VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

THE KAROLINA SYSTEM

BRANISLAV JANSÍK

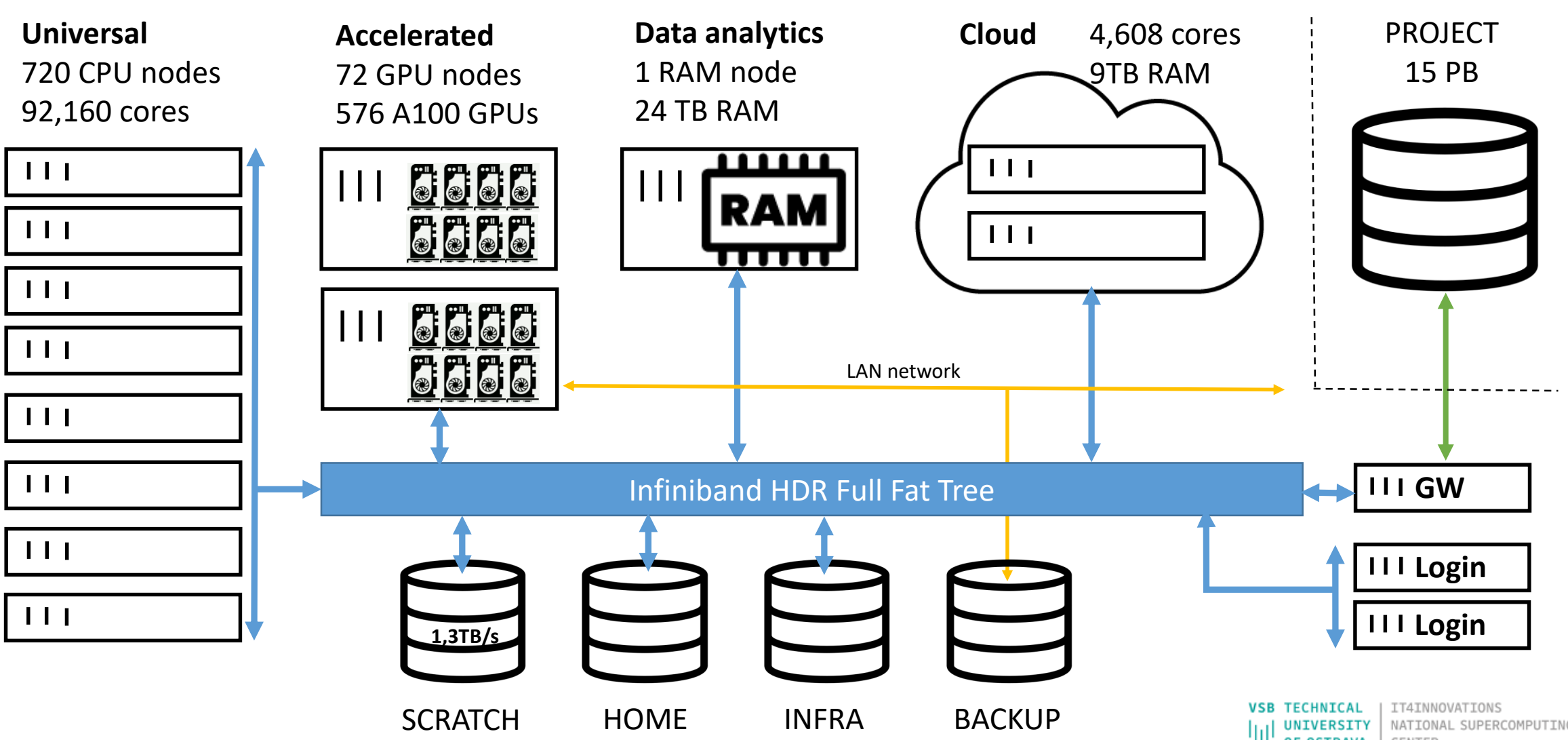
EUROHPC USER DAY
11 DEC 2023

MINISTRY OF EDUCATION, YOUTH AND SPORTS

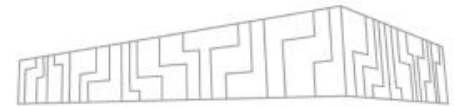
This conference is supported by The Ministry of Education, Youth and Sports from the Large Infrastructures for Research, Experimental Development and Innovations project „IT4Innovations National Supercomputing Center – LM2015070“.

KAROLINA SUPERCOMPUTER

KAROLINA ARCHITECTURE



SLURM SCHEDULER



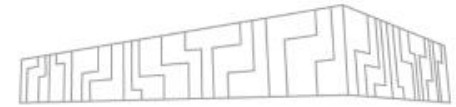
Complete switchover to Slurm

- Slurm on Barbora, Karolina and Complementary systems
- Complete integration to accounting and SCS
- Removed legacy queues, check: $ sinfo -s
- Still optimizing the queue policies: fairsharing and backfilling
- https://docs.it4i.cz/general/job-submission-and-execution/

Scheduler status view

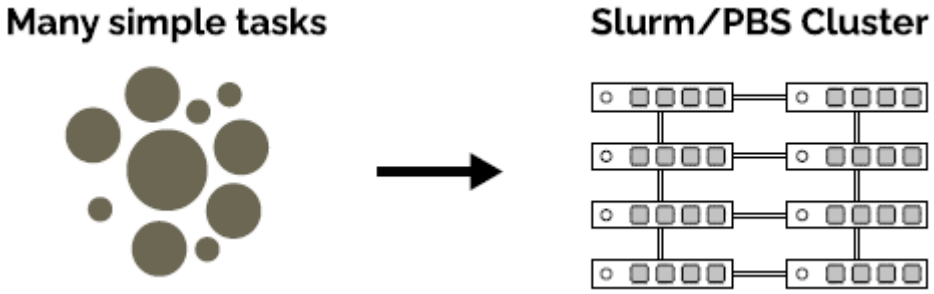
- $ squeue [--time] –all ; squeue –u uname
- https://extranet.it4i.cz/rsweb/karolina/queues?user=uname
  (Includes electricity stats, see O. Vysocky 17:00)

HYPERQUEUE



Problem:

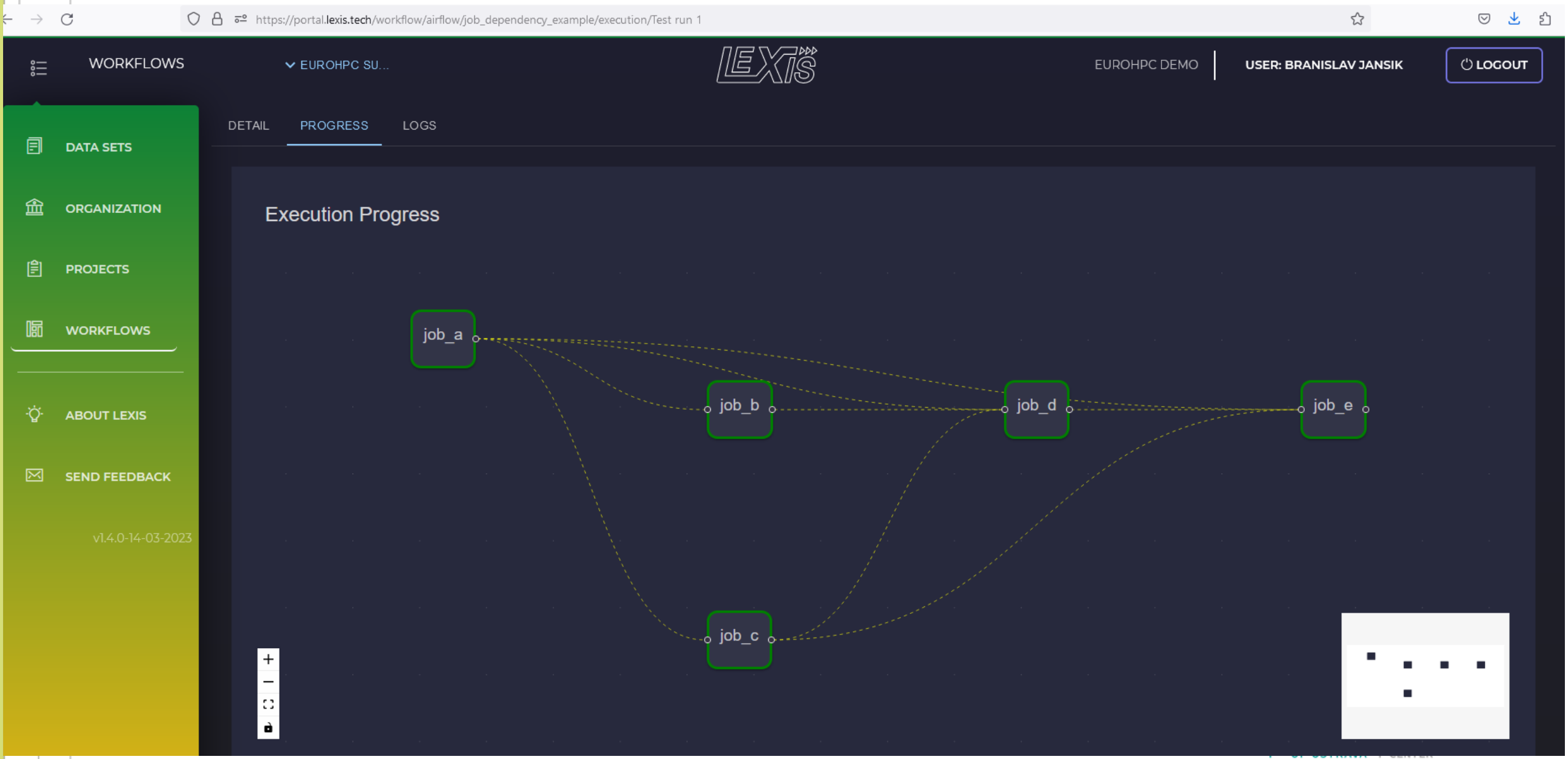
- Large amount of jobs?
- Very large amount of small jobs?
- Only 1 or few cores per job needed?
- Complicated job dependencies?



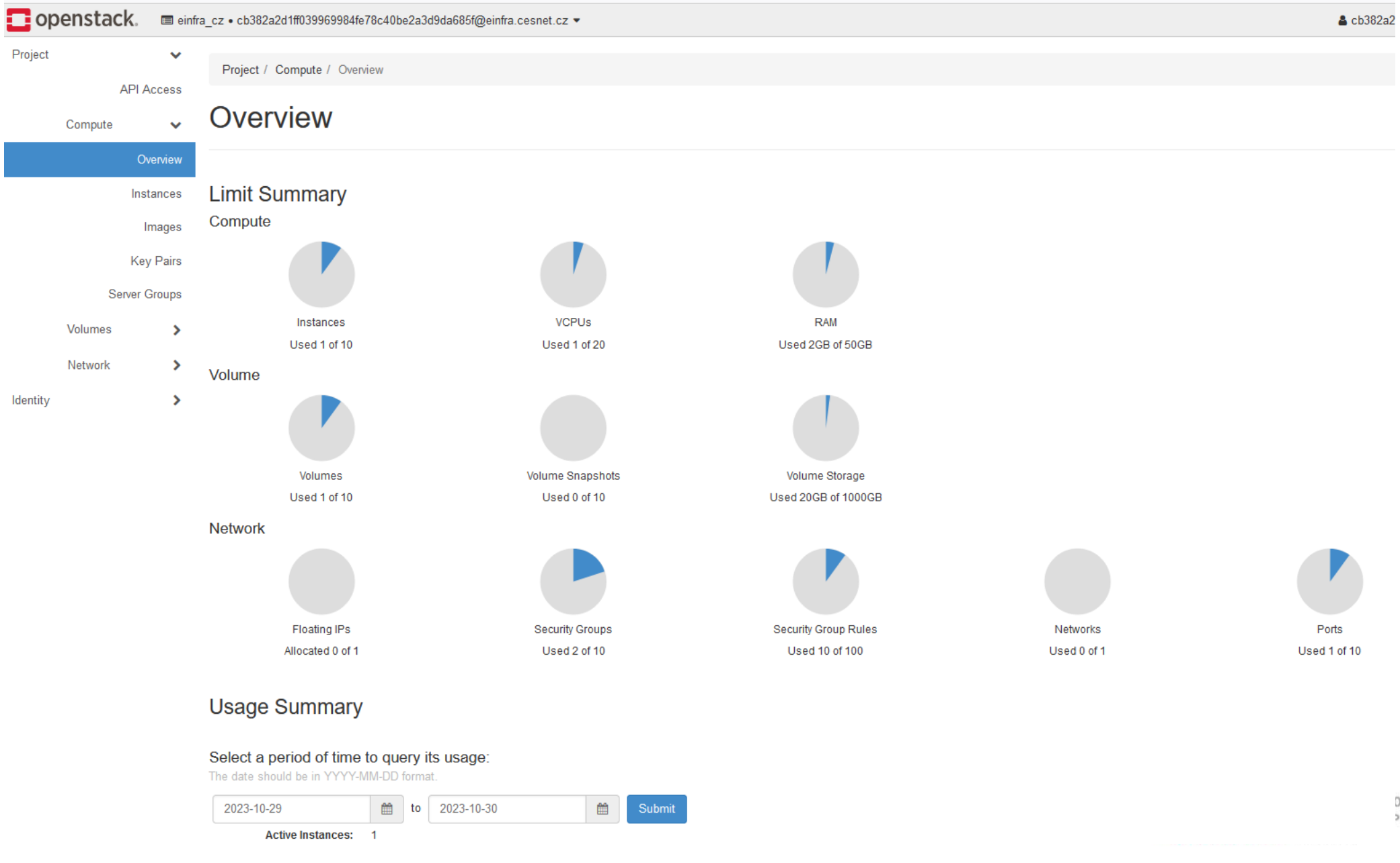
Answer: Use hq!

https://it4innovations.github.io/hyperqueue/stable/
https://docs.it4i.cz/general/capacity-computing/#hyperqueue

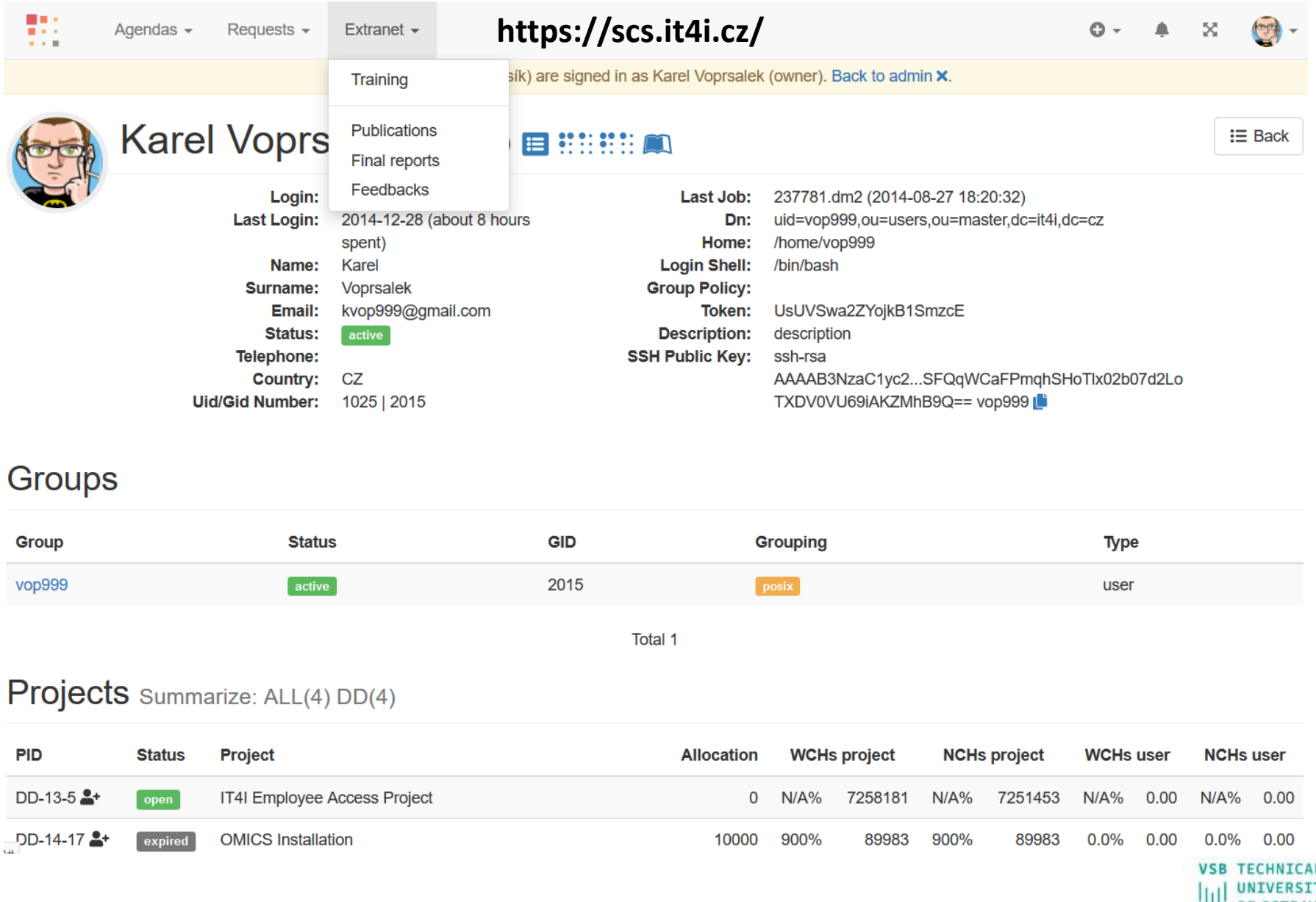
COMPUTE TRENDS - LEXIS PLATFORM



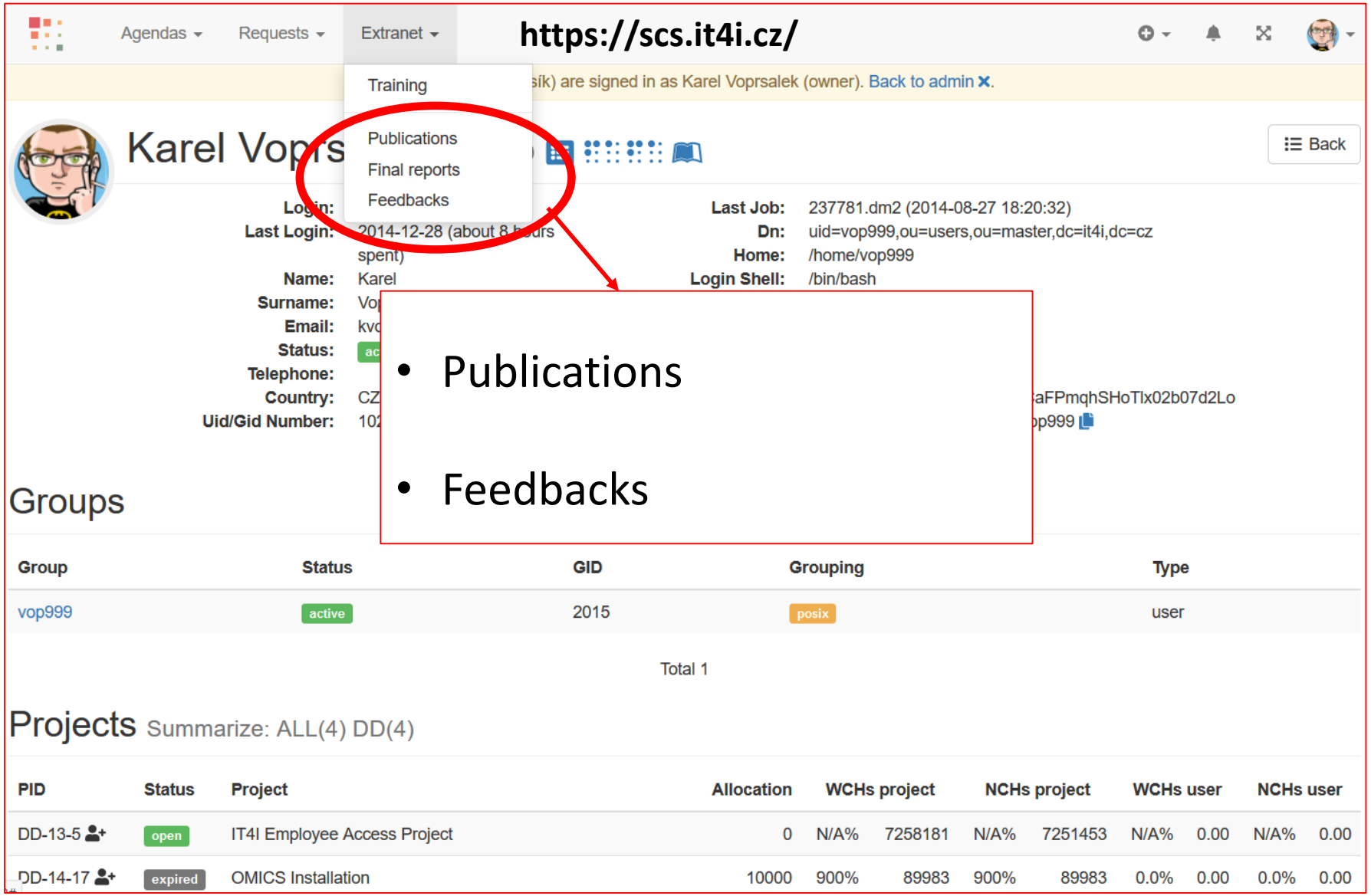
CLOUD AT KAROLINA



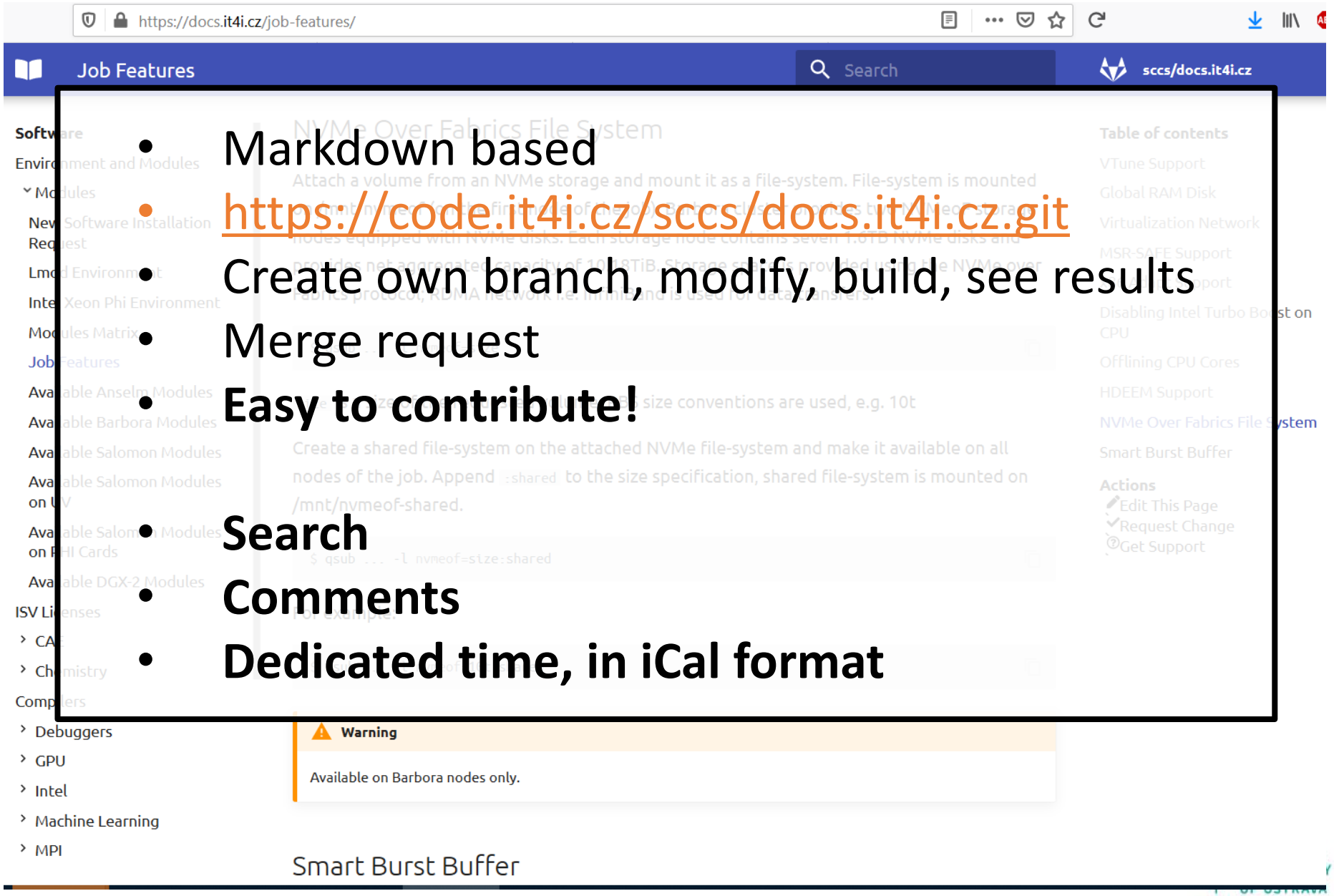
SELF SERVICE PORTAL



SELF SERVICE PORTAL



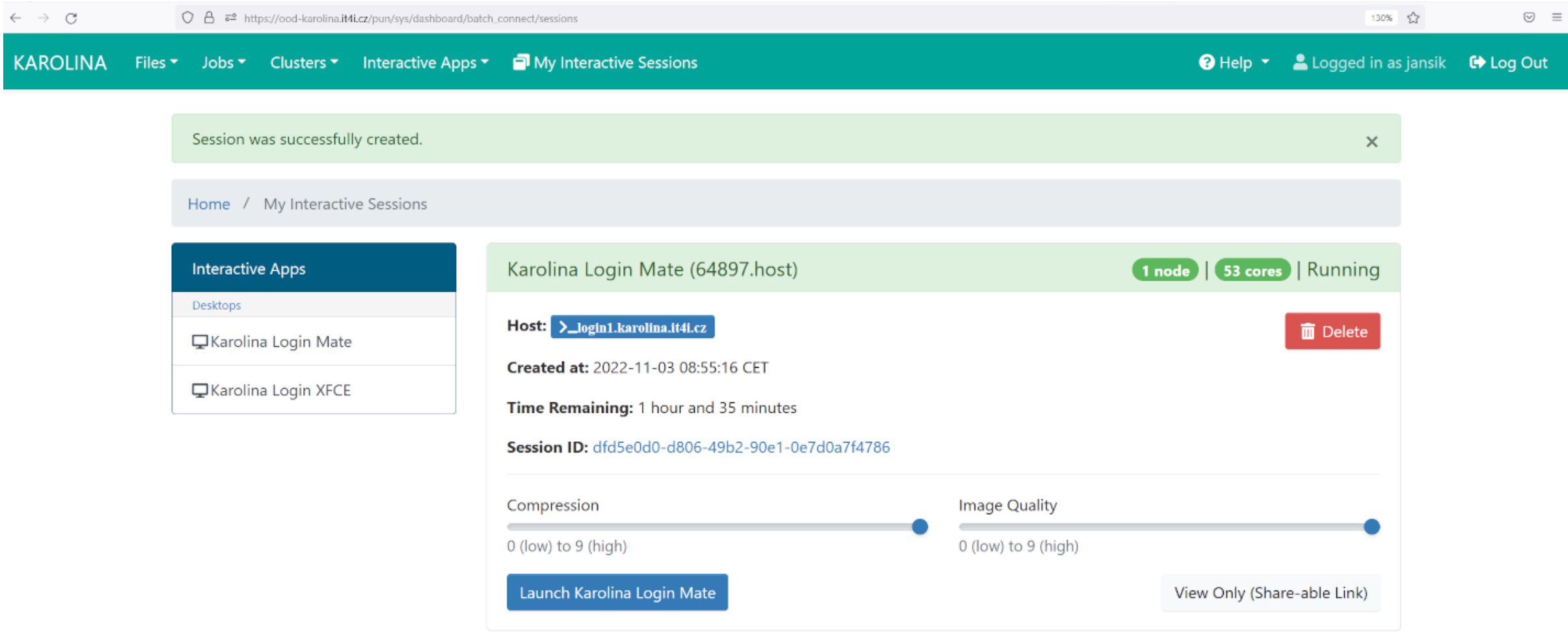
DOCUMENTATION



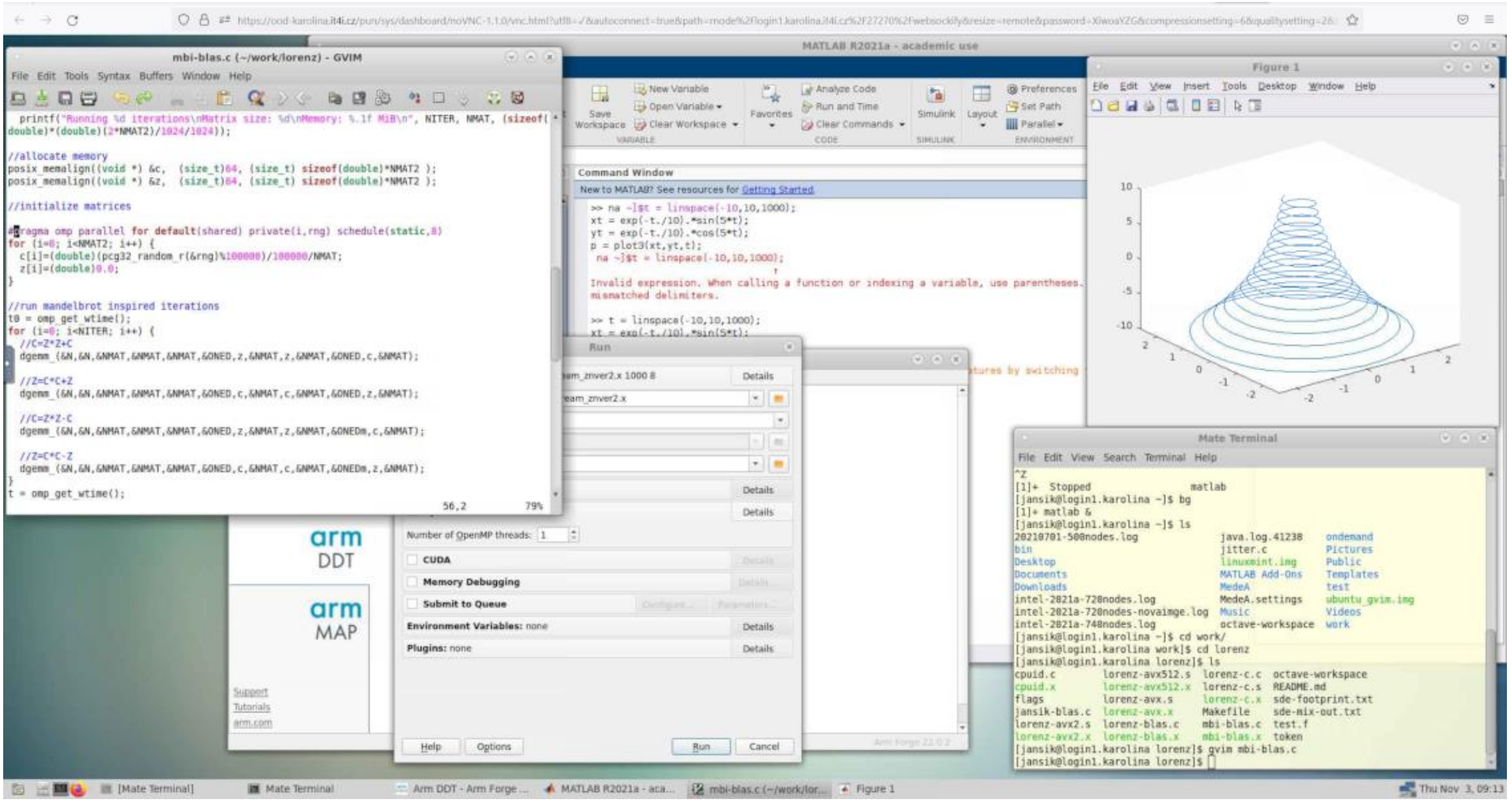
- Markdown based
- https://code.it4i.cz/sccs/docs.it4i.cz.git
- Create own branch, modify, build, see results
- Merge request
- **Easy to contribute!**
- **Search**
- **Comments**
- **Dedicated time, in iCal format**

KAROLINA GUI - OPENONDEMAND

FortiClient VPN + **https://ood-karolina.it4i.cz/**

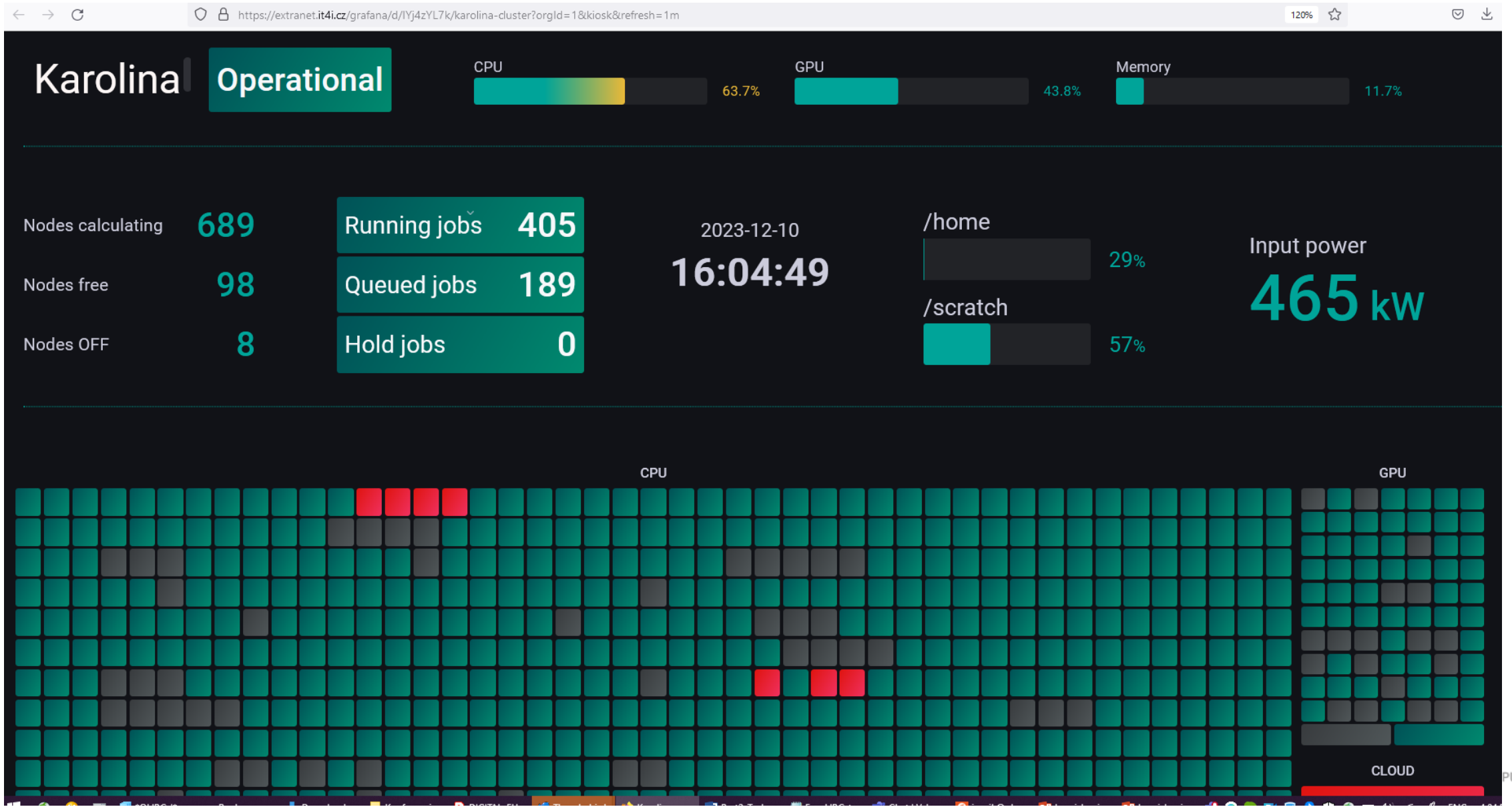


KAROLINA GUI - OPENONDEMAND



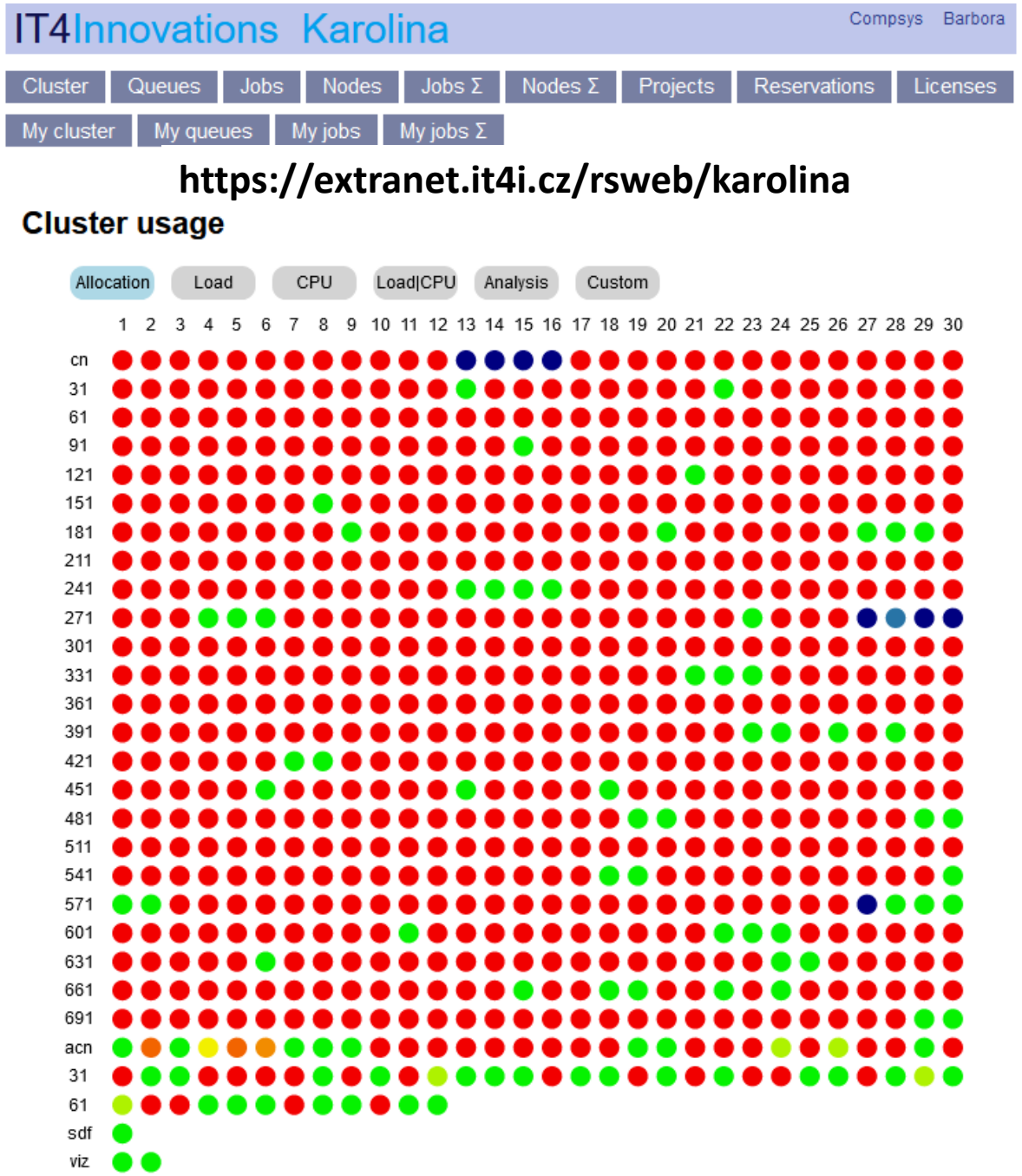
DASHBOARD

https://https://extranet.it4i.cz/grafana/d/IYj4zYL7k/karolina-cluster?orgId=1&kiosk&refresh=1m



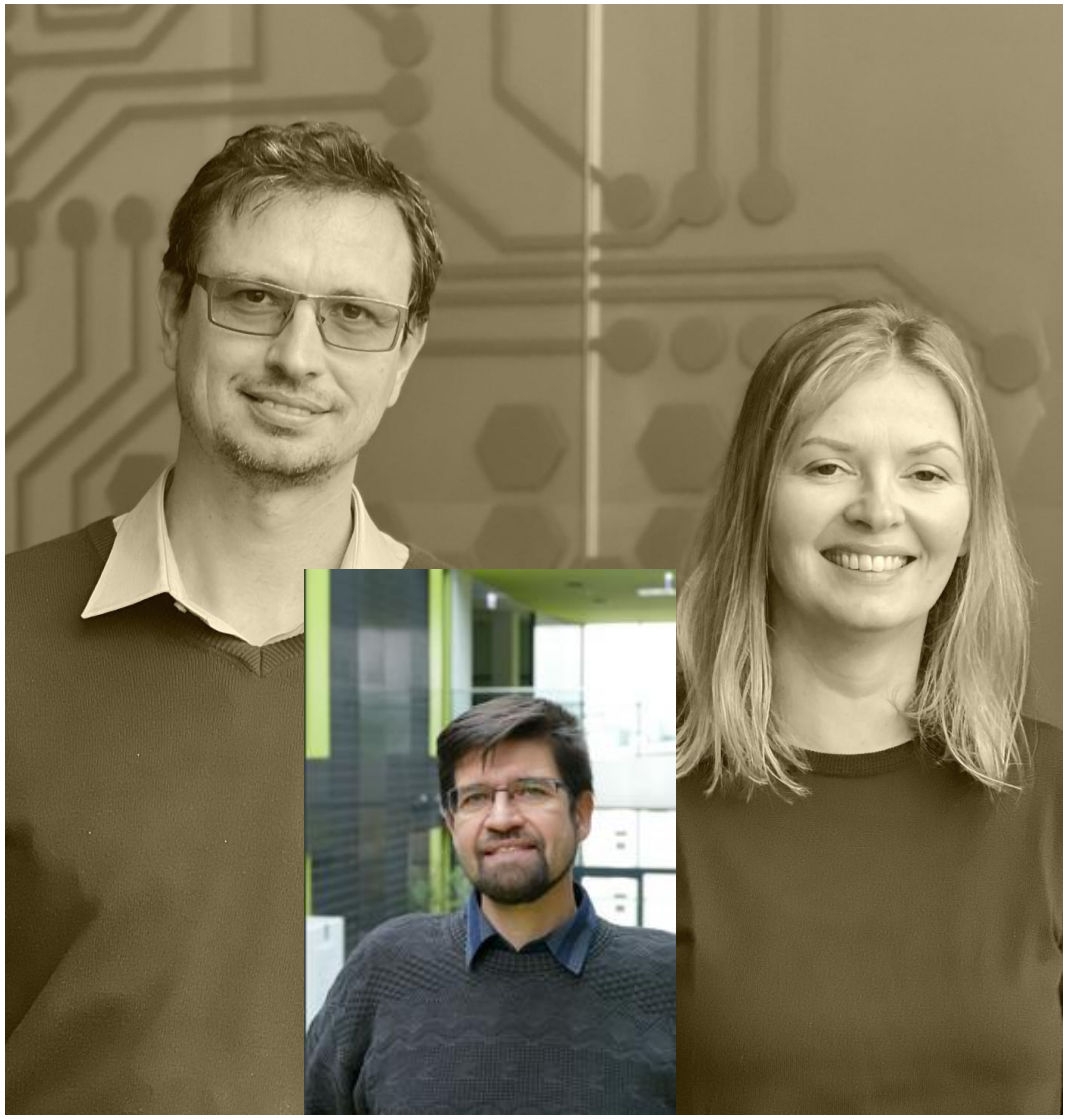
DASHBOARD

- Check node status
- Check job status
- Check graphs and stats
- Check energy consumption



IT4I SUPPORT TEAM

- Ing. Petra Lyčková Navrátilová
  - Team Head
- Ing. Vojtěch Gubani
  - Team Member
- Supported by IT4I SCS admin team members and LUST team members (CZ: Jan Vicherek)



IT4I SUPPORT TEAM BASIC TASKS

Kontakt:
support@it4i.cz
https://support.it4i.cz/

1. Accounts

- Creating accounts (IT4I, Einfra, Lumi) - questions, links
- Administrative questions (conditions, requirements)
- Questions of a technical nature (certificates, technical connection problems, documentation)

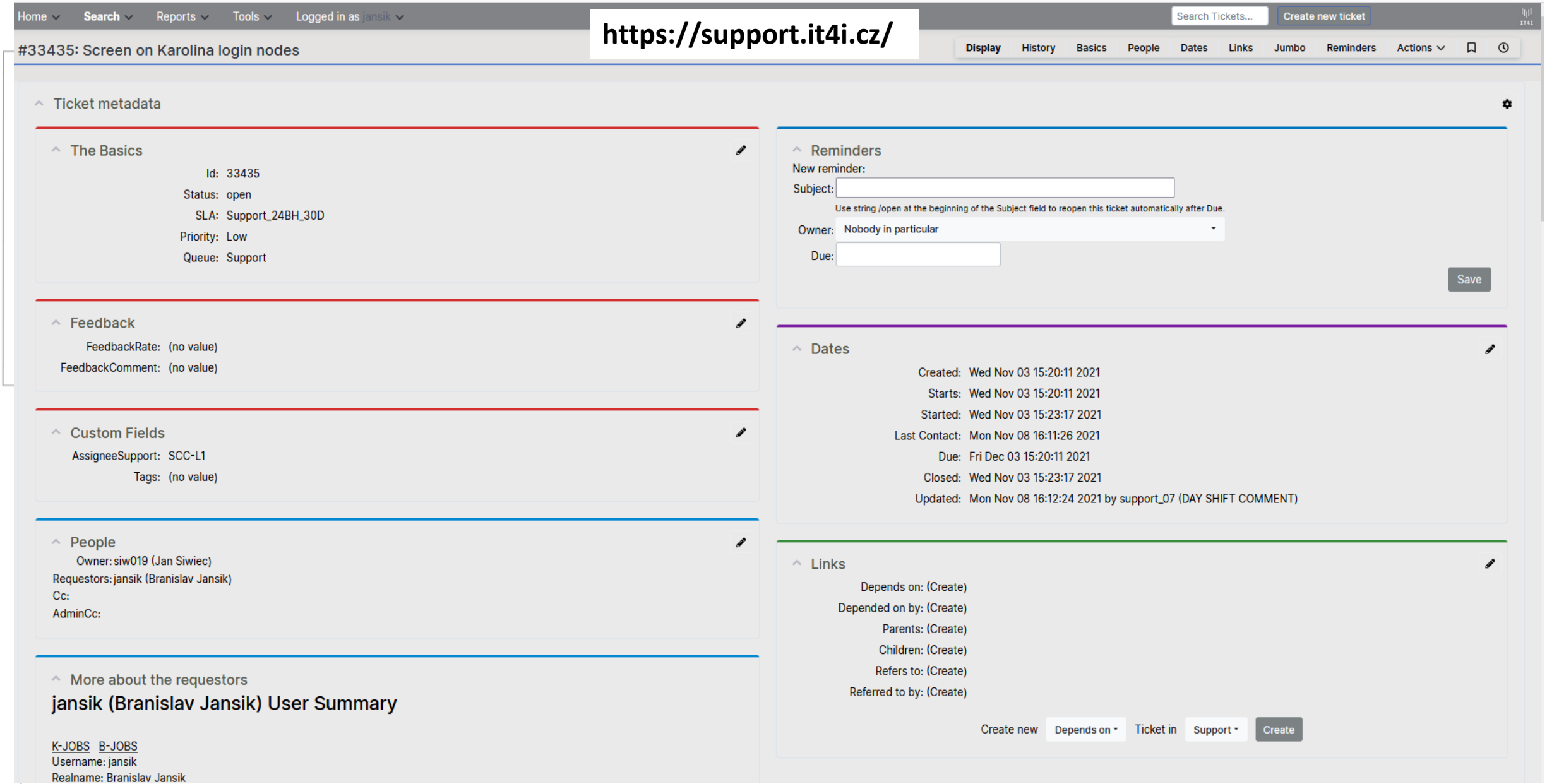
2. Projects

- Project establishment (life cycle)
- Establishing and managing projects within EuroHPC
- Extension and editing, checking the current status of the project
- Adding users to the project

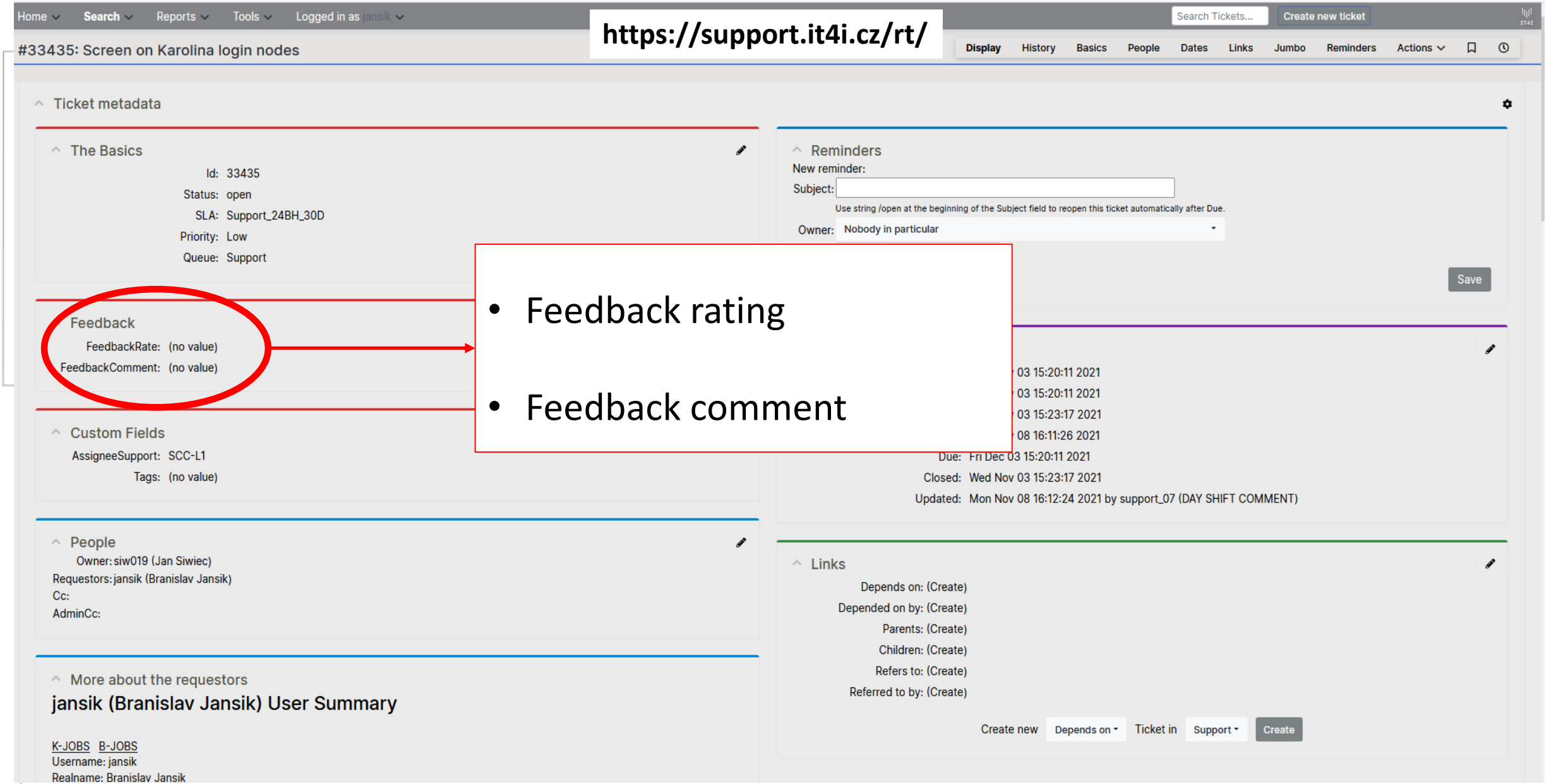
3. Technical questions – cooperation with L2, L3

- Installing and compiling the program
- Counting/status checking
- Reporting of counting errors
- Reservation of computing capacity

REQUEST TRACKER PORTAL

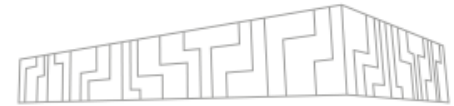


REQUEST TRACKER PORTAL



- Feedback rating
- Feedback comment

TRAINING



https://events.it4i.cz/

Dec 13 - Dec 14
[HYBRID] High-Level Synthesis for FPGA

Dec 12 - Dec 13
[ONLINE] Programming in heterogeneous environments with SYCL and OpenMP offloading

Branislav Jansík
branislav.jansik@vsb.cz

VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

IT4Innovations National Supercomputing Center
VSB – Technical University of Ostrava
17. listopadu 2172/15
708 00 Ostrava-Poruba, Czech Republic
www.it4i.cz

DEUCALION

Comissão Europeia

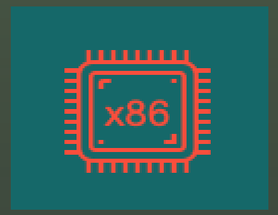
EuroHPC Joint Undertaking

fct Fundação para a Ciência e a Tecnologia

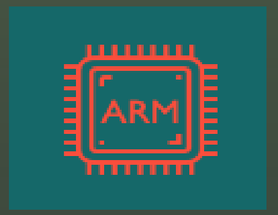
Universidade do Minho

DEUCALION INSTALLATION

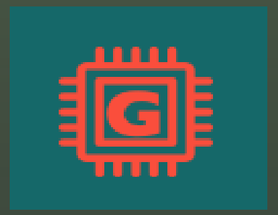
10 PF Supercomputer



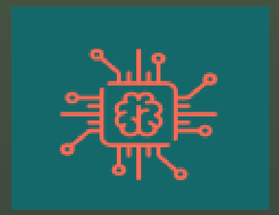
Next generation x86 general purpose system as the powerhouse for conventional digital simulations



State-of-the-art Arm general purpose system with the short-term objective of building HPC science and engineering capacity

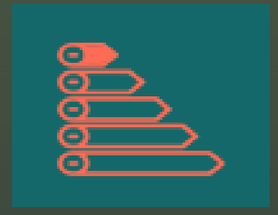


Gpu accelerators on 10% of the x86 subsystem mainly devoted to data science applications



Experimental technologies towards innovative architectures with potential for exascale

High performance dependable storage system with 10 pb net capacity



Ambitious power usage effectiveness (pue) of 1.1

DEUCALION

EuroHPC Joint Undertaking

fct Fundação para a Ciência e a Tecnologia

Universidade do Minho

DEUCALION INSTALLATION

DEUCALION

Comissão Europeia

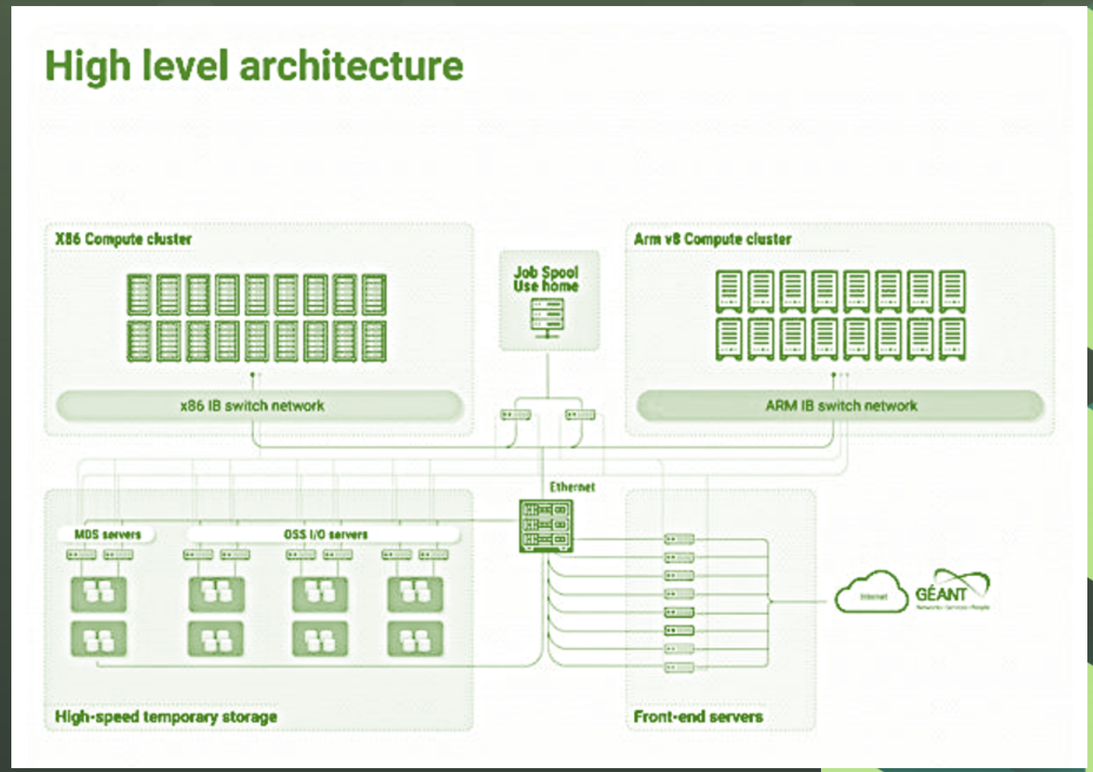
EuroHPC Joint Undertaking

fct Fundação para a Ciência e a Tecnologia

Universidade do Minho

DEUCALION INSTALLATION

- 1670 nodes ARM A64FX
- 500 Nodes AMD Rome
- 33 Nodes AMD Rome + 4 NVidia A100
- 1670 nodes ARM



DEUCALION

EuroHPC Joint Undertaking

fct Fundação para a Ciência e a Tecnologia

Universidade do Minho

DEUCALION INSTALLATION

Partition Name	Node Range	Resource Value
small-arm	cna[0001-1632]	Min node = 1 node Max node = 16 nodes (768 cores) Default walltime = 5 min Max walltime = 2 hours DEFAULT = Yes
normal-arm	cna[0001-1632]	Min node = 1 node Max node = 96 nodes (4608 cores) Default walltime = 5 min Max walltime = 48 hours (2 days)
large-arm	cna[0001-1632]	Min node = 1 node Max node = 256 nodes (12288 cores) Default walltime = 5 min Max walltime = 48 hours (2 days)

DEUCALION

Comissão Europeia

EuroHPC Joint Undertaking

fct Fundação para a Ciência e a Tecnologia

Universidade do Minho

DEUCALION SLURM QUEUEING

Partition Name	Node Range	Resource Value
small-x86	cnx[001-500]	Min node = 1 node Max node = 8 nodes (1024 cores) Default walltime = 5 min Max walltime = 2 hours DEFAULT = Yes
normal-x86	cnx[001-500]	Min node = 1 node Max node = 24 nodes (3072 cores) Default walltime = 5 min Max walltime = 48 hours (2 days)
large-x86	cnx[001-500]	Min node = 1 node Max node = 64 nodes (8192 cores) Default walltime = 5 min Max walltime = 48 hours (2 days)
a100-40	gnx[501-516]	Min node = 1 node Max node = 2 node (8 GPU cards) Default walltime = 5min Max walltime = 48 hours (2 days)
a100-80	gnx[517-533]	Min node = 1 node Max node = 2 node (8 GPU cards) Default walltime = 5min Max walltime = 48 hours (2 days)

DEUCALION

Comissão Europeia EuroHPC Joint Undertaking fct Fundação para a Ciência e a Tecnologia Universidade do Minho

USER SUPPORT PORTAL
(UNDERDEVELOPMENT)

- System Status and Health

- User Account Request
- User Management
  - Contact Profile Information
  - SSH Keys Management
  - TOPT Device Management

- Project Allocation Management (EuroHPC/National)
  - Available Resources
  - Team Members
  - Resource Consumption Report
  - Energy Consumption Report

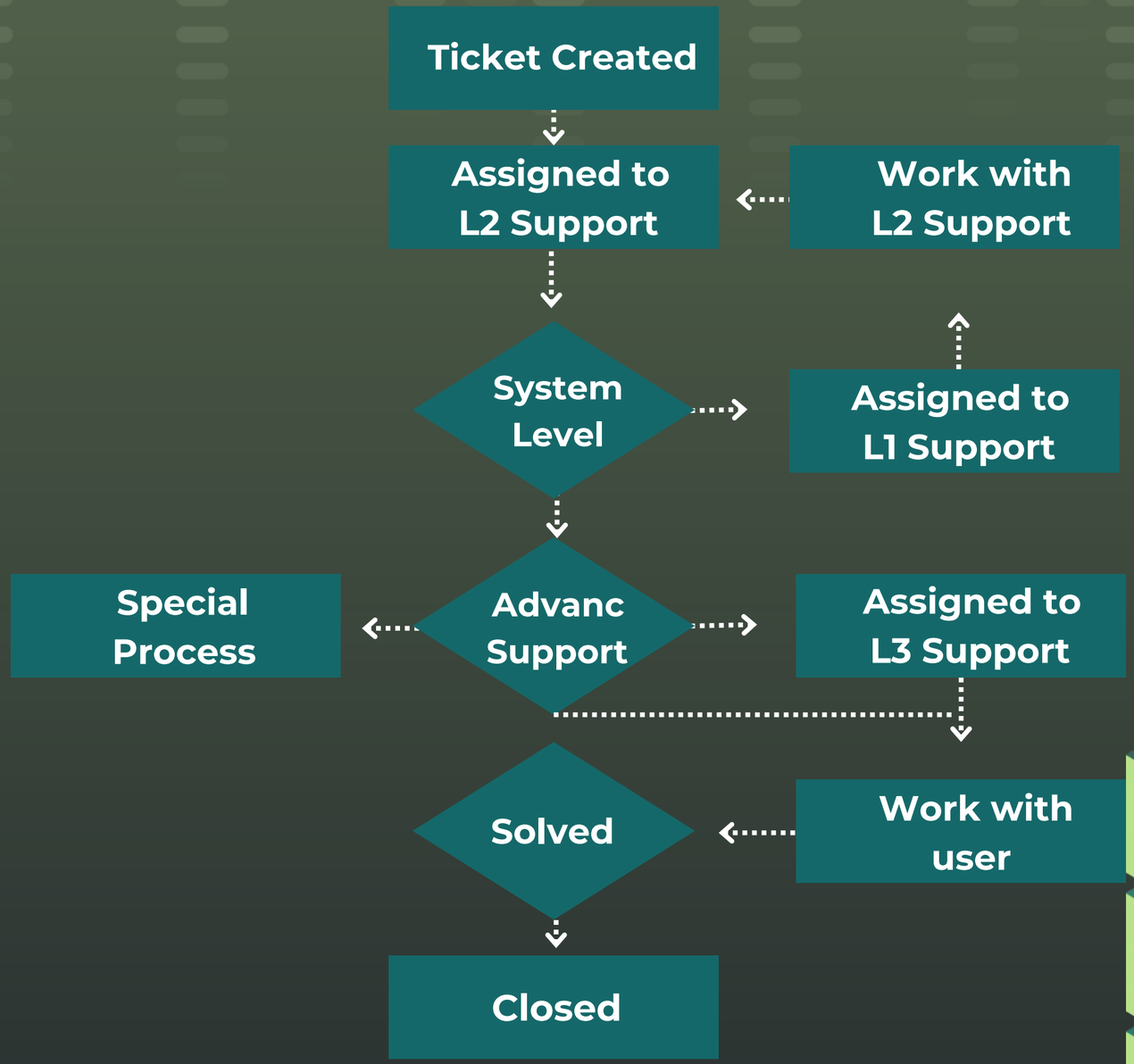
DEUCALION

Comissão Europeia EuroHPC Joint Undertaking fct Fundação para a Ciência e a Tecnologia Universidade do Minho

USER SUPPORT PORTAL
(UNDERDEVELOPMENT)

- Knowledge Base
  - Tutorials
  - Documentations
  - Best Practices

- User Ticket System
  - User Tickets
  - Support Chat



DEUCALION

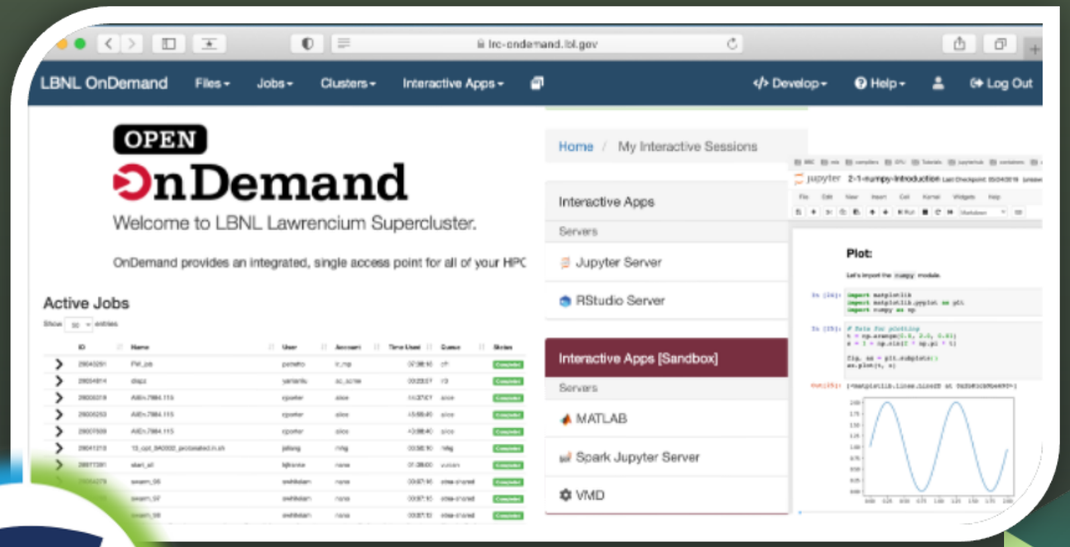
EuroHPC Joint Undertaking

fct Fundação para a Ciência e a Tecnologia

Universidade do Minho

DEUCALION ACCESS + ADDITIONAL SERVICES

- Usual SSH Access
  - SSH Key + TOPT
- Open OnDemand
  - Jupiter Notebooks/Lab
  - R Studio
  - X/Gnome GUI Environment
- Singularity Containers



DEUCALION

EuroHPC Joint Undertaking

fct Fundação para a Ciência e a Tecnologia

Universidade do Minho

AVAILABLE RESOURCES

	X86	ARM	GPU
Total	**560.640.000**	**686.220.000**	**1.156.320**
EuroHPC	**196.224.000**	**24.087.700**	**404.712**

DEUCALION

Comissão Europeia

EuroHPC Joint Undertaking

fct Fundação para a Ciência e a Tecnologia

Universidade do Minho

THANK YOU

João Barbosa
jbarbosa@macc.fccn.pt

DEUCALION

Comissão Europeia

EuroHPC Joint Undertaking

fct Fundação para a Ciência e a Tecnologia

Universidade do Minho

ACHIEVEMENTS

- Building of the Year 2021 in Bulgaria
- 91st in the world's TOP 500 Supercomputers (oct 2023 166)
- Completion in time – 6 months
- 1,10 MW cooling
- Full automation
- 100% free cooling (depending on the weather conditions)
- Heat utilization for common use

GREEN & ENVIRONMENTAL FRIENDLY

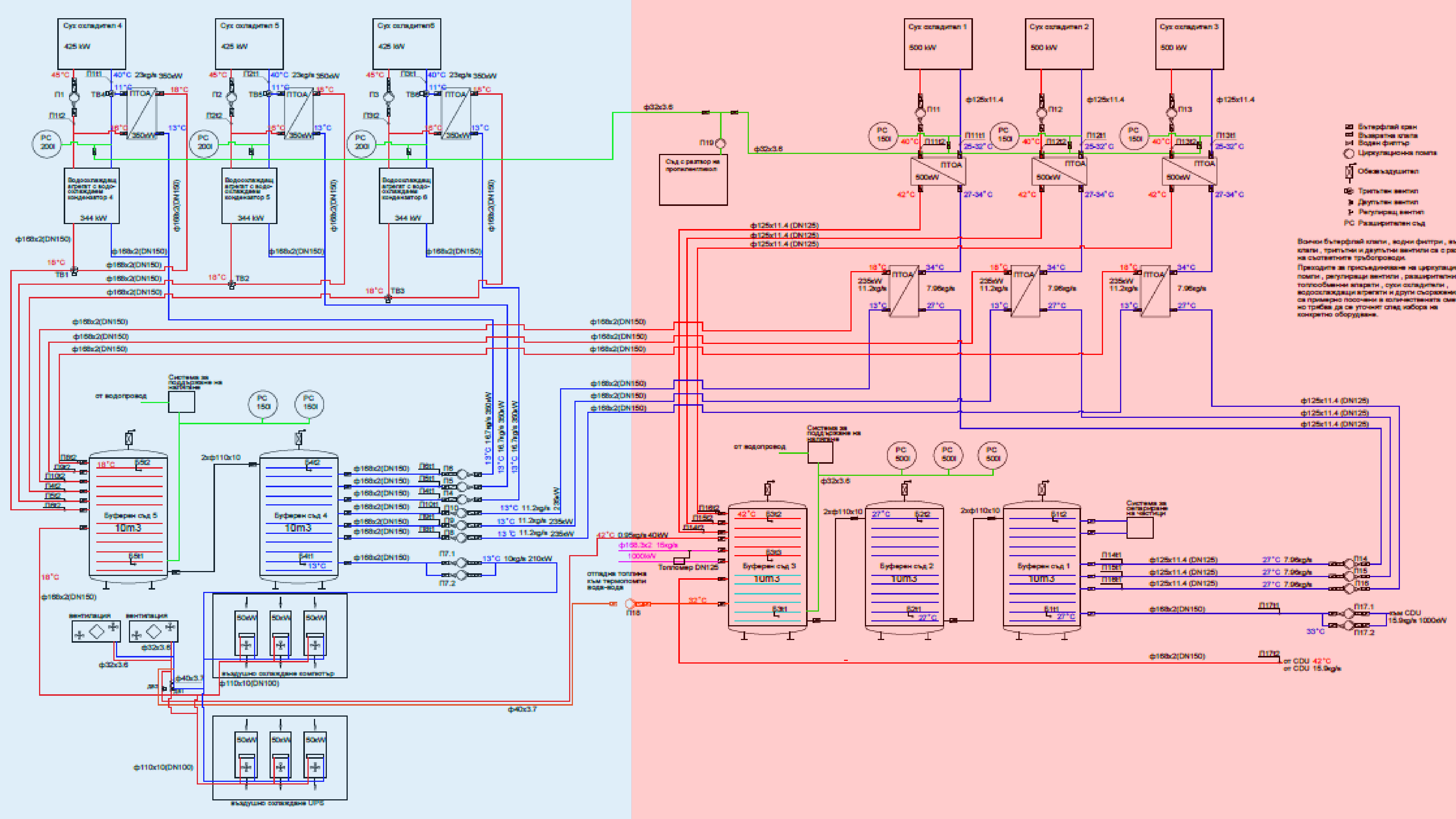
COOLING SYSTEM CHALLENGES

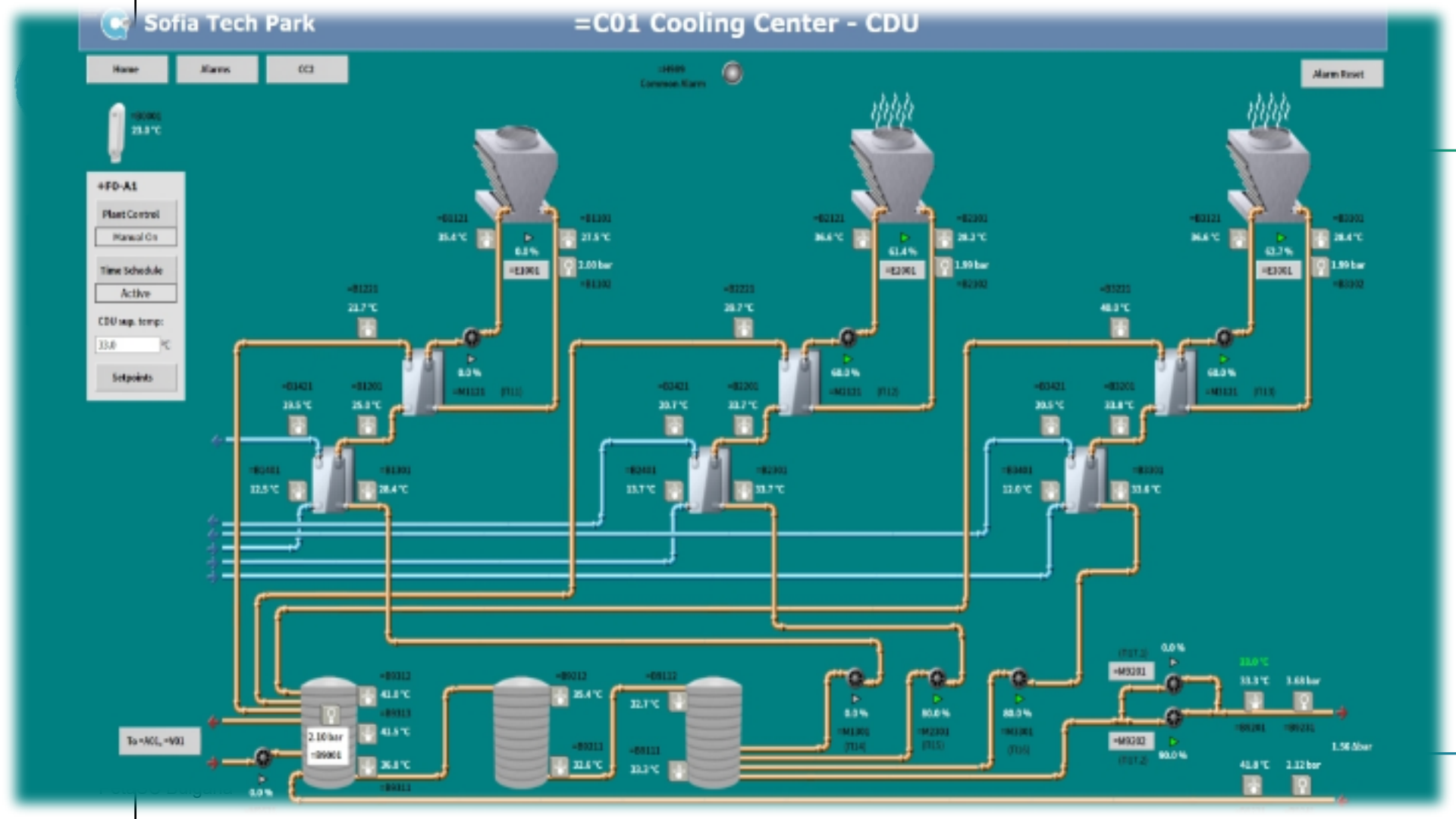
Planning in 2020

- Extreme quality benchmarks
- Sustain the future – Make it green
- Iterative approach due to constantly improving technology of the Supercomputer
- Build to suit
- Customized engineering solutions
- N+1 Redundancy

Doing in 2021

- COVID-19 & Lockdown
- Supply chain obstructions
- Winter season
- Tight time schedule
- International stakeholders' collaboration
- Strict EU & State government supervision


Сух охладител 4
425 kW
Сух охладител 5
425 kW
Сух охладител6
425 kW
Сух охладител 1
500 kW
Сух охладител 2
500 kW
Сух охладител 3
500 kW
ПТОА
Водоохлаждащ агрегат с водоохлаждаем кондензатор 4
344 kW
Водоохлаждащ агрегат с водоохлаждаем кондензатор 5
344 kW
Водоохлаждащ агрегат с водоохлаждаем кондензатор 6
344 kW
Съд с разтвор на пропиленгликол
Буферен съд 5
10m3
Буферен съд 4
10m3
Буферен съд 3
10m3
Буферен съд 2
10m3
Буферен съд 1
10m3
Система за поддържане на налягане
от водопровод
Система за сепариране на частици
вентилация
въздушно охлаждане компютри
въздушно охлаждане UPS
Топломер DN125
отпадна топлина към термопомпи вода-вода
към CDU
от CDU
Бътерфлай кран
Възвратна клапа
Воден филтър
Циркулационна помпа
Обезвъздушител
Трипътен вентил
Двупътен вентил
Регулиращ вентил
РС Разширителен съд


Sofia Tech Park
=C01 Cooling Center - CDU
Home
Alarms
CCI
Alarm Reset
+FD-A1
Plant Control
Manual On
Time Schedule
Active
CDU sup. temp:
Setpoints

What is DISCOVERER

Sofia Petascale Supercomputer

- **Discoverer** is a Petascale supercomputer can execute:
  - 4,6 PetaFlops Rmax
  - 6,0 PetaFlops Rpeak

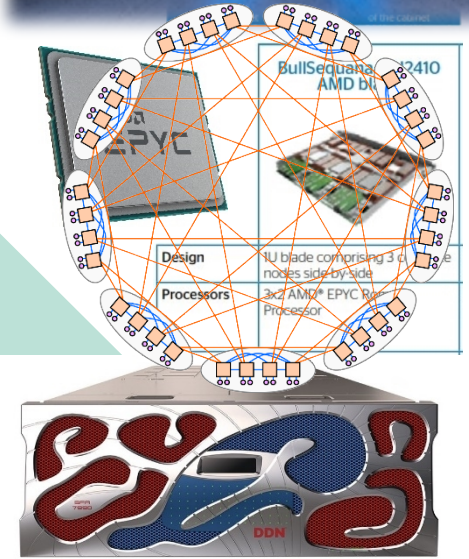
  [1 PetaFlops= 10^{15} Flops = 10^{6} Flops x 10^{9} Flops]
- **In** in oct 2021 **Discoverer** was ranked at 91st place among the worlds top 500 supercomputers (in oct 2023 it is 166th)
- Discoverer's infrastructure is **co-funded by EuroHPC JU (35%) and by PetaSC** and the Bulgarian government (65%).
- **PetaSC Bulgaria** is a legal consortium combining the knowledge and 15 years of expertise
- **Discoverer's mission & vision:**
  - **T**o foster better science for society
  - **T**o facilitate innovations by establishing deeper collaborations between academic institutions and the business
  - **T**o help training the next generation IT talent

166th (91st)

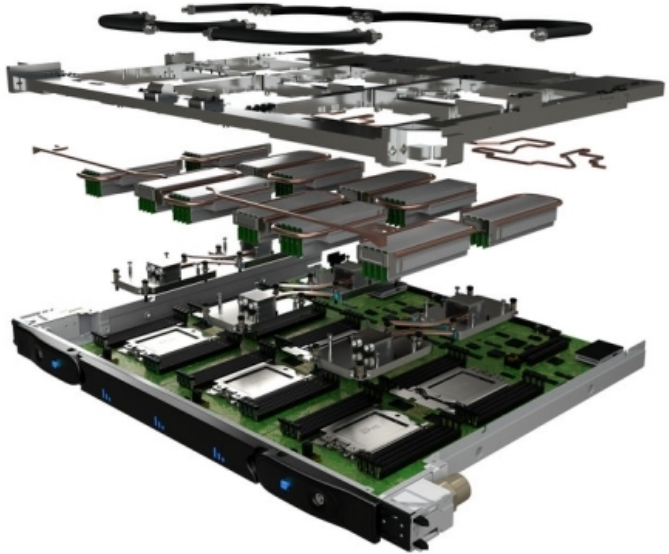
TOP 500 The List.

System Architecture & Specs

DISCOVERER
Sofia Petascale Supercomputer



- **12** Direct Liquid Cooling **BullSequana XH2000 Racks** with up to **32 blades per rack. i.e. Discoverer has 376 blades (12x32-8=376)**
- **376 blades** x **3 nodes/blade** = **1128 computing nodes**
  - **2** x **AMD EPYC processors per node i.e. Discoverer has in total of 2256 AMD EPYC CPUs (2.6 GHz normal freq)**
  - **256 GB shared memory per node with 18 fat nodes with 1024GB RAM (3200MT/s DR)**
- **es**
- (1128x256+18x(1024-256)=316
- **n total capacity of 2 PB (LustreFS)**
- and storage) infrastructure
- **DR with DragonFly+ topology**
- ower failure using an ut of **1 MW.**
- **total power consumption of**



Additional Resources & Forthcoming upgrade

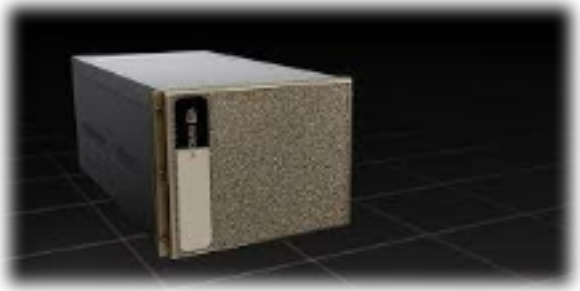
Additional Infrastructure (EuroHPC independent)

- 2 Palo Alto FW 5250 running in active-passive mode
- Separate DiscoCloud portion and small storage for containers (Kubernetes) – targeted commercial use and costumers with special recrements
- Access to Discoverer resources via Nimbix (HPC as a Service – commercial use only)
- Secondary (block NFS) storage - 130TB NVMe and 92TB SSD

Firewall throughput2 (App-ID enabled)	35.9 Gbps
Threat prevention throughput3	20.3 Gbps
IPsec VPN throughput	14 Gbps
Max sessions	8,000,000

Forthcoming Discover+ upgrade

- Small GPU partition based on NVIDIA H100 (4 DGX with 8x100H or equivalent)
- Storage upgrade (mixed block and object)– 1-0.5 TB NVMe and 4-6TB SSD/HDD
- Secondary UPS (200kW)
- Security software



Discoverer Access Policy

- **EuroHPC has 35% resource share from the Discoverer's resources, which will be allocated via regular calls for:**
  - Regular HPC projects (~80-90%) (for projects that require significant amount of HPC resources)
  - HPC project benchmarking (few %) (to determine project suitability & scalability, which is prerequisite to apply for a regular project)
  - HPC Software Development & benchmarking (few %)
  - Fast Track applications (~10%) (for example to get extra data quested during the peer-review process of previous project)
- EuroHPC calls will be organised via **PRACE** and are targeting pan-EU HPC projects including academic & industrial applications (with 80:20 split – see bellow)

- **PetaSC's resource share is 65% and it will be allocated using similar to EuroHPC access policy**
- PetaSC calls will be organised by PetaSC's Scientific & Applicant Board and the applications will be assessed by national & international scientific panel.

- For both EuroHPC (35%) and PetaSC (65%) share, there an additional **(min)80:(max)20** split between **free access** for purely scientific/non-profit and **paid access** for industrial/for-profit applications.
  - the distinction between free & paid access will be made on the base of results dissemination, where free access projects have obligation to openly share the all results from the project, while paid access calls can retain (part of) their results and use them for-profit /business proposes.
  - i.e. consortium from academic & industrial partners can qualify for free access to the Discoverer, provided they prepare suitable project proposal, apply for a call and get selected by the panel and after project end to openly share all the data and results from the project.

Resources and Control Policy

- **Slurm is employed as a queue manager and controls the utilization of computation resource:**
  - By default, (if no account is specified) the jobs are assigned to the default account “normal”, where no resources are available
  - For each HPC project all team members utilize a pool of resources (limits) defined by a Slurm account: GrpTRES, GrpTRESMins, MaxJobs, MaxSubmit, MaxWall
  - At least one QoS is associated to every Slurm account to control the way the resources are available to spent or to re-define some of the limits (the default QoS just imposes the flags DenyOnLimit and NoDecay)
  - Additional QoS is associated to certain Slurm accounts upon request if some of the limits described in the Slurm account need to be redefined, for instance running the job or more nodes than the account allows, but with reduced wall time (re-defines MaxWall)
  - Some of the development projects may require QoS-es to set different job priority in case they need a test queue with lesser job priority and shorten wall time
- **Good vs bad behaviour**
  - Estimating and maximizing the productivity of long runs based on series of short runs is considered **good behaviour** (less electricity spending to derive more results)
  - Using the optimized libraries and other application software available in the public software repository (maintained by our support team) in favour of creating an uncontrolled number of Conda virtual environments (leads to extra occupation of the file system) is also considered a **good behaviour**
  - Reporting problems immediately after they are detected is an example of a **good behaviour**
  - Running jobs on the login nodes is considered a **very very bad behaviour**
  - Running poorly optimized software and simulations is always considered **very very bad behaviour**

Resources, Helpdesk & Energy

- **Due to some limitation related to the employed object storage (Lustre), we currently support only Singularity containers**
  - Currently only one project actively uses Singularity containers (related with image processing from the telescope in Rozhen observatory)
  - A new storage system is being implemented and it will allow to provide local registries with pre-defined rootless Podman containers
  - NVMe over Fabrics is already available on Discoverer's secondary block storage (until the upgrade available only for selected number of EuroHPC users)
- **Helpdesk is organized to be time-effective:**
  - In 99.3% of the cases, the reported issues are solved within 24 hours
  - In 0.5% of the cases it takes up to 72 hours to address the issue (compiling and testing new libraries upon request, bugtracking software, profiling)
  - In 0.2% of the cases it takes more than 72 hours to solve the reported problems
- **Documentation is public and available online:**
  - We support the web-portal https://docs.discoverer.bg whereupon most of the information needed by the users for accessing discoverer and running jobs is available
  - The documentation portal will receive a massive update in January
  - **Python-related documentation** is available as a separate web-portal at https://py.discoverer.bg
- **In H1 2024 we will start tracking the energy use:**
  - Slurm can collect information about the power consumption related to each running job across the cluster
  - To estimate the indirect power consumption we want to include in the estimate the electricity spent on cooling and running the auxiliary systems (switches, monitoring hosts, UPS-es)

Software & Target Application Areas

Open-source Software (short list)

- Bioinformatics / Genomics (BLAST/RAY/EXCALATE/(HAD)DOCK/ROSETTA)
- Computational & Quantum Chemistry (CP2K/CPMD/Quantum Expresso/ GAMES)
- Molecular Dynamics & Mesoscale Modelling, Monte Carlo (GROMACS/NAMD/LAMMPS/DL POLY)
- Computational Fluid Dynamic / Finite Elements Methods (Open FOAM/Alya/SALOME)
- AI / Big Data Analytics (Tensor Flow/Python ML Libraries /NEURON)

Application Areas:

- *In-silico* Drug Discovery
- Structure-Property Predictions & Molecular Discovery
- Digital Product Formulation & Optimization
- Climate & Weather Forecasting / Environmental Modelling
- Simulated Environments in Automotive & Civil Engineering
- FinTech/MarkeTech & Big Data (DL/ML/AI)

HPC Vega quick facts

- **1st operational EuroHPC JU system**
- **In production since April 2021**
- **HPC Infrastructure 17 MIO EUR**
- **Operation est. costs 30 MIO EUR in 8 years**

- Atos BullSequana XH2000
- Performance 6.9 PFLOPS
- 1020 Compute nodes (60 with 4xA100 GPUs), Infiniband 100Gb/s
- 18 PB Large Capacity Storage Ceph
- 1 PB High-Performance Storage Lustre
- Hyper-Connected 600 Gb/s

TOP500 LIST - JUNE 2021

R_{max} and R_{peak} values are in TFlops. For more details about other fields, check the TOP500 description.

R_{peak} values are calculated using the advertised clock rate of the CPU. For the efficiency of the systems you should take into account the Turbo CPU clock rate where it applies.

Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
106	**VEGA HPC CPU** - BullSequana XH2000, AMD EPYC 7H12 64C 2.6GHz, Mellanox InfiniBand HDR100, Atos IZUM Slovenia	122,880	3,822.0	5,367.0	
134	**VEGA HPC GPU** - BullSequana XH2000, AMD EPYC 7H12 64C 2.6GHz, NVIDIA A100, Infiniband HDR, Atos IZUM Slovenia	33,600	3,096.0	4,680.0	

- Power consumption < 1MW
- Power Usage Efficiency, PUE < 1.15

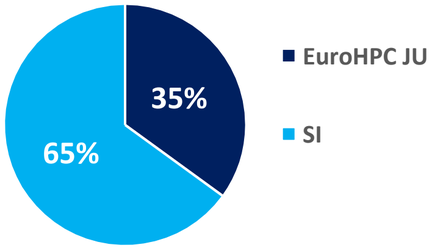
Share of Resources

65% of all capacities – national share, divided:

- up to 70% for Open Access
- up to 20% for Commercial Access (Industry and other)
- from 10% upwards is Hosting Entity (IZUM) reservation
  - Community Access, Urgent Priority Usage
  - Continuous Maintenance

35% of all capacities – EuroHPC JU share:

- any EU organization can apply for calls
- independent calls for allocations
- users and projects must be approved by EuroHPC JU



https://doc.vega.izum.si/shares/

Resources available for EuroHPC JU

Partition	Nodes	Core hours	Storage
CPU standard	269	301.400.064	6,3 PB
CPU large memory	67	75.125.760	\
GPU partition	21	23.546.880	\

EU Projects supported

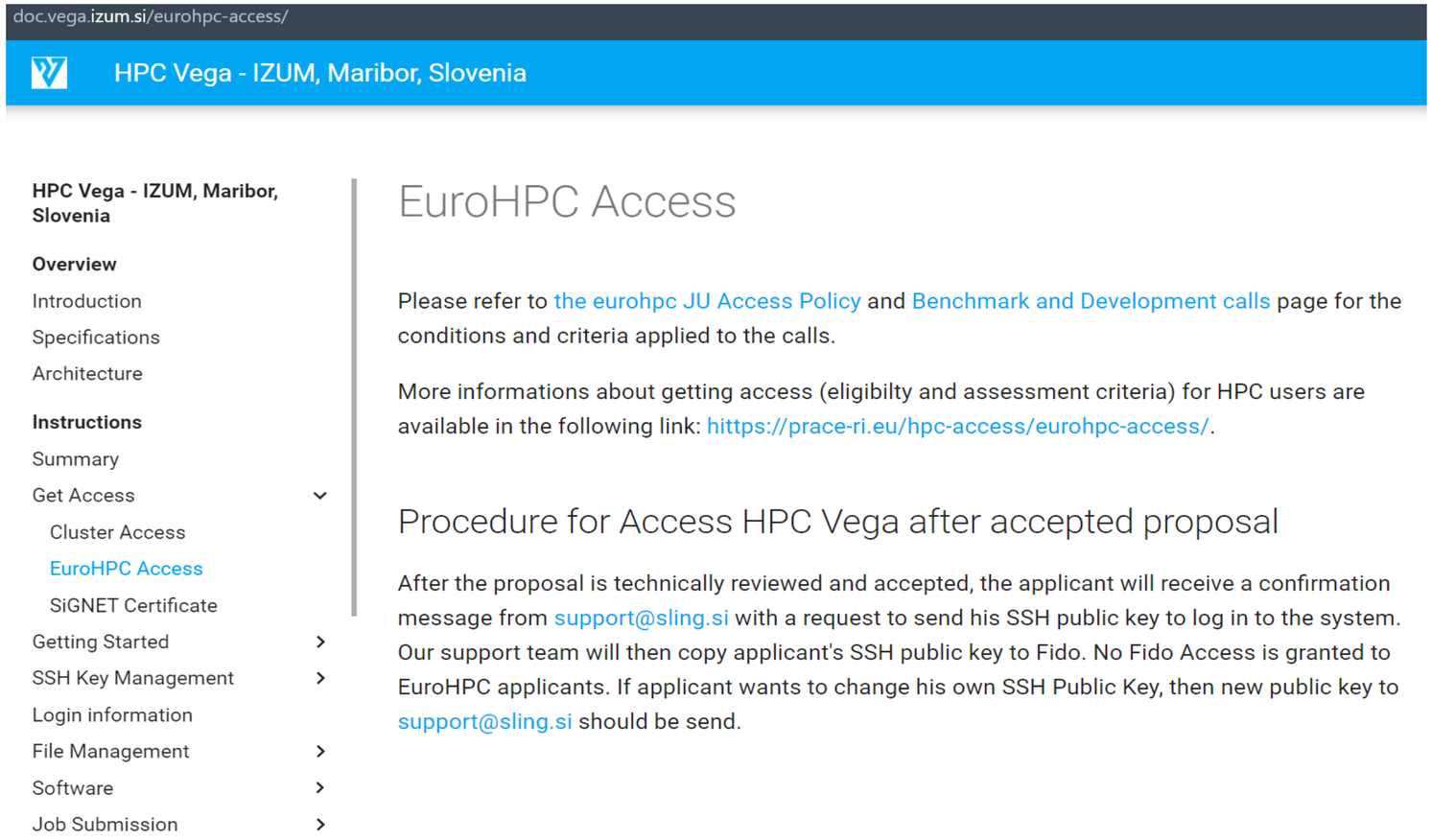
Funded projects supported:

- InterTwin:
  - Full exploitation of HPC Vega infrastructure
- EPICURE (Adv. App. Support)
  - 2 FTEs (1 IZUM, 1 JSI), starts 1. 2. 2024
- SMASH (MCSA COFUND)
  - On-boarding first Postdocs
- EUmaster4HPC:
  - Preparing offers for internships

Non-funded projects/activities supported:

- EuroCC SLING
  - Promotional activities, focused on Industry users
- MaX3, MultiXscale CoEs:
  - Configuration of Environment
  - Using National and EuroHPC JU share
- Others:
  - European Digital Infrastructure Consortium (EDIC) – national resources reserved
  - High-level App. Support help for Leonardo
  - CASTIEL2, Container Forum

User Documentation and Support



https://doc.vega.izum.si/

Technical Support

- **First-level support** (info, access, login, etc.) is provided by SLING. Users always send emails only to support@sling.si.
- **Second-level support** is intended for HPC Vega specifics (e.g., missing software on the cluster, problems with compiling, scalability). First-level support sends tickets to Vega system administrators and development engineers.
- **Third-level support** is intended for science disciplines in specific research domains (e.g., issues with provided results). It is organized with domain experts within the National Competence Center or Slovenian EuroCC.

EUROHPC JU SUPERCOMPUTERS

State-of-Play and Upcoming Systems

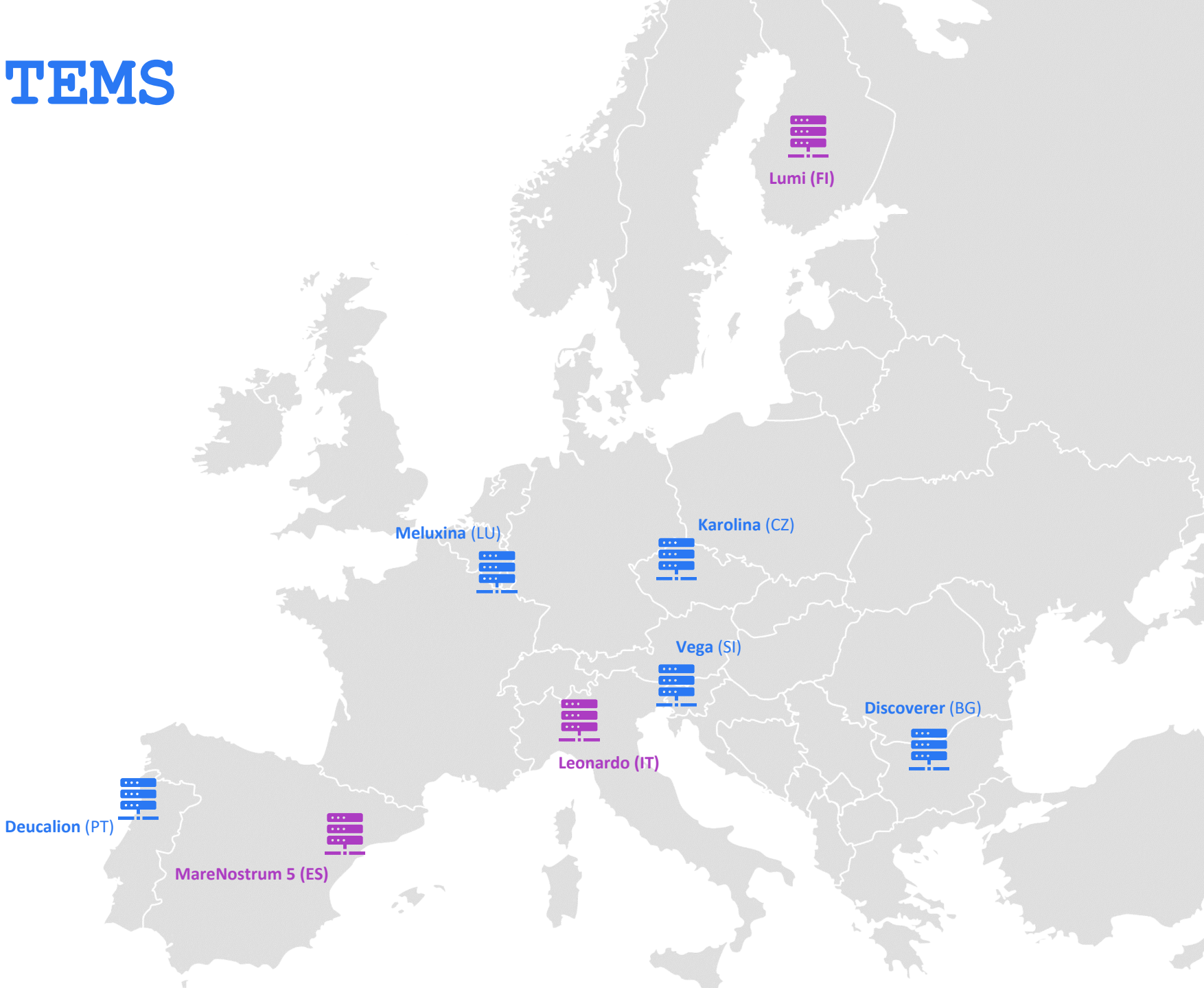
Vangelis Floros | 11 Dec 2023 | EuroHPC User Day

EuroHPC
Joint Undertaking

EUROHPC SYSTEMS
2019 → 2023

PRE-EXASCALE

PETASCALE



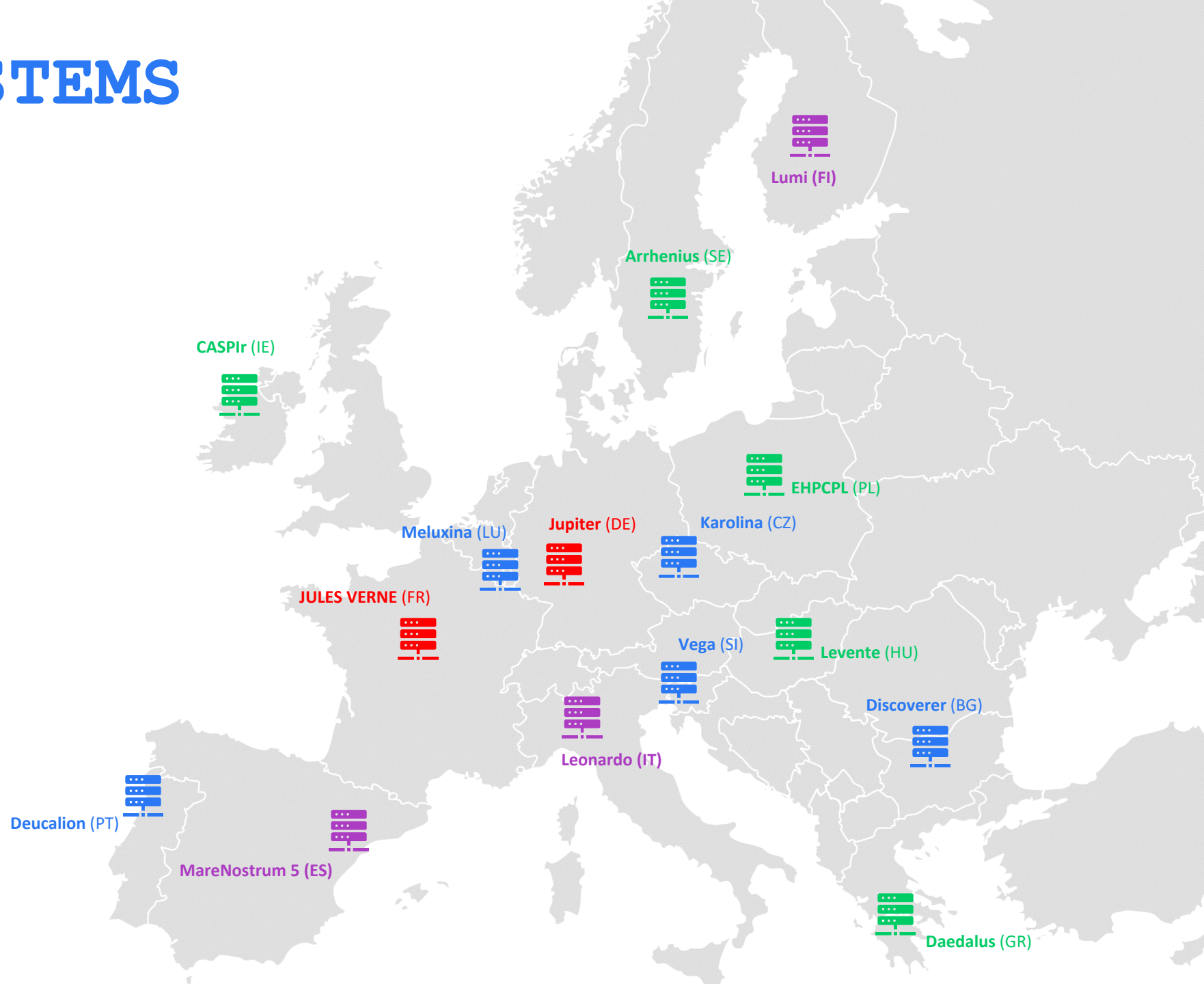
EUROHPC SYSTEMS
2024 → 2026

EXASCALE

PRE-EXASCALE

PETASCALE

MID-RANGE



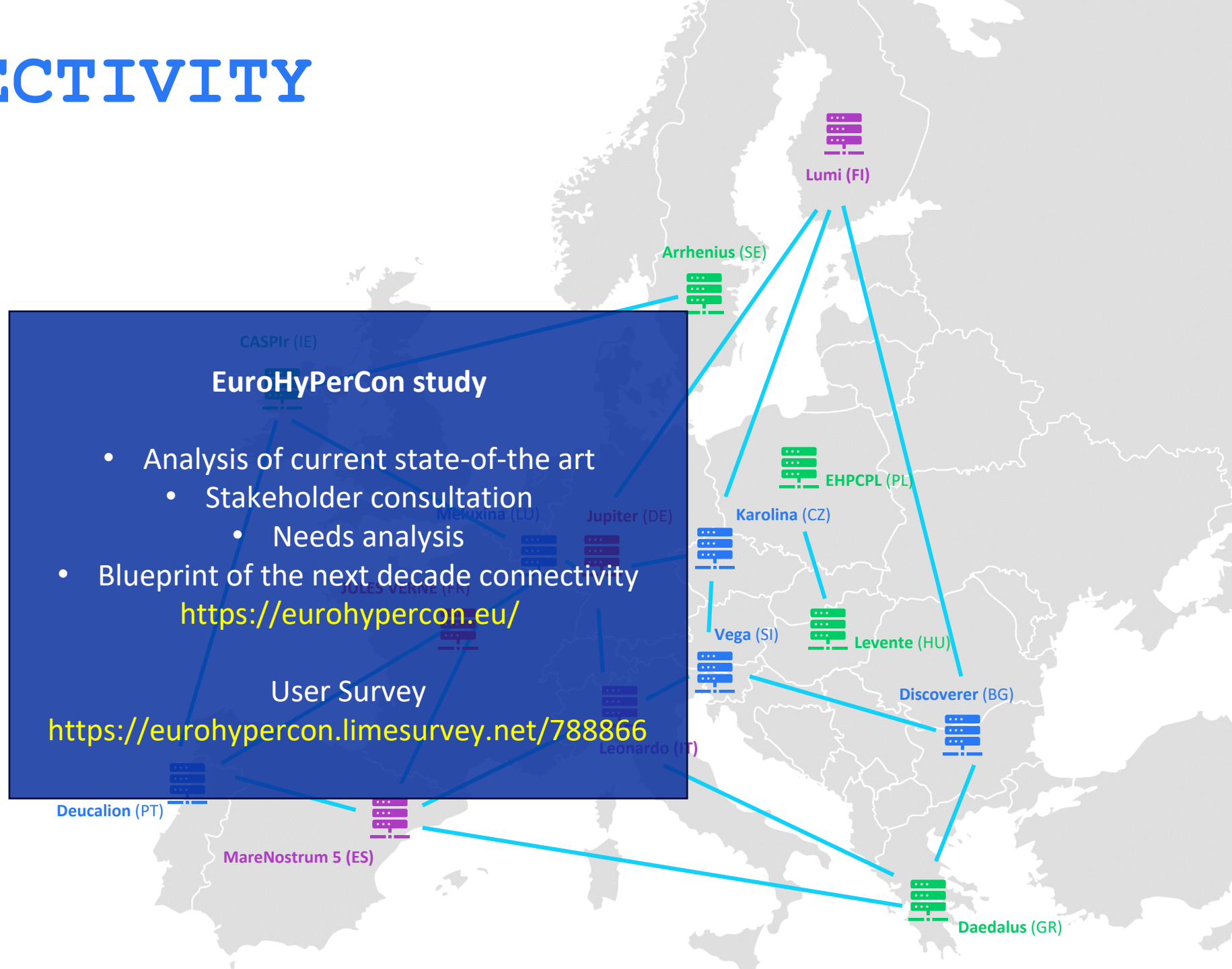
HYPERCONNECTIVITY 2025

EXASCALE

PRE-EXASCALE

PETASCALE

MID-RANGE



EuroHyPerCon study

- Analysis of current state-of-the art
- Stakeholder consultation
- Needs analysis
- Blueprint of the next decade connectivity

https://eurohypercon.eu/

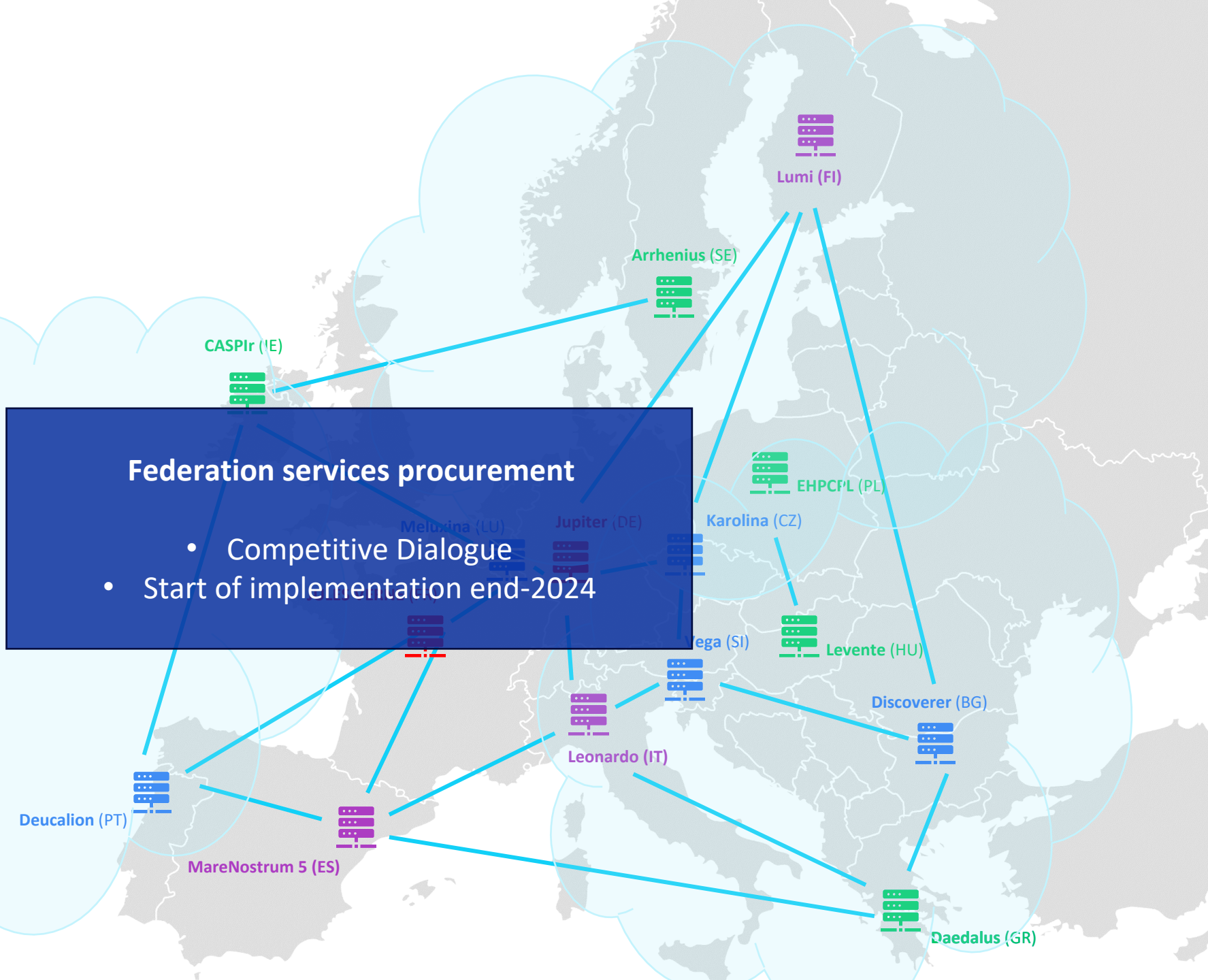
User Survey

https://eurohypercon.limesurvey.net/788866

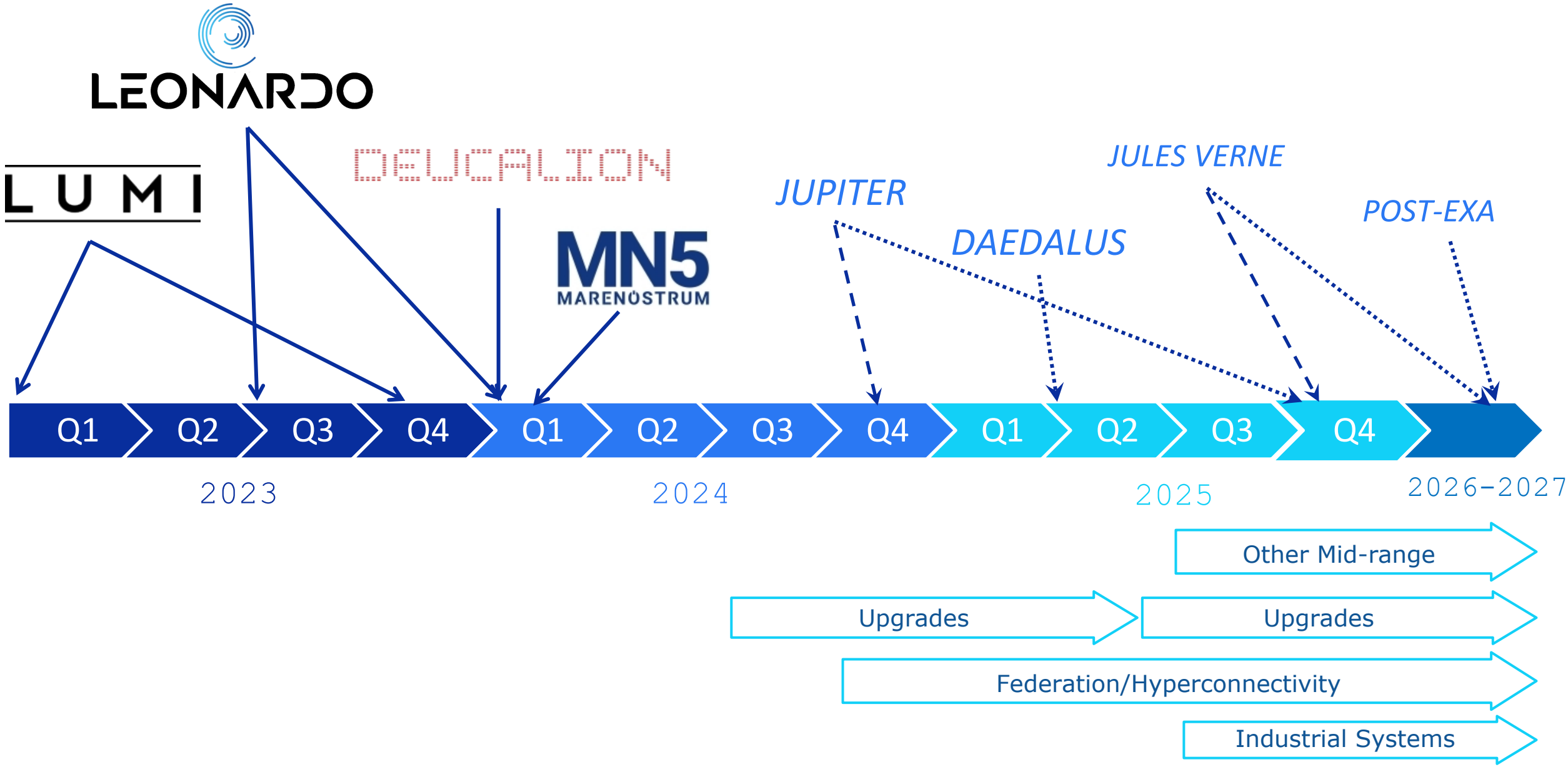
FEDERATION 2024+

Federate HPC resources across all EuroHPC systems

- **Authentication, Authorization and Identification services (AAI)**
- **Computing services**
  - Interactive Computing
  - Cloud access – Virtual Machines - Containers
- **Data services**
  - Archival Services and Data repositories
  - Data mover / transport services
- **User and Resource management**



TIMELINE OF UPCOMING SYSTEMS



THANK YOU

EuroHPC Joint Undertaking

For more information, feel free to visit our website and social media:

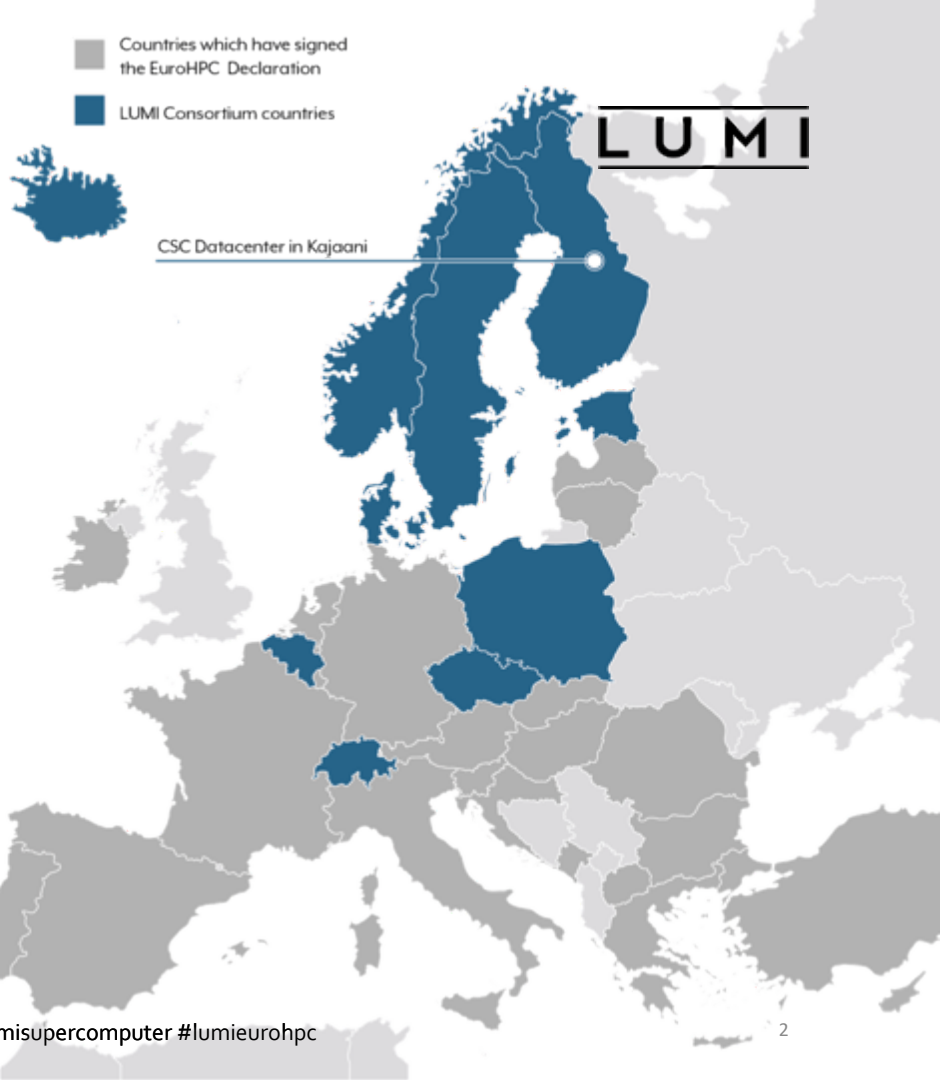
eurohpc-ju.europa.eu @euroHPC_JU eurohpc-ju @eurohpc-ju

LUMI – EuroHPC User Day

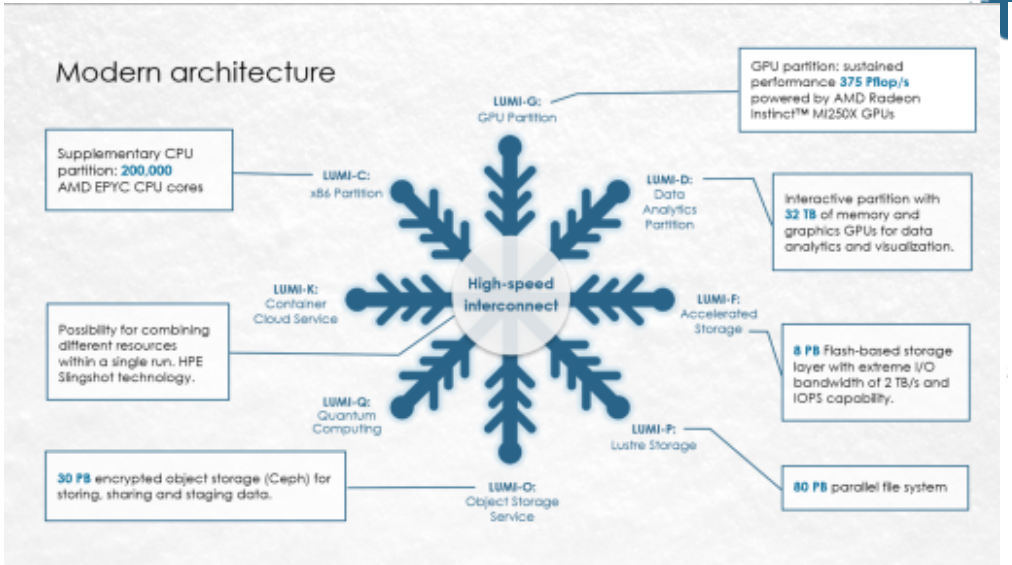
Brussels, Belgium - 11 December 2023

LUMI – Introduction

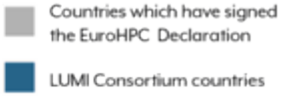
- LUMI is one of the three European pre-exascale supercomputers.
- HPE Cray EX supercomputer consisting of several hardware partitions, all connected via a HPE Slingshot 11 high-speed interconnect.
- As of 11/2023, LUMI ranks fifth on the Top500 list and is currently the fastest supercomputer in Europe. The aggregated HPL Linpack performance of LUMI-G is 379.70 PFlop/s.
- More information: https://www.lumi-supercomputer.eu/



LUMI – Hardware



- **Job scheduling system**: Slurm based on a fair share policy configuration with a priority rate based on job age.



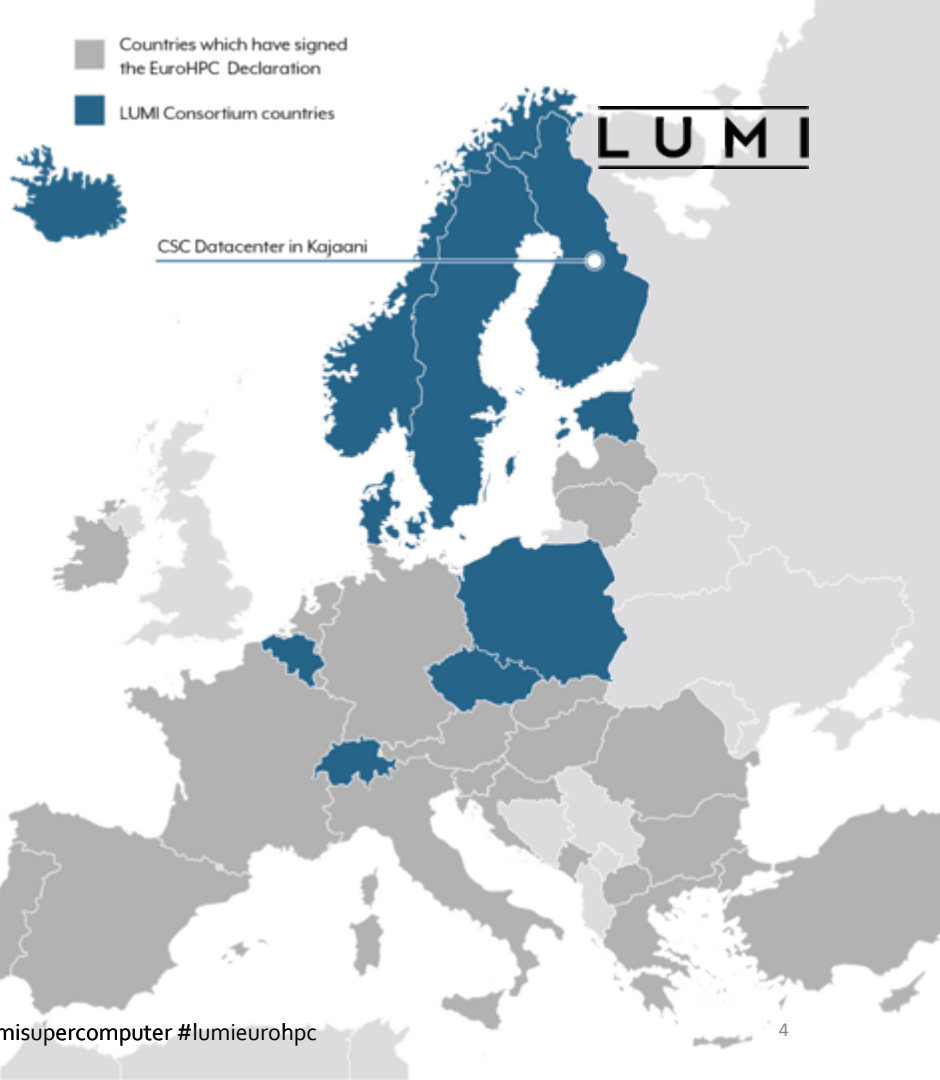
LUMI

- The primary compute power is found in the LUMI-G hardware partition which consists of 2978 nodes with 4 AMD MI250x GPUs and a single 64 cores AMD EPYC "Trento" CPU.
- The LUMI-C hardware partition consists of 2048 CPU based compute nodes, each of them equipped with two AMD EPYC 7763 CPUs with 64 cores each running at 2.45 GHz for a total of 128 cores per node. The cores have support for 2-way simultaneous multithreading (SMT) allowing for up to 256 threads per node.
- The LUMI-D partition consists of 16 nodes with large memory capacity, of which 8 nodes also feature Nvidia visualization GPUs. LUMI-D is intended for interactive data analytics and visualization.
- Storage: LUMI-P (80PB Lustre parallel filesystem), LUMI-F (9PB accelerated flash-based storage layer) and LUMI-O (30PB encrypted object storage)
- More information: https://docs.lumi-supercomputer.eu/hardware/

LUMI – Energy

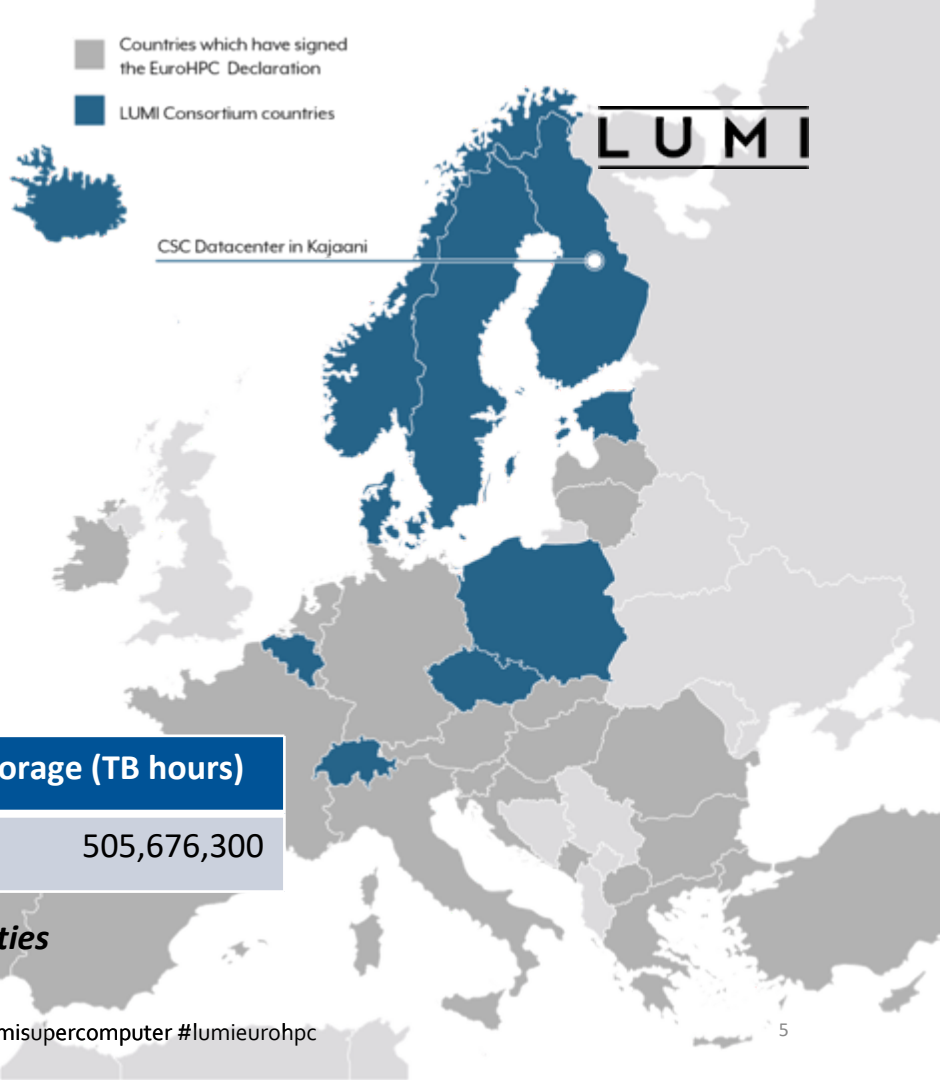
100% HYDROELECTRIC ENERGY

LUMI is using 100% hydro-powered energy. LUMI's waste heat is used to heat hundreds of households in the city of Kajaani.



A European collaboration

- A joint endeavour between EuroHPC JU and 11 consortium members: Belgium, Czech Republic, Denmark, Estonia, Finland, Iceland, Norway, Poland, Sweden, Switzerland and The Netherlands.
- The resources of LUMI are allocated per the investments. The share of the EuroHPC JU (50%) is allocated by a peer-review process:

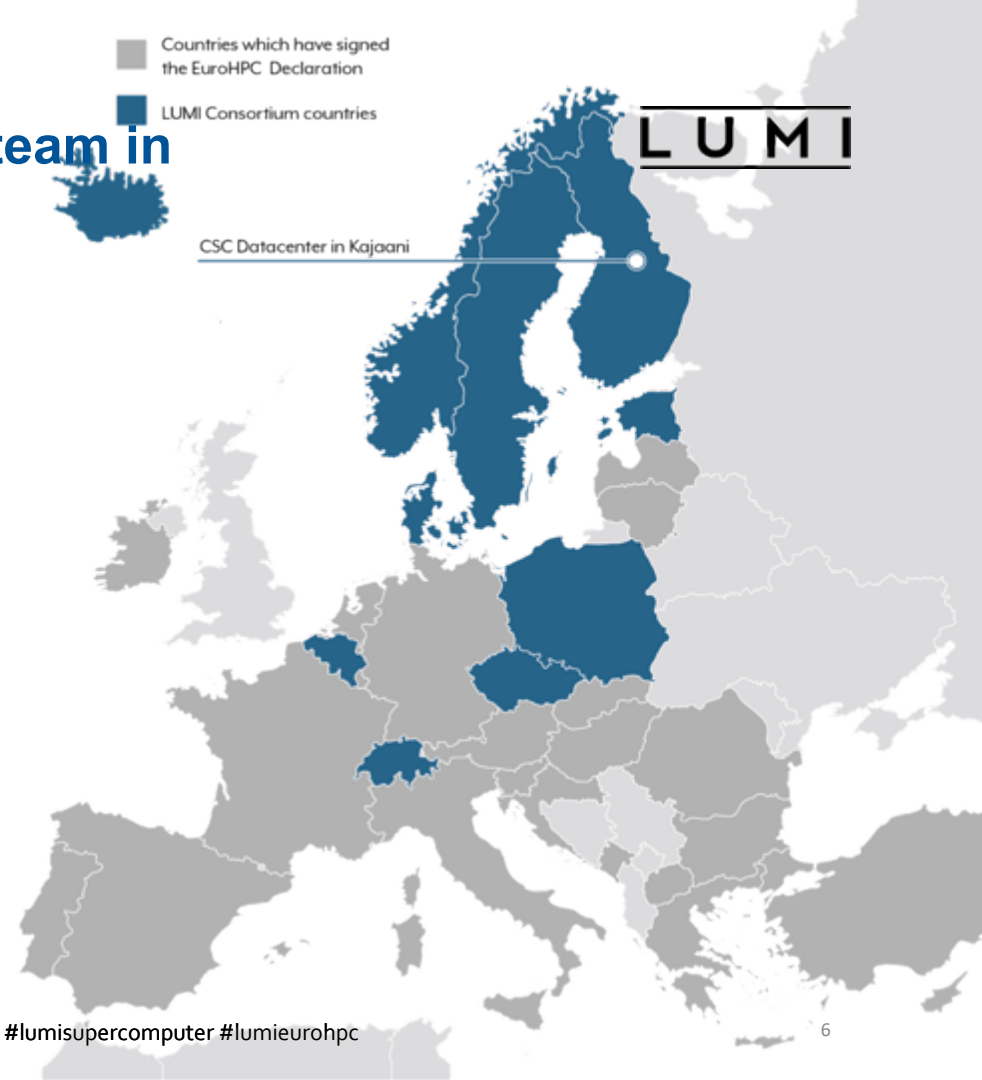


CPU-core hours	GPU-node hours	Storage (TB hours)
1,460,417,900	95,720,200	505,676,300

2024 EuroHPC JU allocation capacities

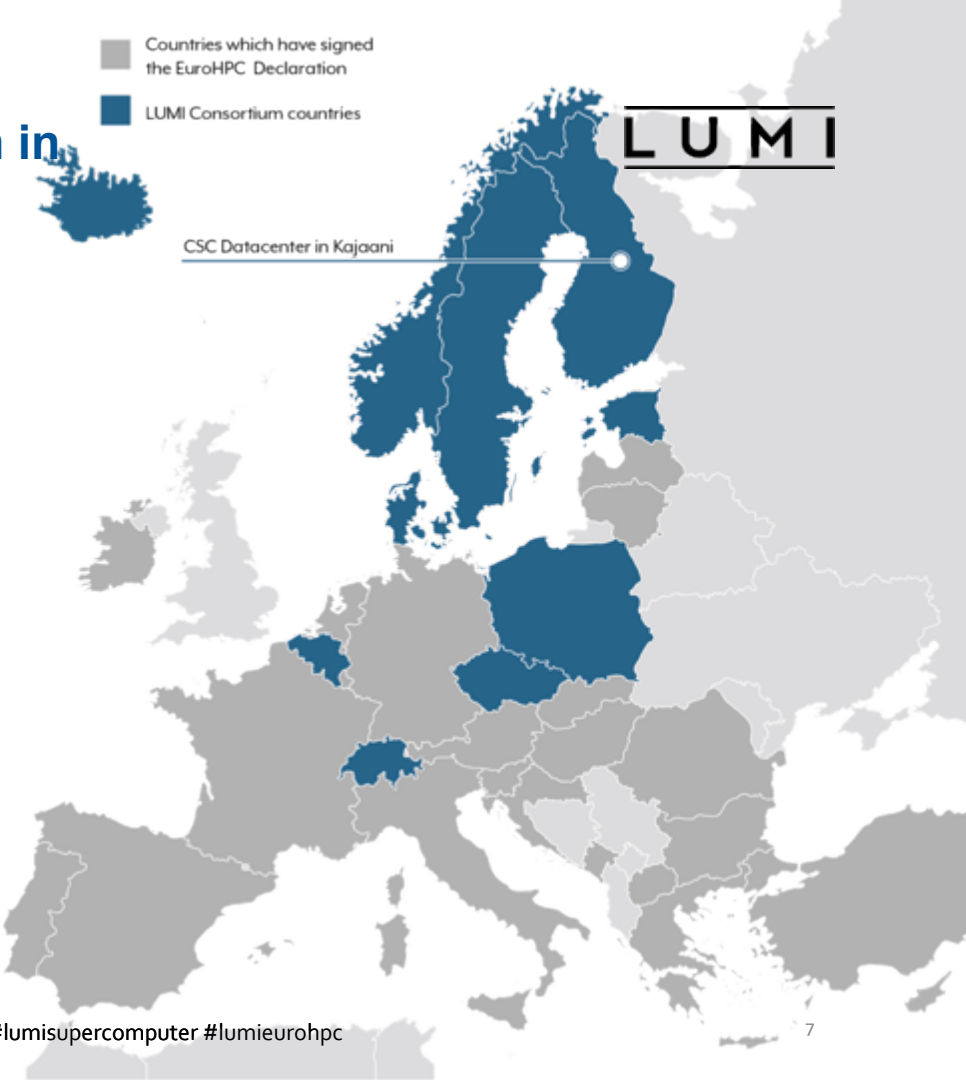
LUST – a unique user support team in high-performance computing

- Centralized virtual help desk run by the distributed LUMI User Support Team
- The model is based on a network of dedicated HPC experts
- Each partner provides 1 FTE/year
- Level-3 support via local centers, EuroHPC Competence Centers, HPE and AMD
- National support for issues with accounts and allocations



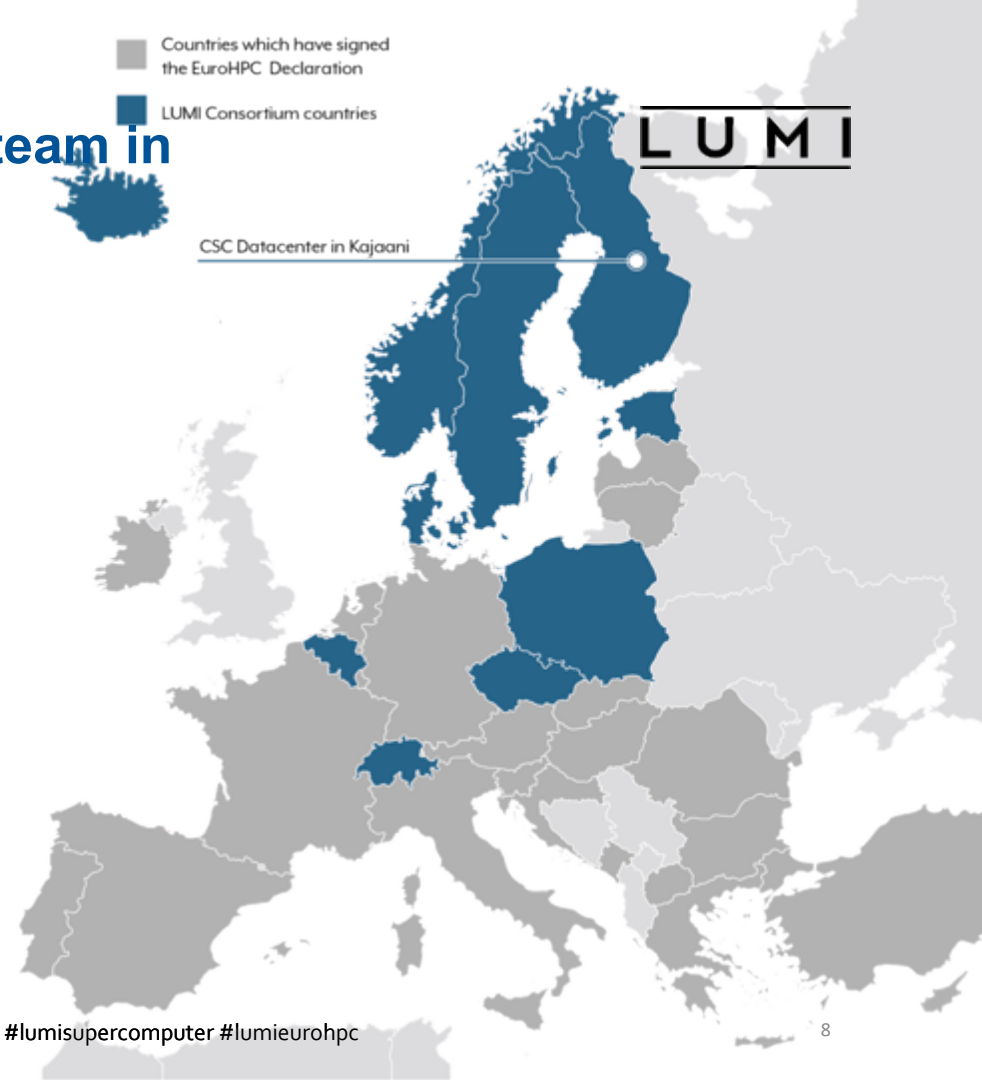
LUST – a unique user support team in high-performance computing

- LUST main activities
  - Help desk from Monday to Friday 8-18 CET (9-19 EET)
  - User documentation
  - Computing environment
  - User training
  - Hackathons
  - Benchmarking, porting, optimization consultancy
  - Full machine runs



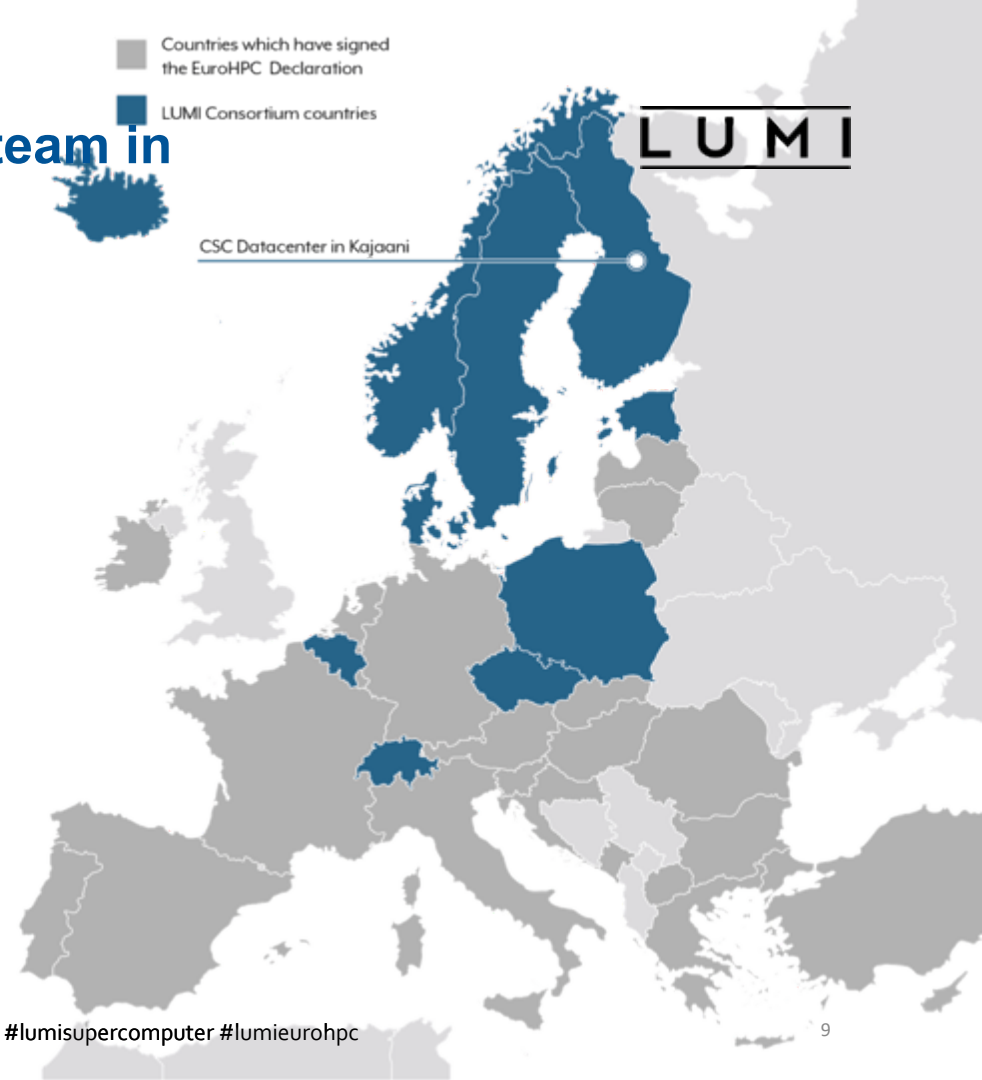
LUST – a unique user support team in high-performance computing

- Computing environment
  - The LUMI software stacks contain the software that are already installed on LUMI.
    - **CrayEnv** offers the Cray Programming Environment (PE)
    - **LUMI** is an extensible software stack that is essentially managed through EasyBuild.
    - Container repository
    - **Local software stack** for local organizations to manage their own software installation



LUST – a unique user support team in high-performance computing

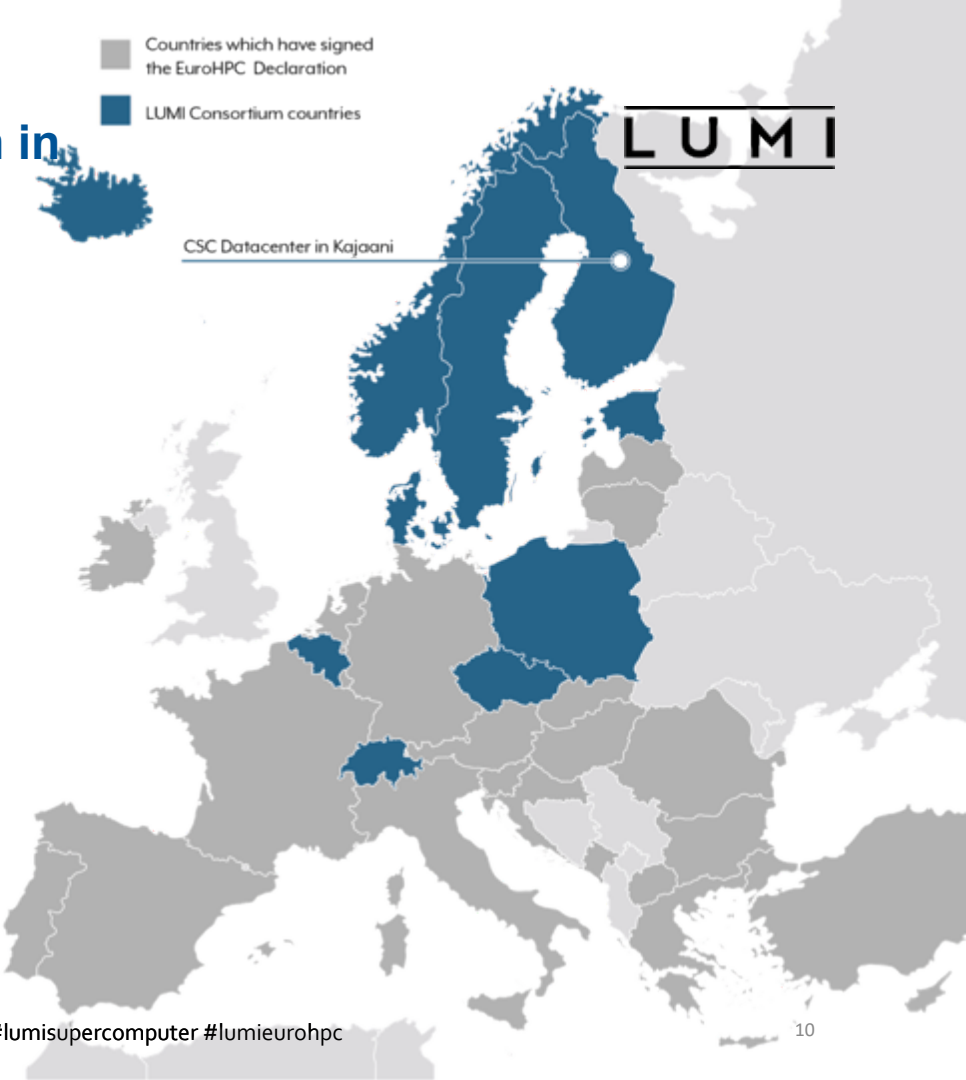
- User training
  - In collaboration with HPE and AMD
  - LUMI-C environment and architecture
  - LUMI-G hardware and programming environment
  - Hackathon
  - More specific trainings: EasyBuild, ML frameworks, profiling…



LUST – a unique user support team in high-performance computing

- Porting & optimization

  - CoE (HPE & AMD) + LUST effort
  - 6 projects in 2022 (e.g. QuantumEspresso, Megatron-LM, SLIM, tmLQCD & PLEGMA, GPAW) and 5 projects in 2023 (TurboGAP, Vlasiator, Genesis, HMSC & SIESTA)
    - Interviews
    - Consultancy
    - Testing
    - Optimization



LUMI

Follow us

Twitter: @LUMIhpc

LinkedIn: LUMI supercomputer

YouTube: LUMI supercomputer

www.lumi-supercomputer.eu

contact@lumi-supercomputer.eu

www.lumi-supercomputer.eu #lumisupercomputer #lumieurohpc

The acquisition and operation of the EuroHPC supercomputer is funded jointly by the EuroHPC Joint Undertaking, through the European Union's Connecting Europe Facility and the Horizon 2020 research and innovation programme, as well as the Participating States FI, BE, CH, CZ, DK, EE, IS, NO, PL, SE.

Leonardo in production

EuroHPC User Day Brussels 11.12.2023

Andrew Emerson, HPC Dept, CINECA

LEONARDO CINECA

Leonardo *https://leonardo-supercomputer.cineca.eu*

- Jointly funded by Italy and EuroHPC
- Located at the Bologna Tecnopole (“Science Park”)
- Many innovations including warm water cooling (37 °C -> °47 C)

Built on the site of an ex-tobacco factory designed by Pier-Luigi Nervi

Remains of roman road found when preparing for Leonardo. Tombs, wells and water channels were also discovered.

TOP500

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,730,112	1,102.00	1,685.65	21,100
2	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899
3	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	2,220,288	309.10	428.70	6,016
4	**Leonardo** - BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 SXM4 64 GB, Quad-rail NVIDIA HDR100 Infiniband, EVIDEN EuroHPC/CINECA Italy	1,463,616	174.70	255.75	5,610

November 2022

5	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	2,752,704	379.70	531.51	7,107
6	**Leonardo** - BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 SXM4 64 GB, Quad-rail NVIDIA HDR100 Infiniband, EVIDEN EuroHPC/CINECA Italy	1,824,768	238.70	304.47	7,404
7	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148.60	200.79	10,096
8	**MareNostrum 5 ACC** - BullSequana XH3000, Xeon Platinum 8460Y+ 40C 2.3GHz, NVIDIA H100 64GB, Infiniband NDR200, EVIDEN EuroHPC/BSC Spain	680,960	138.20	265.57	2,560

November 2023

dropped only two places in 1 year

Numbers

LEONARDO'S NUMBERS

155 SYSTEM RACKS	**4992** COMPUTING NODES	**250** PETAFLOPS	**2800** TB OF RAM
6 MW IN OPERATIONS	**110** PB OF STORAGE	**600** M^2 FOOTPRINT	**>95%** HEAT DISSIPATION VIA DLC

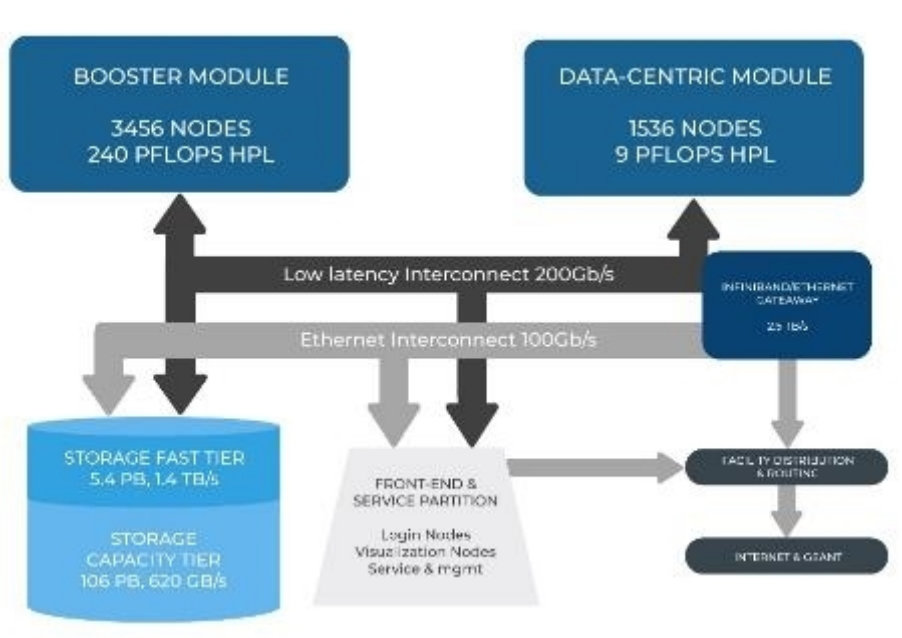
BOOSTER

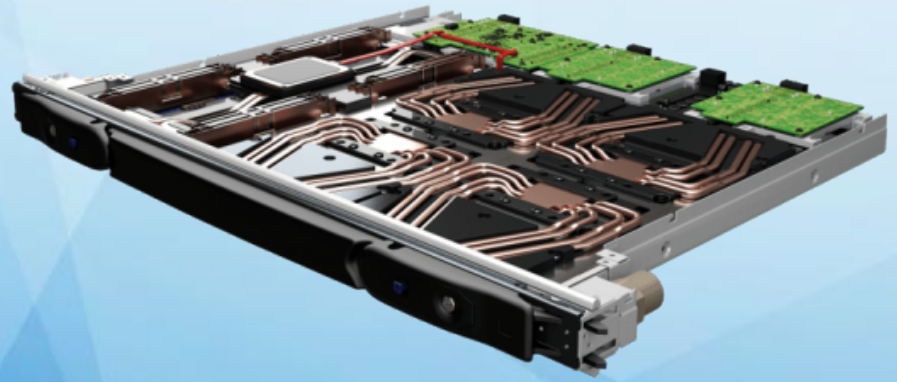
PRE-PRODUCTION : April 2023

Full PRODUCTION FROM: MAY 1st 2023

LEONARDO CINECA

In detail





Booster Module

- **3456** nodes consisting of 4xNVIDIA **A100** GPUs and 1 x32-core Intel Ice Lake CPU
- 512 GB RAM/node
- 89.4 TFLOPs peak perf per node

Data Centric Module

- 1536 nodes with two Intel Sapphire Rapids CPU/node (40 cores*)

Network

- Infiniband network of 200 Gb/s
- Dragonfly topology

Data Storage

- Over 200 PB total storage
- Fast Tier - 24 x 7,68 TB SSD NVMe with encryption support
- Capacity Tier -82 x 18 TB HDD

Data – Centric partition

1500 nodes consisting of BullSequana X2610 compute blades.

Each node contains:

- 2x Intel Sapphire Rapids, 56 cores, TDP 350 W
- 512 (16 x 32) GB RAM DDR5 4800 MHz
- 1 x NVidia HDR100 100 Gb/s card
- 8 TB NVM

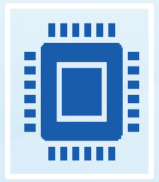
Coming Soon

Next steps

LISA (Leonardo Improved Supercomputing Architecture)

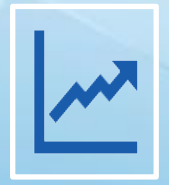
New Capacity partition

- High memory-bandwidth computing nodes



New Accelerated module

- New acceleration module based on next generation GPUs



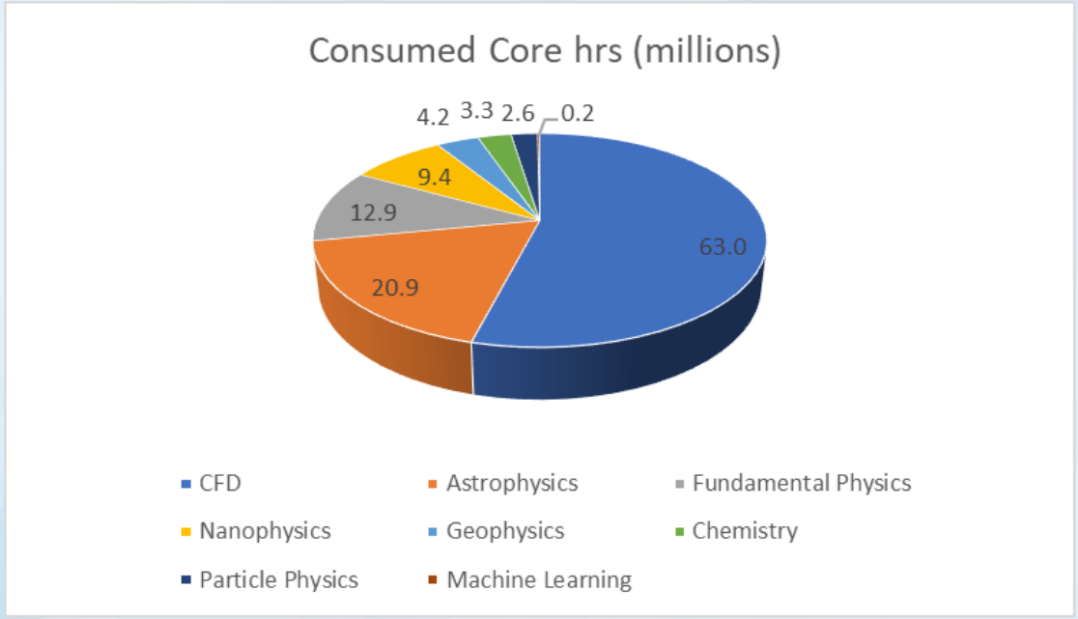
Will result in an increase in performance of 100 Pflops

Pre-production

LEAP Leonardo **E**arly **A**ccess **P**rogram

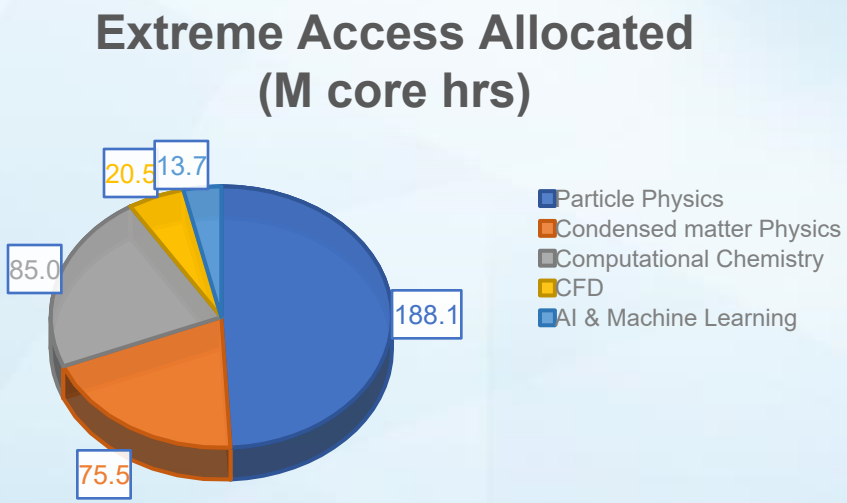
- In agreement with EuroHPC, CINECA opened a **call for early access** to the Booster module of EuroHPC Leonardo (Jan – May 2023)
- **Requirements**: production-ready, highly efficient applications able to exploit fully the hardware, potentially the whole cluster
- **Pre-production** – no guarantee of providing all resources requested, or non-predicted downtimes.

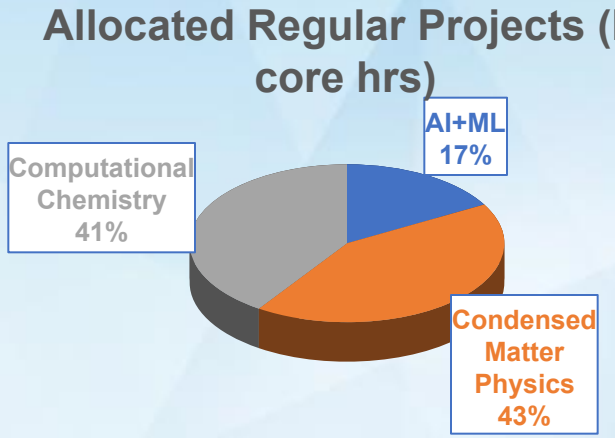
83 projects submitted
20 projects accepted



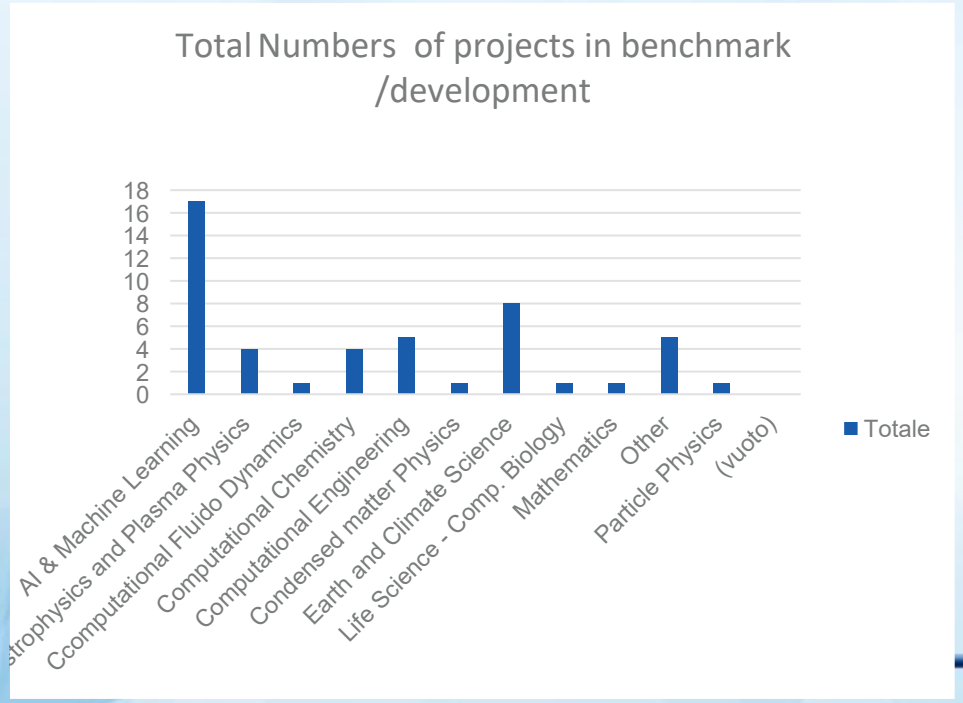
** 32 core hours= 1 node hour = 4 GPU hours*

Production – EuroHPC Resource allocations





Many new requests for AI and ML – often with very high storage and data transfer requests



DESTINATION EARTH
18.05 M core hrs allocated

Example Horizons Europe project : LIGATE

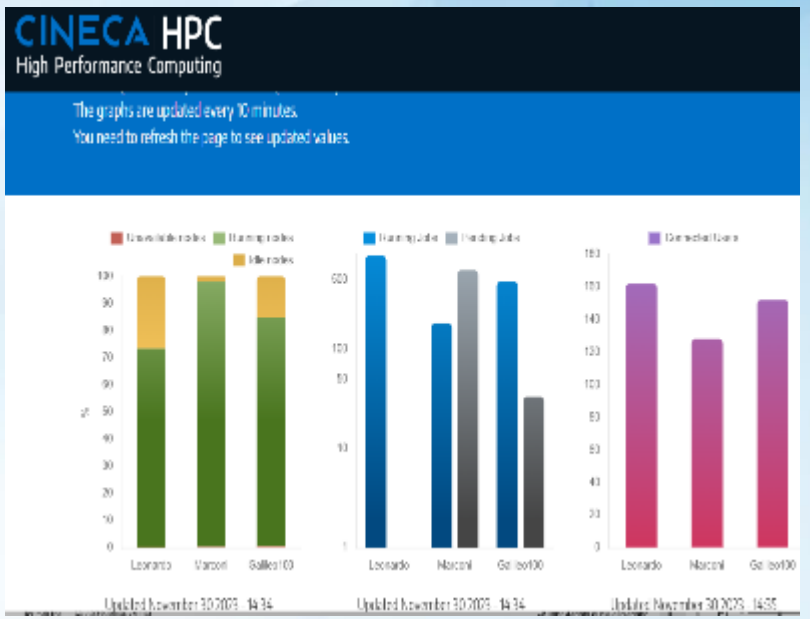
- LIGATE aims for an integrated solution for drug discovery using HPC, virtual screening, workflow software and AI.
- Key difficulty is the ***scoring function*** which identifies which molecules bind to the target.
- On Leonardo we aim to perform **500-1000 Gromacs molecular dynamics simulations / month**, each 1 μs.
- The results will be used to train an accurate ML model for screening new drugs.

LIGATE

LEONARDO CINECA

Leonardo for Users

- LEONARDO is in full production
- Access for users via SSH and 2 factor authentication (2FA is imposed by Italian security regulations)
- User registration completely online
  - One username can access multiple budgets
- Home, Scratch, Work and Project areas allow for different data storage and sharing.
- Currently lacking long term archive storage (DRES) and fast network to CINECA main systems (Leonardo is at a different site)
- Real-time status of HPC systems via web portal (new)
- Use Support via tickets system
- Extensive training opportunities for staff and users
- On-site Nvidia staff for training and support



`superc@cineca.it`

for support requests

THANK YOU!

EuroHPC
Joint Undertaking

“The acquisition and operation of the EuroHPC supercomputer is funded jointly by the EuroHPC Joint Undertaking under grant agreement No cnect.ddg1.c.2(2019)8804531 - LEONARDO supercomputer, through the European Union’s Connecting Europe Facility and the Horizon 2020 research and innovation programme, as well as the Italian Ministry of University and Research.

This project has received funding from the European High Performance Computing Joint Undertaking (JU) under grant agreement No cnect.ddg1.c.2(2019)8804531 - LEONARDO supercomputer. The JU receives support from the European Union’s Horizon 2020 research and innovation programme.

LEONARDO
CINECA

Generalitat de Catalunya
Departament de Recerca
i Universitats

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN

UPC
UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

UNIÓN EUROPEA
Fondo Europeo de Desarrollo Regional

MareNostrum5

The acquisition and operation of the EuroHPC supercomputer is funded jointly by the EuroHPC Joint Undertaking, through the European Union's Connecting Europe Facility and the Horizon 2020 research and innovation programme, as well as the Participating States Spain, Portugal and Türkiye

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

MareNostrum5

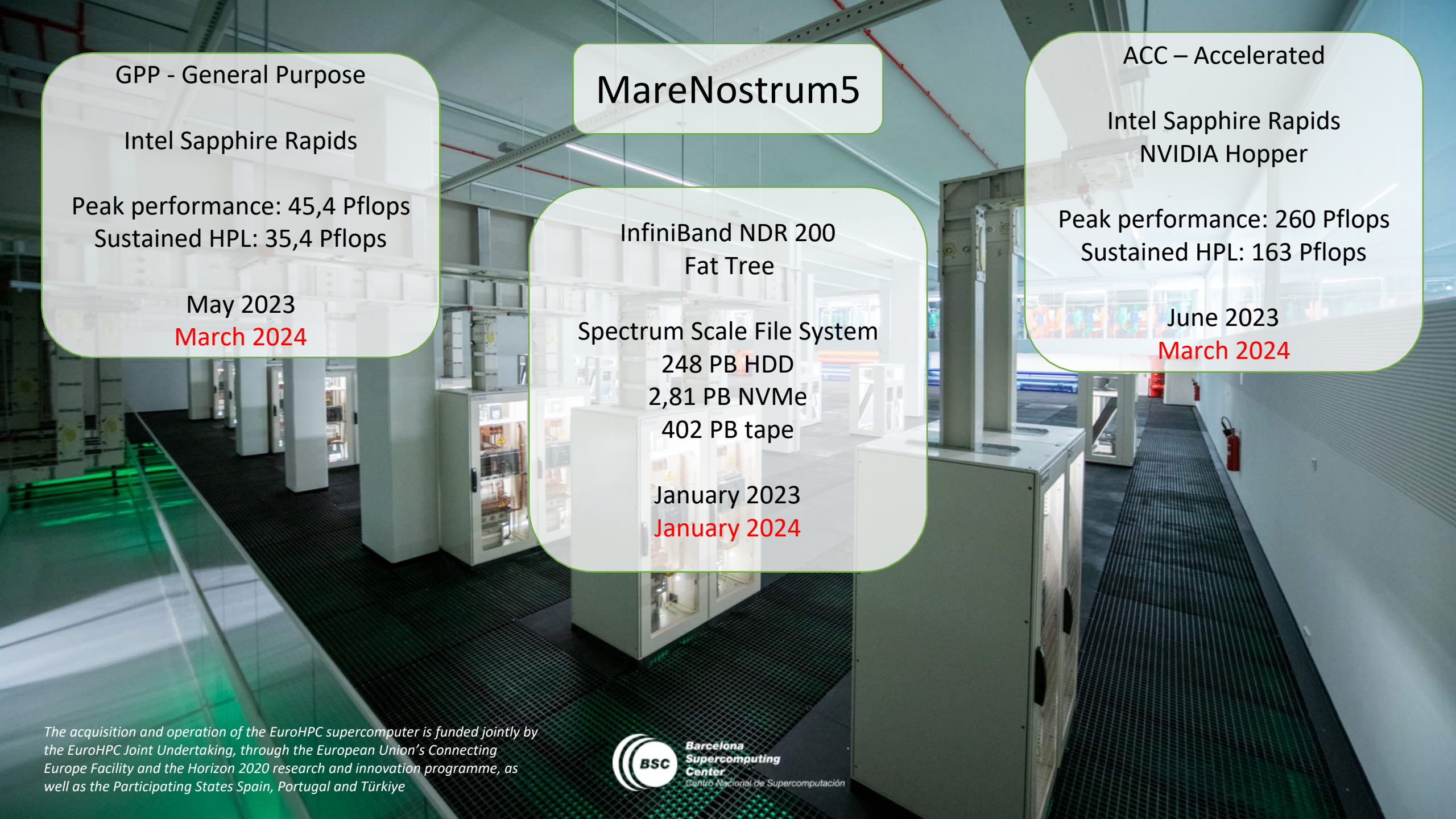
InfiniBand NDR 200
Fat Tree

Spectrum Scale File System
248 PB HDD
2,81 PB NVMe
402 PB tape

January 2023
January 2024

The acquisition and operation of the EuroHPC supercomputer is funded jointly by the EuroHPC Joint Undertaking, through the European Union's Connecting Europe Facility and the Horizon 2020 research and innovation programme, as well as the Participating States Spain, Portugal and Türkiye

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación



MareNostrum5

GPP - General Purpose

Intel Sapphire Rapids

Peak performance: 45,4 Pflops
Sustained HPL: 35,4 Pflops

May 2023
March 2024

InfiniBand NDR 200
Fat Tree

Spectrum Scale File System
248 PB HDD
2,81 PB NVMe
402 PB tape

January 2023
January 2024

ACC – Accelerated

Intel Sapphire Rapids
NVIDIA Hopper

Peak performance: 260 Pflops
Sustained HPL: 163 Pflops

June 2023
March 2024

The acquisition and operation of the EuroHPC supercomputer is funded jointly by the EuroHPC Joint Undertaking, through the European Union's Connecting Europe Facility and the Horizon 2020 research and innovation programme, as well as the Participating States Spain, Portugal and Türkiye

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

MareNostrum5

GPP - General Purpose

Intel Sapphire Rapids

Peak performance: 45,4 Pflops
Sustained HPL: 40,1 Pflops

May 2023
March 2024

NGT GPP - Next Generation

NVIDIA Grace

Peak performance: 2,82 Pflops
Sustained HPL: *2 Pflops**

October 2023
March 2024

InfiniBand NDR 200
Fat Tree

Spectrum Scale File System
248 PB HDD
2,81 PB NVMe
402 PB tape

January 2023
January 2024

ACC – Accelerated

Intel Sapphire Rapids
NVIDIA Hopper

Peak performance: 260 Pflops
Sustained HPL: *163 Pflops**

June 2023
March 2024

NGT ACC - Next Generation

~~CANCELLED~~

Intel Emerald Rapids
Intel Rialto Bridge

Peak performance: 6 Pflops
Sustained HPL: 4,24 Pflops*

MORE Coming Soon

December 2023

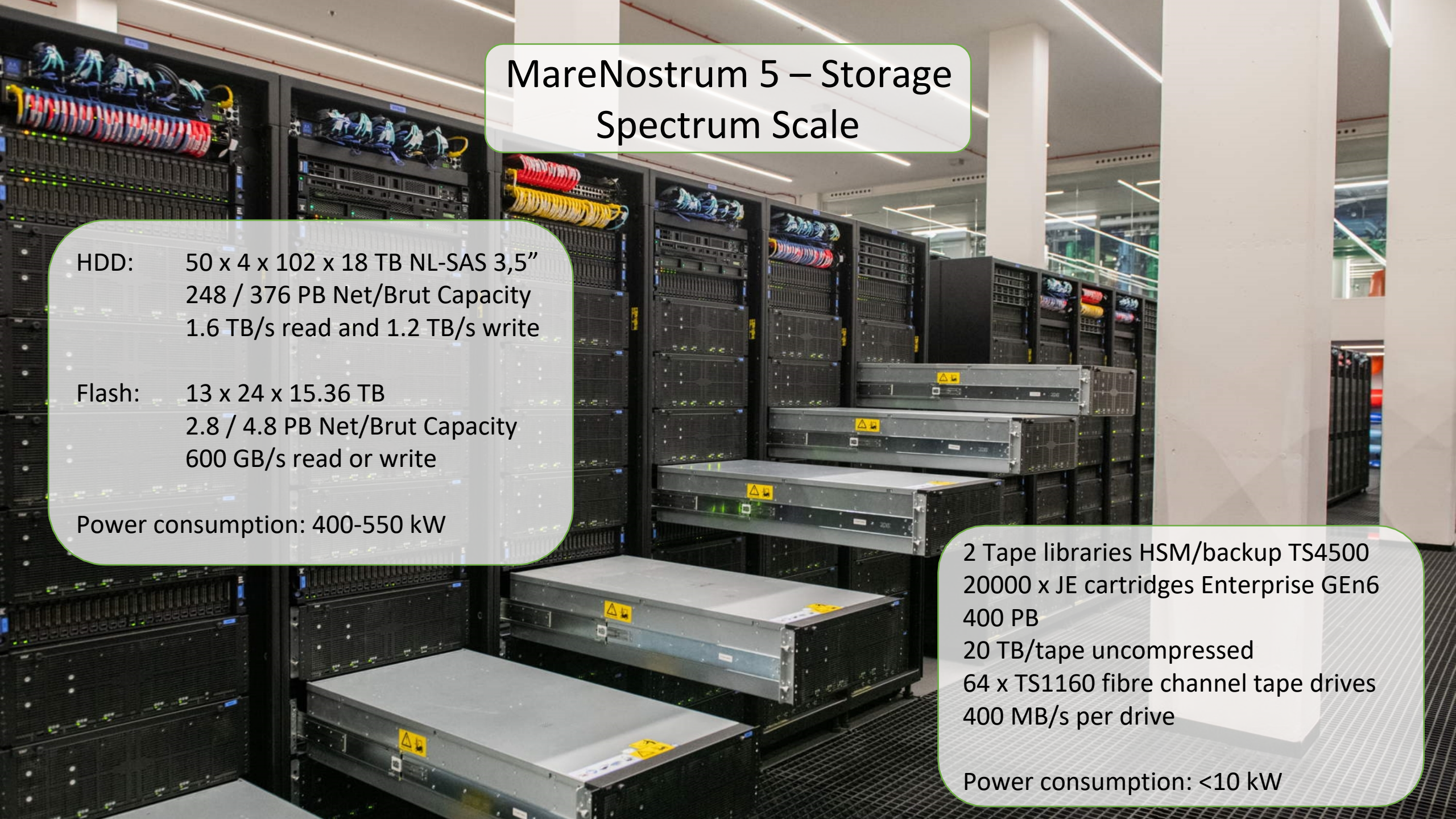
The acquisition and operation of the EuroHPC supercomputer is funded jointly by the EuroHPC Joint Undertaking, through the European Union's Connecting Europe Facility and the Horizon 2020 research and innovation programme, as well as the Participating States Spain, Portugal and Türkiye

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

MareNostrum5 – Software stack

Software type	MN5
Operating system	Red Hat Enterprise Linux
Compiler Suite	Intel OneAPI HPC Toolkit Nvidia SDK (PGI)
Numerical libraries	Intel MKL Nvidia SDK
Debugging/profiler tools	BSC Performance tools ARM DDT Nvidia SDK Intel OneAPI HPC Toolkit (vtune, ...)
Resource and workload manager	SLURM Only one Slurm cluster, with different partitions
Energy Efficiency and Power Management	EAR

MareNostrum 5 – Storage Spectrum Scale



HDD: 50 x 4 x 102 x 18 TB NL-SAS 3,5”
248 / 376 PB Net/Brut Capacity
1.6 TB/s read and 1.2 TB/s write

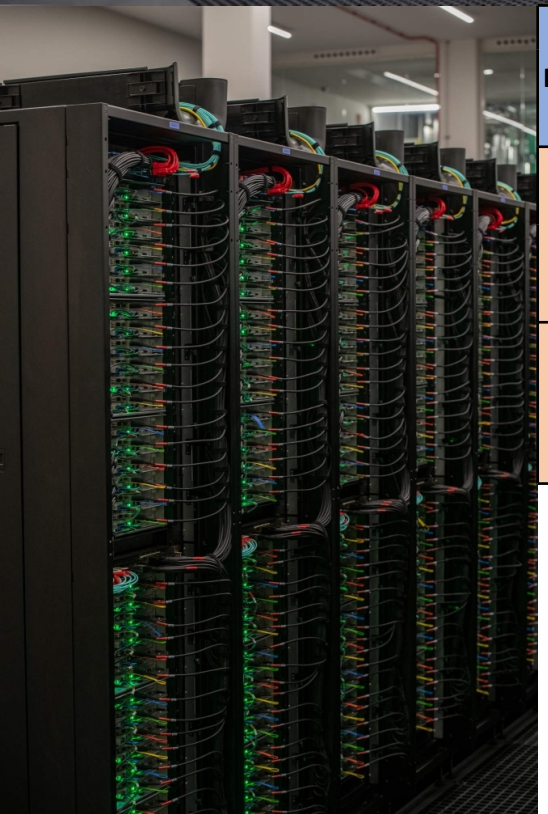
Flash: 13 x 24 x 15.36 TB
2.8 / 4.8 PB Net/Brut Capacity
600 GB/s read or write

Power consumption: 400-550 kW

2 Tape libraries HSM/backup TS4500
20000 x JE cartridges Enterprise GEn6
400 PB
20 TB/tape uncompressed
64 x TS1160 fibre channel tape drives
400 MB/s per drive

Power consumption: <10 kW

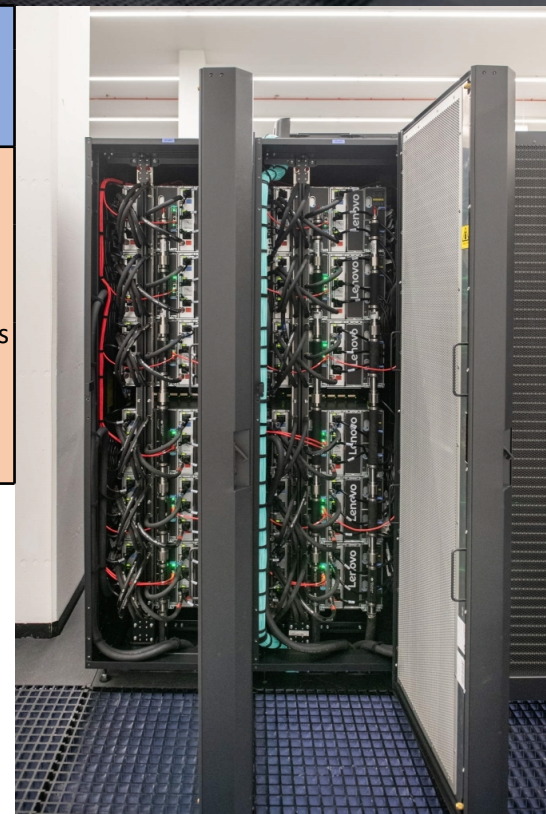
MareNostrum5 GPP



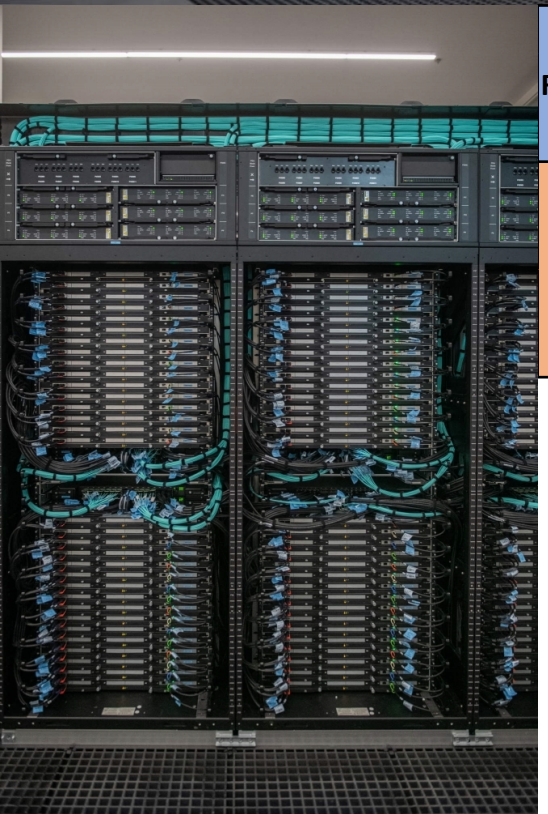
Racks	Cooling	Nodes		Processor/Accelerator		Memory	PFlops (HPL)	Local Drive	High-Perf. Network
		Total	per rack						
89	DLC +RDHX	6192	72 (6x6x2)	2x Intel Sapphire R. 8480+	56c @ 2GHz	>2GB/core 256GB DDR5	40.10	960GB NVMe	1x NDR200 Shared by 2 nodes
		216				>8GB/core 1024GB DDR5			
1		72		2x Intel Sapphire R. 9480	56c @ 1.9GHz	> 0.5GB HBM/core 128GB HBM + 32GB DDR5	0.34		

November 2023

HPL: #19 , #1 x86
HPCG: #24
Green500: #81
5.7 MW under HPL



MareNostrum5 ACC



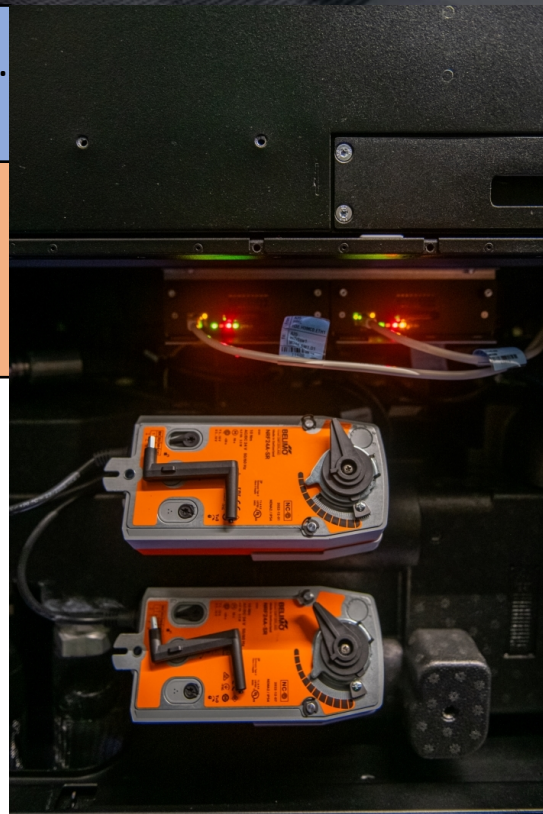
Racks	Cooling	Nodes		Processor/Accelerator		Memory	PFlops (HPL)	Local Drive	High-Perf. Network
		Total	per rack						
35	DLC	1120	32	2x Intel Sapphire R. 8460Y+	40c @ 2GHz	512GB	*163** 138	480GB NVMe	4x NDR200
				4x Nvidia Hopper 64GB HBM					

November 2023

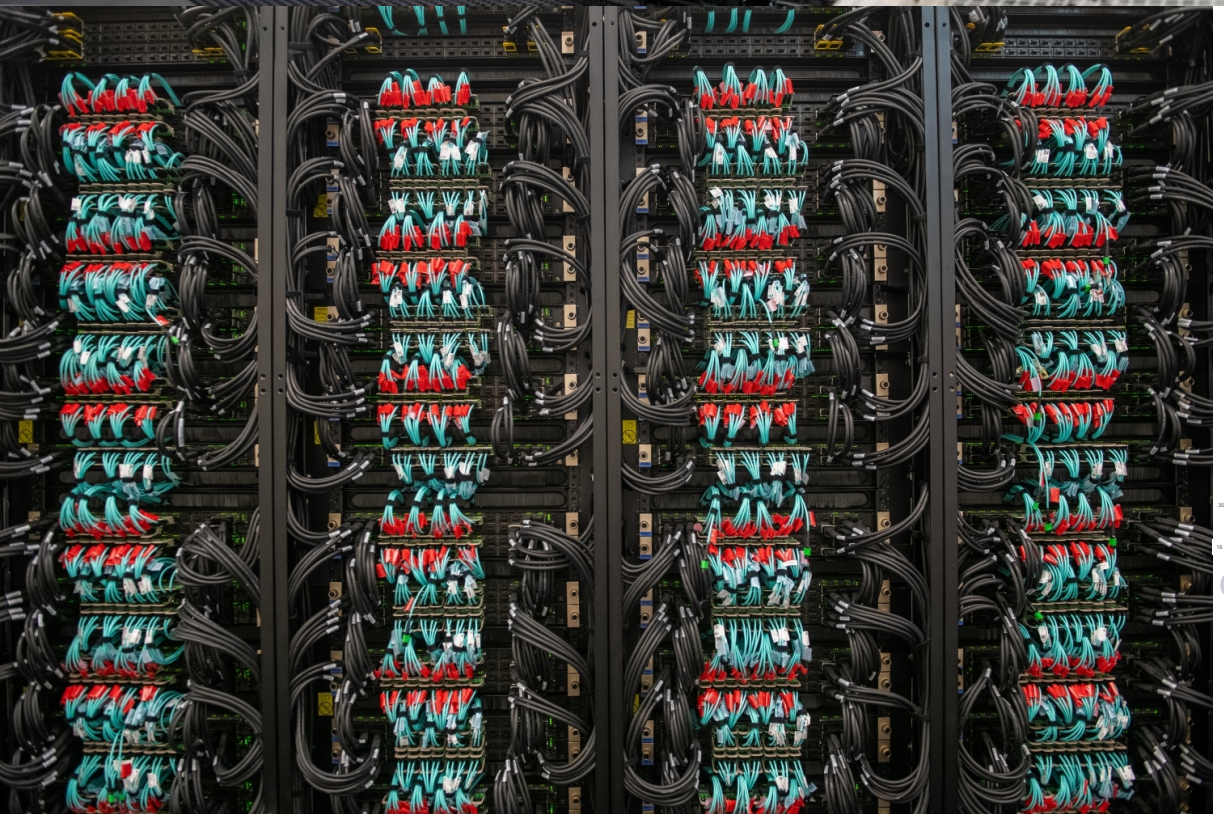
HPL: #8

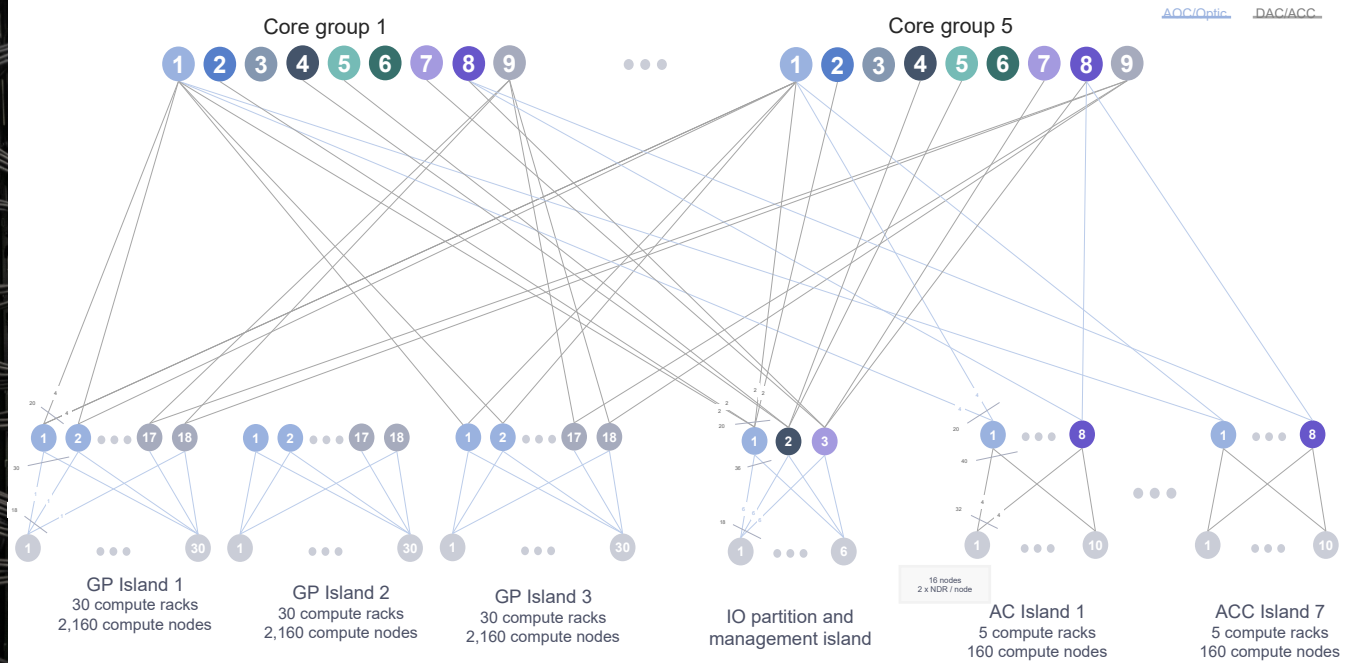
Green500: #5

2.5 MW under HPL



MareNostrum 5 – High Speed network
IB NDR200





Thank you

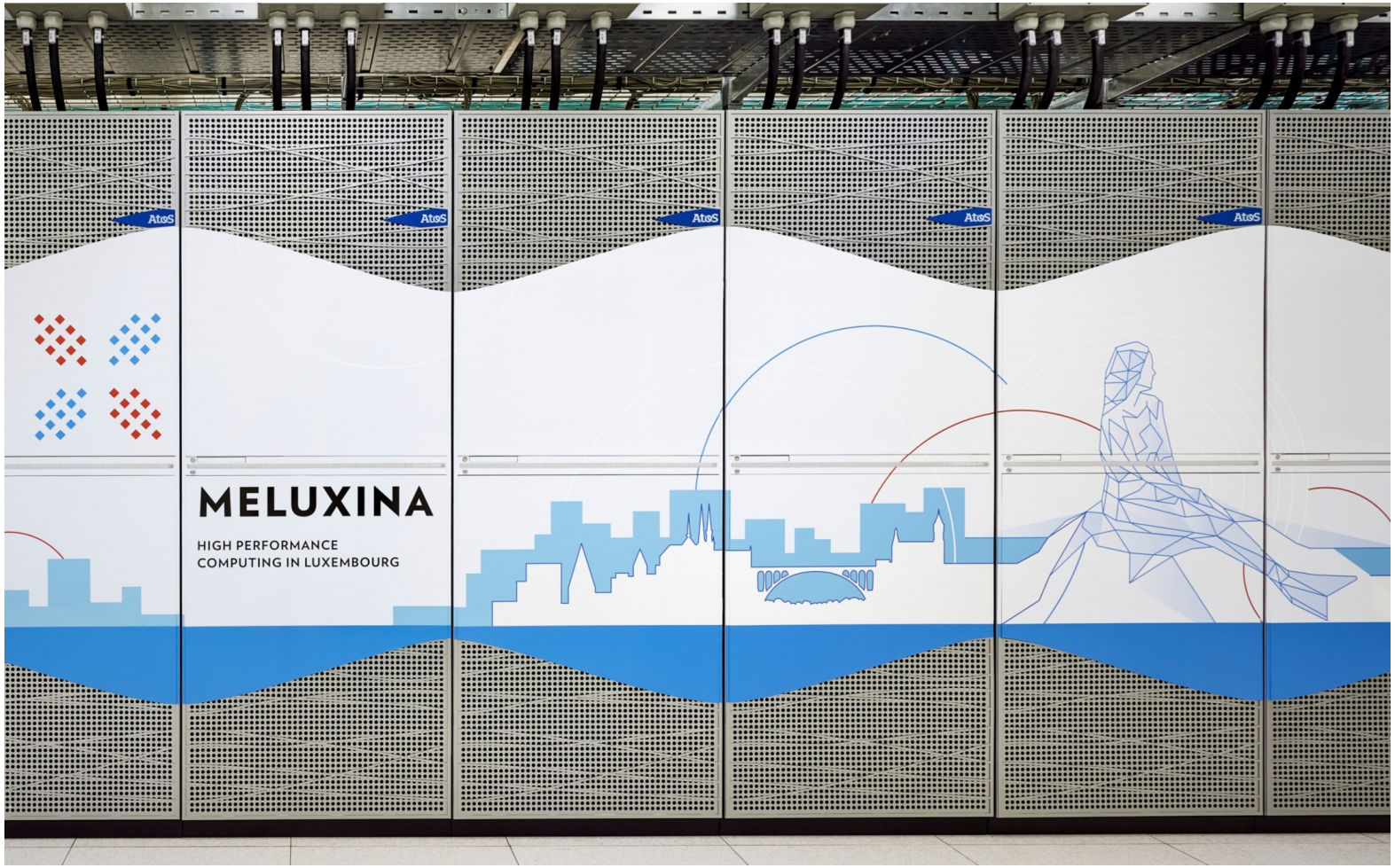
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

sergi.girona@bsc.es

LUX
PROVIDE
NEXT-LEVEL INSIGHTS WITH SUPERCOMPUTING

MeluXina

MELUXINA IN NUMBERS



90.000
HPC CPU cores

800
GPU-AI accelerators

20
Petabytes high performance storage

+450
TB RAM

300+
Tailored software packages

When it was launched in June 2021, the MeluXina accelerator module has been ranked:

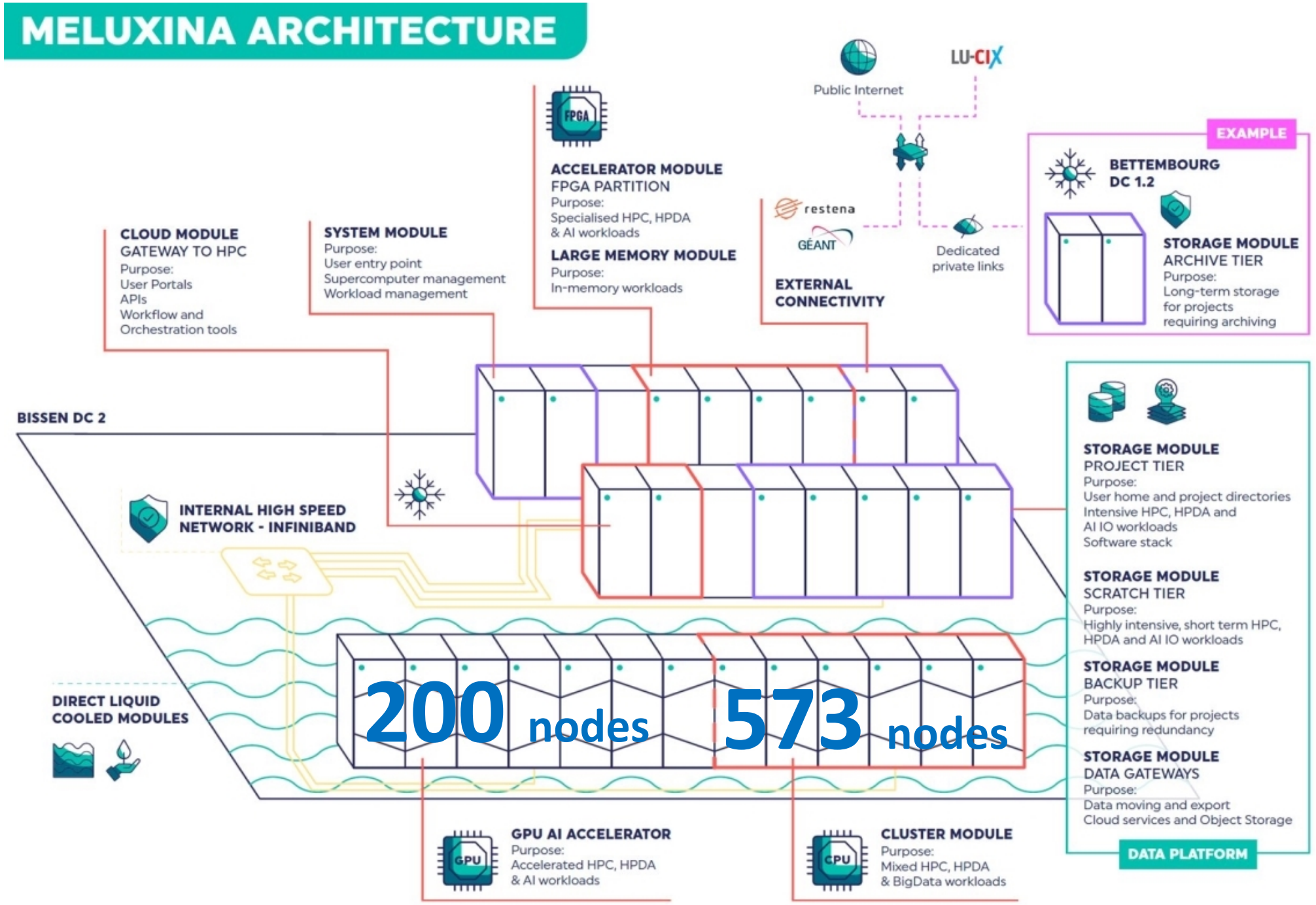
TOP500 in 2021

36th

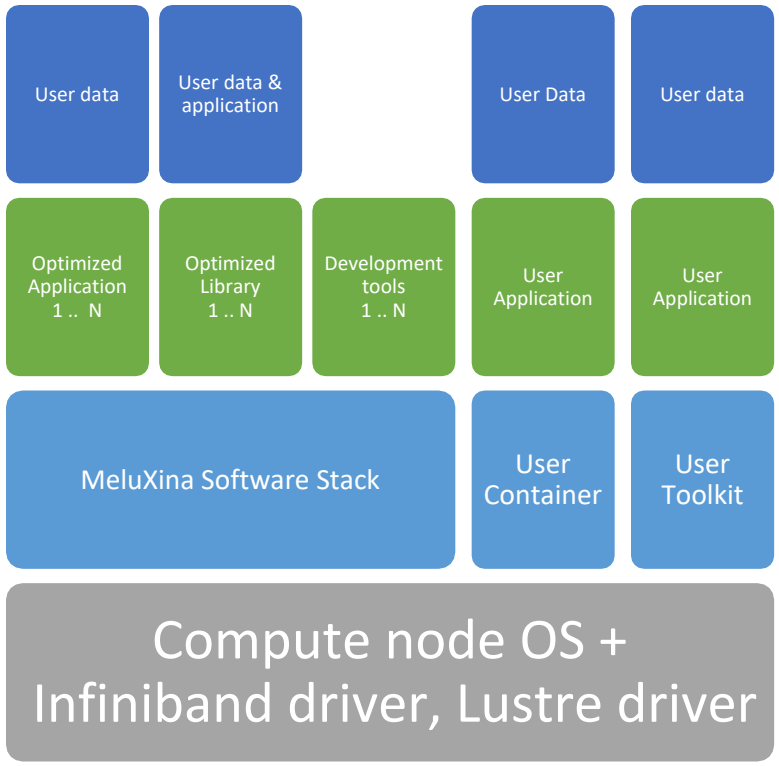
GREEN500 in 2021

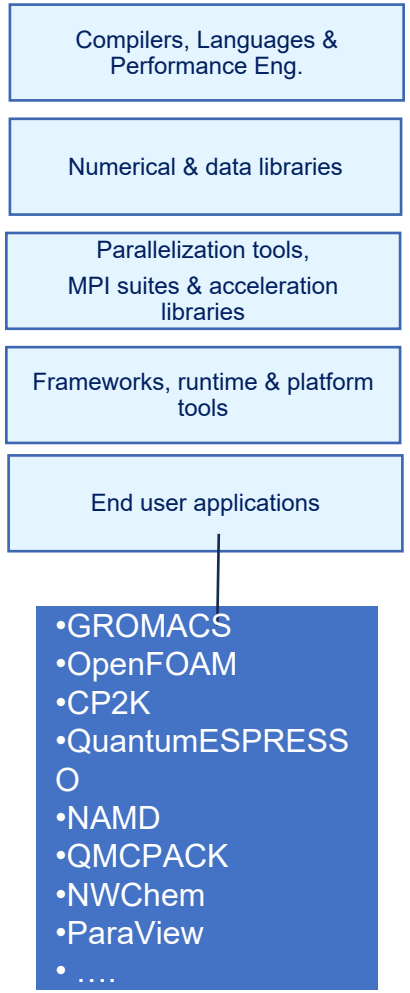
4th (world)

1st (EU)


MELUXINA ARCHITECTURE
Public Internet
LU-CIX
FPGA
ACCELERATOR MODULE
FPGA PARTITION
Purpose:
Specialised HPC, HPDA
& AI workloads
LARGE MEMORY MODULE
Purpose:
In-memory workloads
restena
GÉANT
EXTERNAL
CONNECTIVITY
Dedicated
private links
EXAMPLE
BETTEMBOURG
DC 1.2
STORAGE MODULE
ARCHIVE TIER
Purpose:
Long-term storage
for projects
requiring archiving
CLOUD MODULE
GATEWAY TO HPC
Purpose:
User Portals
APIs
Workflow and
Orchestration tools
SYSTEM MODULE
Purpose:
User entry point
Supercomputer management
Workload management
BISSEN DC 2
INTERNAL HIGH SPEED
NETWORK - INFINIBAND
STORAGE MODULE
PROJECT TIER
Purpose:
User home and project directories
Intensive HPC, HPDA and
AI IO workloads
Software stack
STORAGE MODULE
SCRATCH TIER
Purpose:
Highly intensive, short term HPC,
HPDA and AI IO workloads
STORAGE MODULE
BACKUP TIER
Purpose:
Data backups for projects
requiring redundancy
STORAGE MODULE
DATA GATEWAYS
Purpose:
Data moving and export
Cloud services and Object Storage
DATA PLATFORM
DIRECT LIQUID
COOLED MODULES
200 nodes
573 nodes
GPU
GPU AI ACCELERATOR
Purpose:
Accelerated HPC, HPDA
& AI workloads
CPU
CLUSTER MODULE
Purpose:
Mixed HPC, HPDA
& BigData workloads

MeluXina **User Software Environment (MUSE)**





LXP-managed global software stack

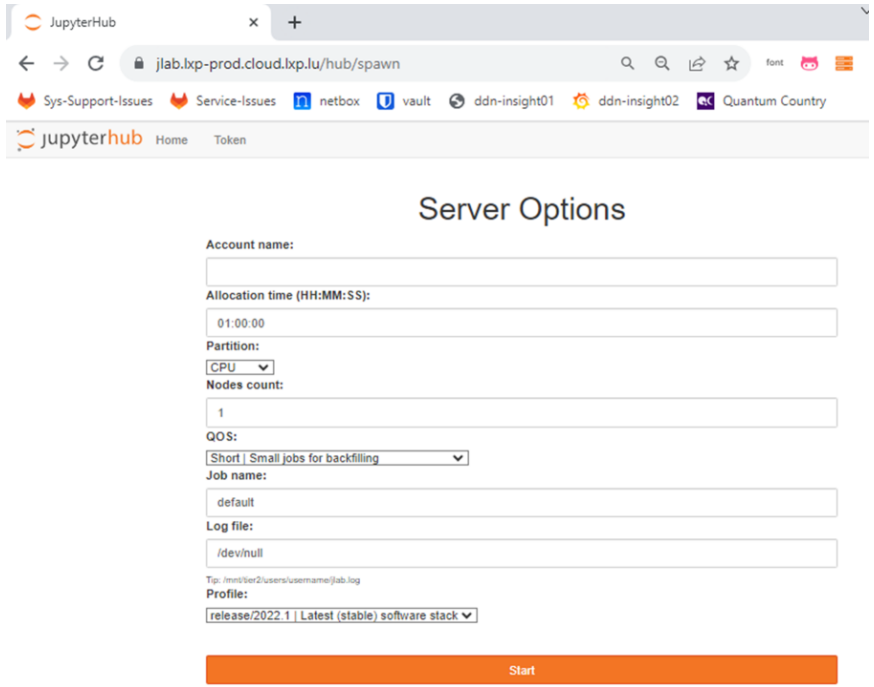
- engineering, data science, AI, chemistry, physics libraries and applications
- managed deployment with software environment management tools (EasyBuild + Lmod)
- Optimized builds and toolchains, incl. integration with parallelization frameworks
- Containers

Bring-your-own software stack

- Easy control of complete stack **reproducibility**
- Create tooling and pipeline on user's infrastructure **run pipeline on MeluXina**

- Docker & OCI compatible
- Non-privileged mode for improved security
- Support for GPU accelerated applications
- Support for creating and running encrypted containers
- Support for trusted containers: PGP signed & verified

Accessing MeluXina



On-premise JupyterLab

- https://jlab.lxp-prod.cloud.lxp.lu/

2FA authentication using KeyCloak

Steps to setup KeyCloak 2FA

- https://docs.lxp.lu/cloud/keycloak/

Jlab service to be reimagined in 2024

ssh *user*@login.lxp.lu –p 8822

- 4 login nodes in total, at the moment round-robin between 2 login nodes
- Login nodes directly accessible (not recommended)
- Can switch between login nodes

Ed25519 key pair-based authentication

- Passphrase on private key highly recommended

Steps to generate appropriate key

- https://docs.lxp.lu/first-steps/connecting/

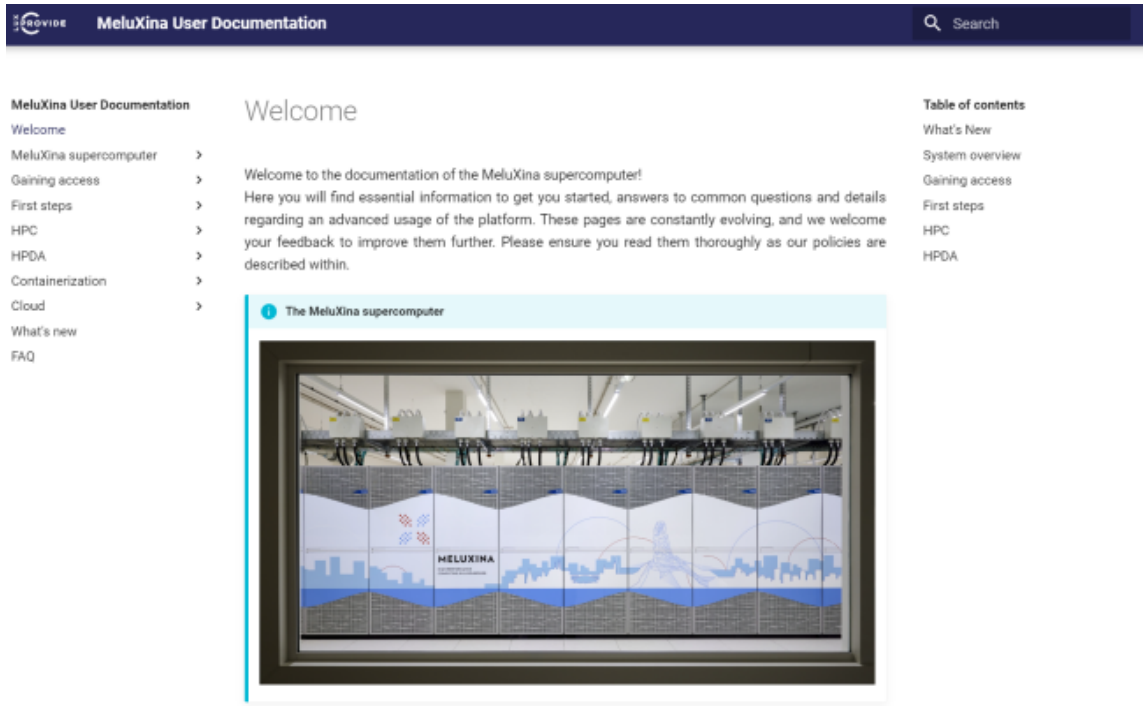
MeluXina workload manager - policies

Policy	Details
Compute budget	Monthly pro rata of overall project compute budget, applied every 1st of the month
Compute quotas	Enforced through SLURM accounting and QOS
Node scheduling	Nodes dedicated to job – no node sharing with other users, 1 core job reserves complete node
Hyperthreading	Hyperthreading enabled by default on compute nodes, can be disabled through SLURM options
Interactive usage	Dedicated resources per partition ensuring rapid development/prototyping

QOS	Max. Time (hh:mm)	Max. nodes per job	Priority	Used for..
dev	06:00	1	Regular	Interactive executions for code/workflow development, with a maximum of 1 job per user; QOS linked to special reservations
test	00:30	5%	High	Testing and debugging, with a maximum of 1 job per user
short	06:00	5%	Regular	Small jobs for backfilling
short-preempt	06:00	5%	Regular	Small jobs for backfilling
default	48:00	25%	Regular	Standard QOS for production jobs
long	144:00	5%	Low	Non-scalable executions with a maximum of 1 job per user
large	24:00	70%	Regular	Very large scale executions by special arrangement, max 1 job per user
urgent	06:00	5%	Very high	Urgent computing needs, by special arrangement, can preempt 'short-preempt' QOS

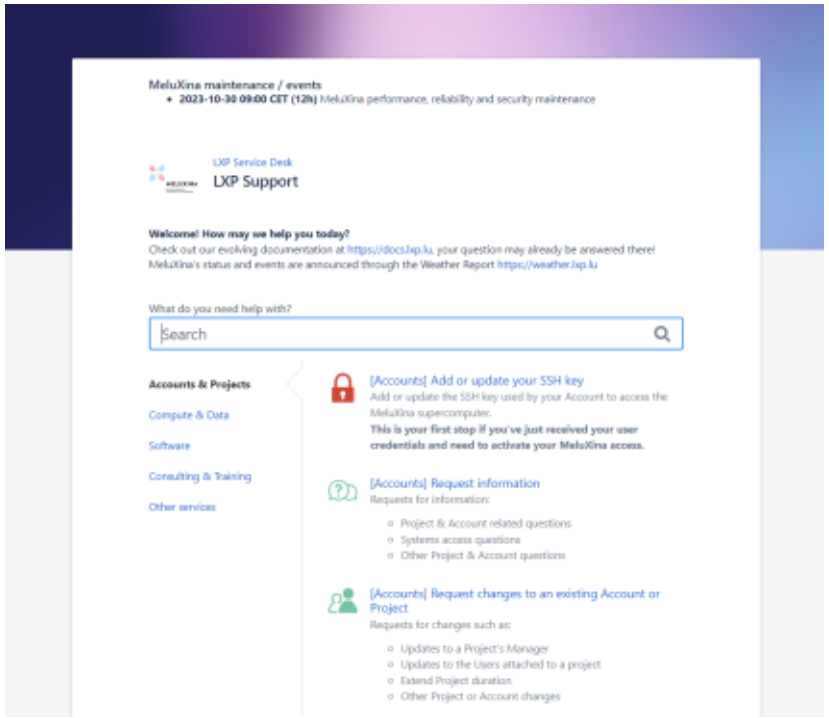
Getting help

https://docs.lxp.lu/



Publicly available

https://servicedesk.lxp.lu

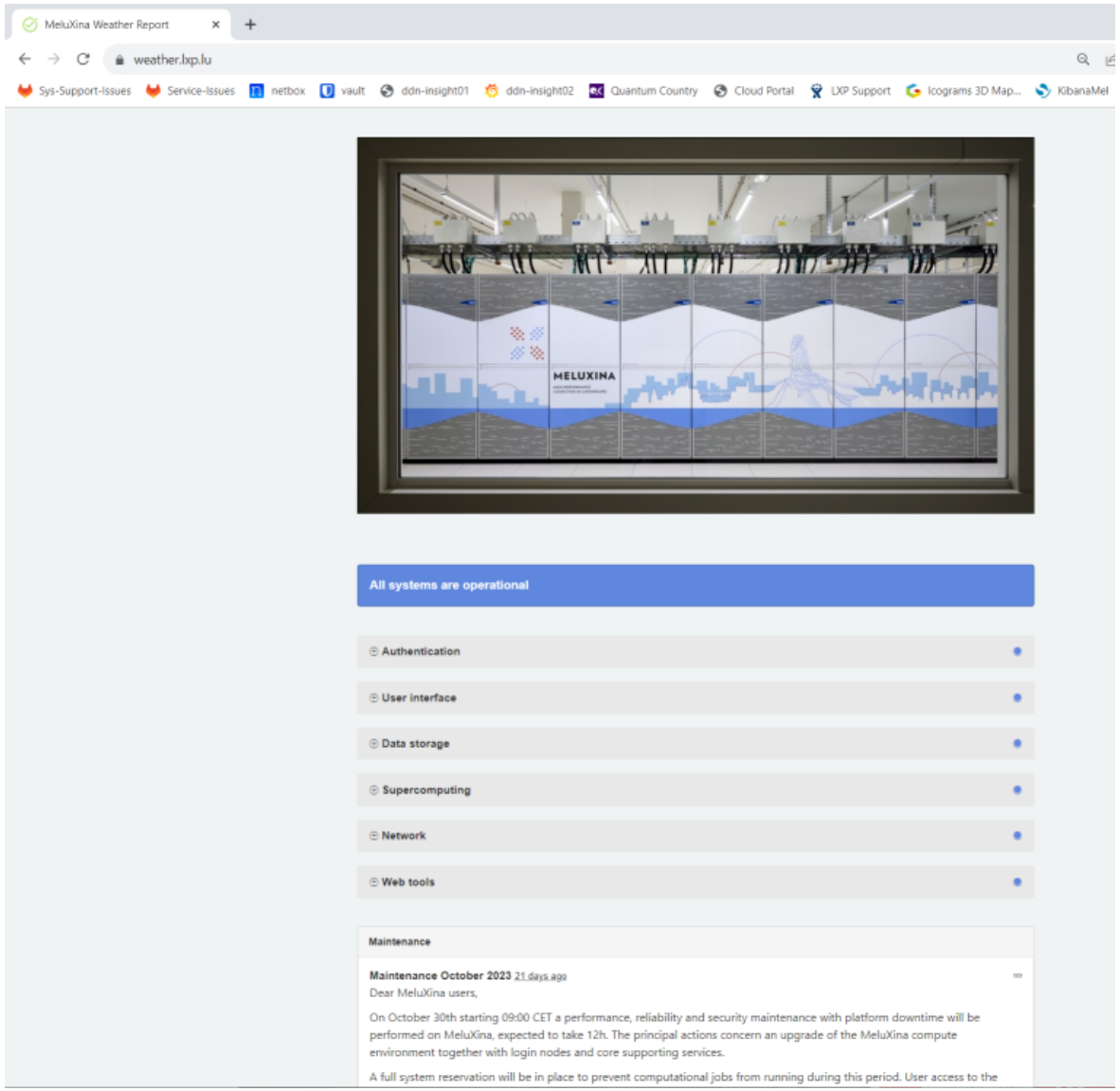


Requires existing user account

MeluXina Weather page

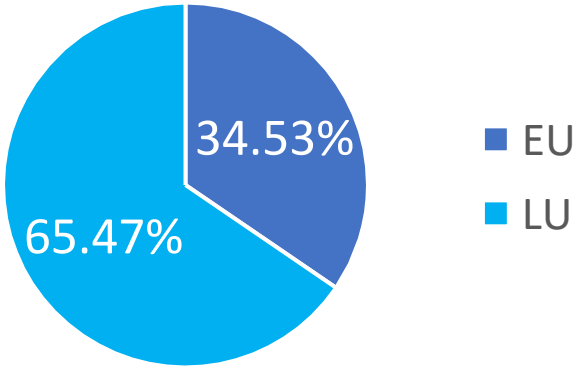
- Current system status
- Ongoing events
- Planned maintenances/updates

weather.lxp.lu



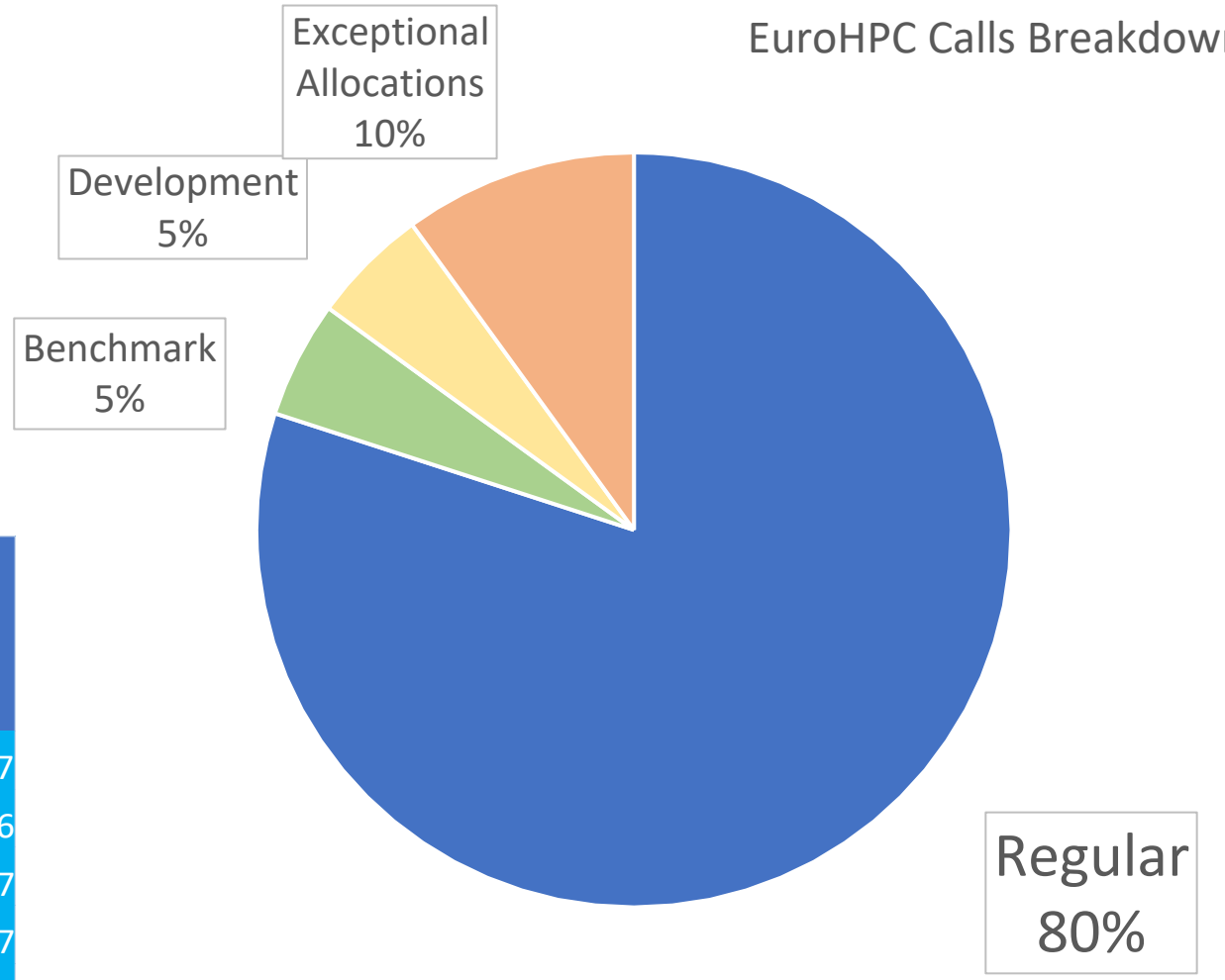
EuroHPC compute time on MeluXina

MeluXina shares



Compute Module	EuroHPC node-hours per year
Cluster	1,733,227
Accelerator – GPU	604,966
Accelerator – FPGA	60,497
Large Memory	60,497
Total	2,459,187

EuroHPC Calls Breakdown



Project budgets

- **Checking compute & data project quotas:** ***myquota*** **tool**

```
[fbongiovanni@login02 ~]$ myquota
                                        COMPUTE ALLOCATIONS FOR CURRENT MONTH
                                         COMPUTE USAGE FROM 2023-10-01 to NOW

Project                              CPU node-hours              GPU node-hours              FPGA node-hours              LargeMem node-hours
                                   Used    Max   %used         Used    Max   %used         Used    Max   %used          Used    Max   %used
lxp                                1223                        1151                         0                            0

                                                   DATA ALLOCATIONS

 Datapath                                    Data GiB                              No. Files
                                       Used |    Max  | Use%              Used |    Max     | Use%
/home/users/fbongiovanni                 17 |     100 |  17%             61352 |     100000 |  61%
/project/scratch/lxp                      - |       - |   -%                 - |          - |   -%
/project/home/lxp                         - |       - |   -%                 - |          - |   -%
/project/scratch/lxp_apps                 - |       - |   -%                 - |          - |   -%
/project/home/lxp_apps                 2893 |    5000 |  57%           7432137 |   10000000 |  74%
[fbongiovanni@login02 ~]$
```

- **Accessible from login/compute nodes** **(on compute node via sticky module)**
- **Accessible to project coordinators and members**

Energy monitoring on MeluXina

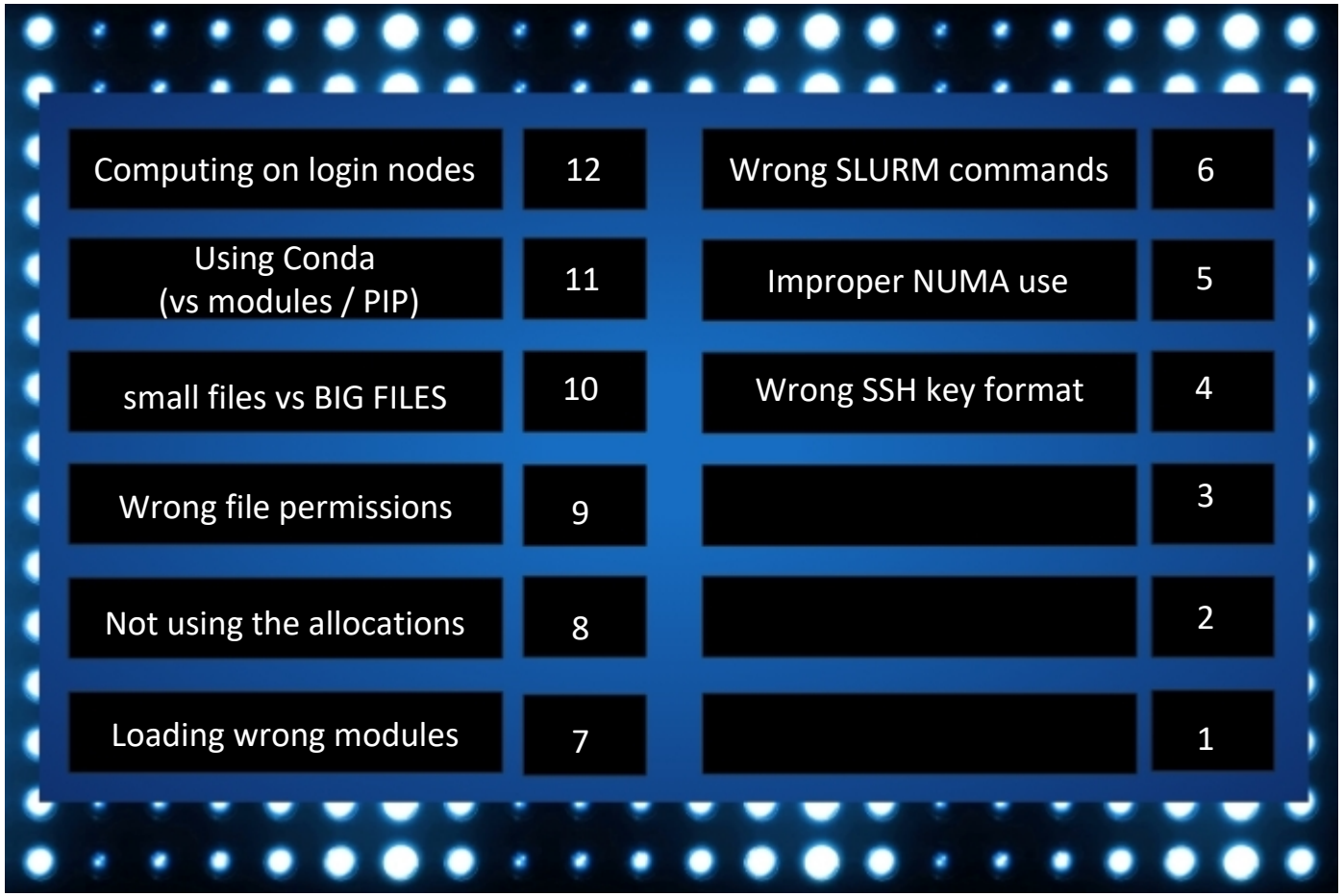
- Enabled at SLURM level since October 2023
  - 30Hz frequency
  - Energy reported in Joules
  - Monitoring via native Slurm commands (sacct, sreport)
- Caveat
  - Need to use **srun** as task launcher!
  - Slurm then monitors the job step

```
$ sacct -j JOBID -o jobid,jobname,partition,account,state,consumedenergyraw
```

```
$ sacct -j 497558 -o jobid,jobname,partition,account,state,consumedenergyraw

JobID           JobName  Partition    Account      State ConsumedEnergyRaw
------     ---------- ---------- ---------- ---------- -----------------
497558       gromacs-g+        gpu        lxp  COMPLETED            116555
497558.batch      batch                   lxp  COMPLETED                 0
497558.0        gmx_mpi                   lxp  COMPLETED            116555
```

Discouraged user behaviours



Encouraged user behaviours

Reading the documentation

Providing supporting documents in support requests

Providing reproducible test cases

Giving feedback