --variable_update=replicated \
 --nodistortions \
 --gradient_repacking=1 \
 --num_gpus=$num_gpu \
 --num_epochs=6 \
 --use_fp16\
 --data_dir=/datadir \
 --weight_decay=1e-4 \
 --train_dir=${CKPT_DIR}\
 --data_name=imagenet"
elif [ $PREC -eq 32 ]; then
CMD="tf_cnn_benchmarks.py \
--batch_size=256 \
--model=resnet50 \
 --optimizer=momentum \
 --variable_update=replicated \
 --nodistortions \
 --gradient_repacking=1 \
 --num_gpus=$num_gpu \
 --nouse_fp16 \
 --weight_decay=1e-4 \
 --data_dir=/datadir\
 --num_epochs=6 \
 --train_dir=${CKPT_DIR}\
 --data_name=imagenet"
```

```
fi

export CUDA_VISIBLE_DEVICES=0,1,2,3

RUN_CMD="singularity exec -B $DATA_DIR:/datadir --nv $CONTAINER python3
${CMD}"

rm -rf ${SINGULARITY_TMPDIR}
```

The two commands differ by the precision used. The first is in 16-bit flow vector precision (FP16) with option "--use_fp16" and other is in 32-bit flow vector precision (FP32).

10.4.5.4 Measurements and results

For all next measurements, run has been made with "--num_epochs=6" to reduce run-time without lose precision in results.

First, we try to see if it was possible to optimize the batch size in the 2 precisions required:

#nodes	#GPU	Precision	Batch_size	Total images/sec
1	4	FP16	256	4051.20
1	4	FP16	384	4270.86
1	4	FP16	512	4319.65
1	**4**	**FP16**	**640**	**4400.74**
1	4	FP16	704	4391.53
1	4	FP16	768	Out-of-memory

Resnet-50: Batch size study for FP16

#nodes	#GPU	Precision	Batch_size	Total images/sec
1	4	FP32	192	1564.21
1	4	FP32	256	1573.67
1	4	FP32	320	1563.87
1	4	FP32	384	Out-of-memory

Resnet-50: Batch size study for FP32

For following measurements, we will used optimal batch size: 640 in FP16.

We study the scalability in single node:

#nodes	#GPU	Precision	Batch_size	Total images/sec
1	1	FP16	640	1145.04

1	2	FP16	640	1986.94
1	4	FP16	640	4400.74

Resnet-50: Scalability

We can see a very good scalability on one node between 1 and 4 GPU. However, performance with 2 GPUs is under what we can expect, so it's more interesting to use 1 or 4 GPU and not 2.

To perform our projections, we studied the dependency of the code to the CPU and GPU frequencies. Our results are given in the two following tables for one node equipped with 4 GPUs:

#nodes	#GPU	Precision	Batch_size	CPU Frequency (GHz)	CPU Frequency (GHz)	Total images/sec
1	4	FP16	640	Turbo	1.53	4444.50
1	4	FP16	640	2.4	1.53	4400.74
1	4	FP16	640	2.2	1.53	4366.00
1	4	FP16	640	2	1.53	4362.48
1	4	FP16	640	1.8	1.53	4369.09

Resnet-50: CPU frequency study

#nodes	#GPU	Precision	Batch_size	CPU Frequency (GHz)	GPU Frequency (GHz)	Total images/sec
1	4	FP16	640	2.4	1.53	4400.74
1	4	FP16	640	2.4	1.32	4283.37
1	4	FP16	640	2.4	1.14	4014.89
1	4	FP16	640	2.4	0.93	3626.99

Resnet-50: GPU frequency study

As expected, we can see a non-significant impact of the CPU frequency, whereas there is a high dependency on the GPU frequency. The extrapolation will be mainly based on the GPU performance.

As dataset of ImageNet is large, we also tested different filesystem to see if the access to data is a parameter which impact the performance. But we saw almost any differences of performances, so it’s not a limiting factor.

We also have reference run with “--num_epochs=20”, which confirm the viability to reduce number of epoch to measurements in scalability and frequency.

#nodes	#GPU	Batchsize	Precision	Total images/sec
1	4	256	FP16	4051.20
1	4	256	FP32	1573.57

Resnet-50: Results on V100

To compute the score, we extrapolate performance on GPU A100-40 according the following table:

Parallelization method	#nodes	#GPU	Batchsize	Precision	Total images/sec
TF replicated	**1**	4	256	FP16	6758
TF replicated	**1**	4	256	FP32	2414

Resnet-50: Projections on 100 with TF replicated parallelization

The previous extrapolations could be achieved with runs using TF replicated as parallelization method. In addition, we also extrapolate the benchmarks with Horovod as parallelization method and could achieve higher numbers:

Parallelization method	#nodes	#GPU	Batchsize	Precision	Total images/sec
Horovod	**1**	4	256	FP16	8237
Horovod	**1**	4	256	FP32	2929

Resnet-50: Projections on 100 with Horovod parallelization

From which we compute the values committed in the excel sheet:

Req. No.	Value
TEN1	**294 710**

Resnet-50: Value for evaluation

following the formula $Value_{TEN1}$ = (R_A*N_GPU)/4 = (4911,84*(4*60))/4 and R_A = SQRT(fp_16*fp_32) = SQRT(8237*2929).

11 Acceptance Procedure

Atos agrees to the acceptance test defined with details which are specified in this proposition (and more specifically in the Benchmark Report for the part related to benchmarks).

12 Value for Money

Atos has filled the “Performance” and “Value for Money” sheets in Appendix B for IZUM to compute the value for money.

13 Index of Figures

14 Index of tables