RAISE
Center of Excellence

Modular and heterogeneous supercomputing architectures - Status and next steps

Gulllaume Houzeaux with contributions

from the partner HPC centers

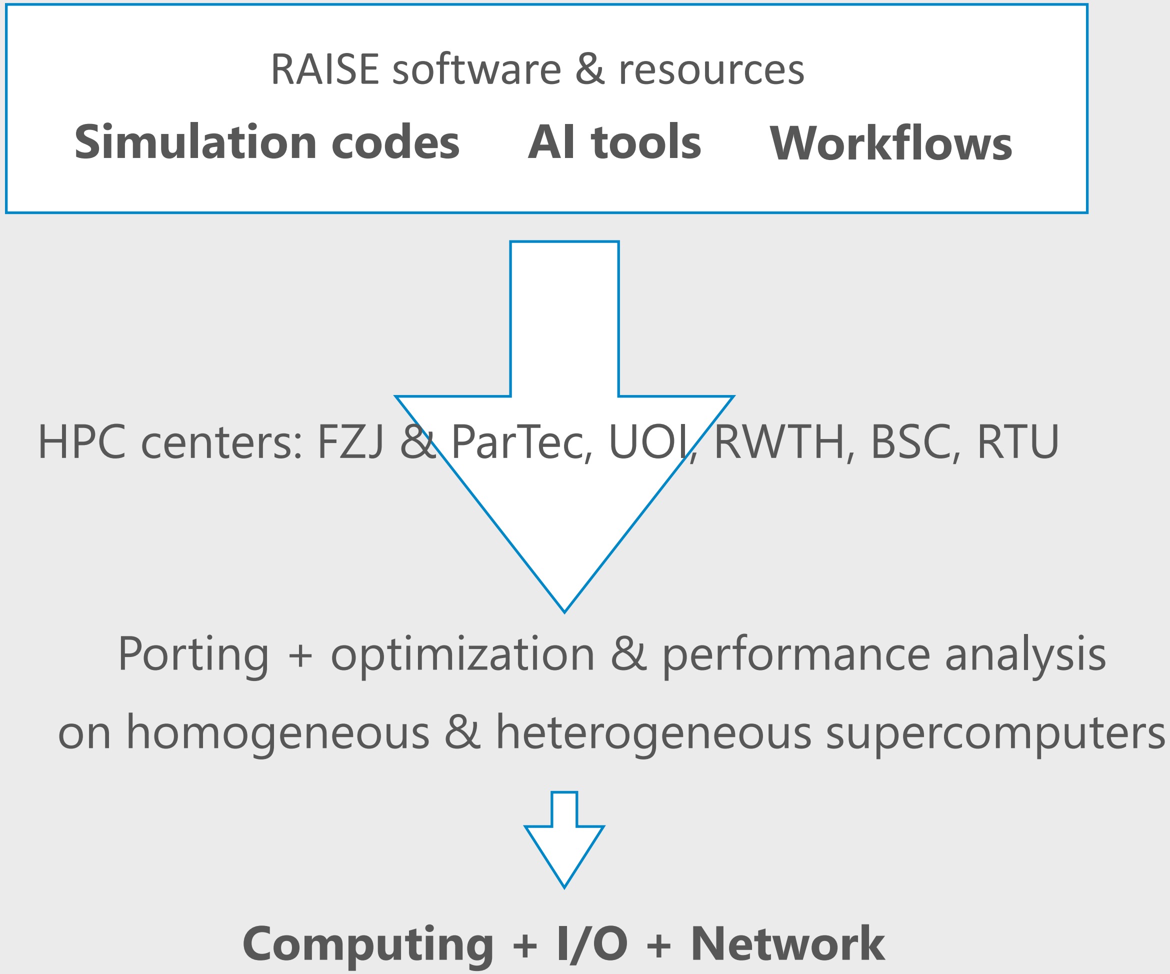
Modular and heterogeneous supercomputing architectures

Task 2.1 Modular and heterogeneous supercomputing architectures <Leader: BSC> <M1- M36>
Contributors: FZJ & ParTec, UOI, RWTH, BSC, RTU Outputs: D2.1, D2.2, D2.3, D2.4

In this task, the HPC centers of RAISE provide their computational resources for development and testing of RAISE's software. This includes the homogeneous and heterogeneous HPC systems found at the Tier-2 and Tier-3 centers of the consortium (UOI, RWTH, and RTU) as well as the cutting- edge HPC systems of the Tier-0 and Tier-1 providers (FZJ and BSC). That is, FZJ and BSC give access to their MSA/heterogeneous supercomputers JURECA, JUWELS, and MareNostrum 4 (MN4) general cluster and its three prototypes (see Secs. 4.1.1 and 4.1.5). Together with the use-case providers, the representative applications, where necessary, are jointly prepared for such heterogeneous systems together with the HPC centers of the consortium. The AI experts further exploit the HPC architectures and file systems for relevant algorithms, i.e., they work on the enhancement of the scalability of existing ML/DL methods on dedicated or shared HPC components using, e.g., GPGPUs or other accelerator platforms, and test I/O performance. The developers are supported by the HPC experts by means of performance engineering activities, best practice guidelines, and system-specific tutorials and manuals. The training courses of the HPC centers are offered to the use-case providers in line with Task 6.1.

D2.1 (BSC) Best practice guidelines and tutorials for the various HPC systems are available (M2)
D2.2/3/4 (BSC) Report on porting and performance engineering activities (M12,M24,M36)

Modular and heterogeneous supercomputing architectures



RAISE software & resources

Simulation codes **AI tools** **Workflows**

HPC centers: FZJ & ParTec, UOI, RWTH, BSC, RTU

Porting + optimization & performance analysis

on homogeneous & heterogeneous supercomputers

Computing + I/O + Network

Modular and heterogeneous supercomputing architectures

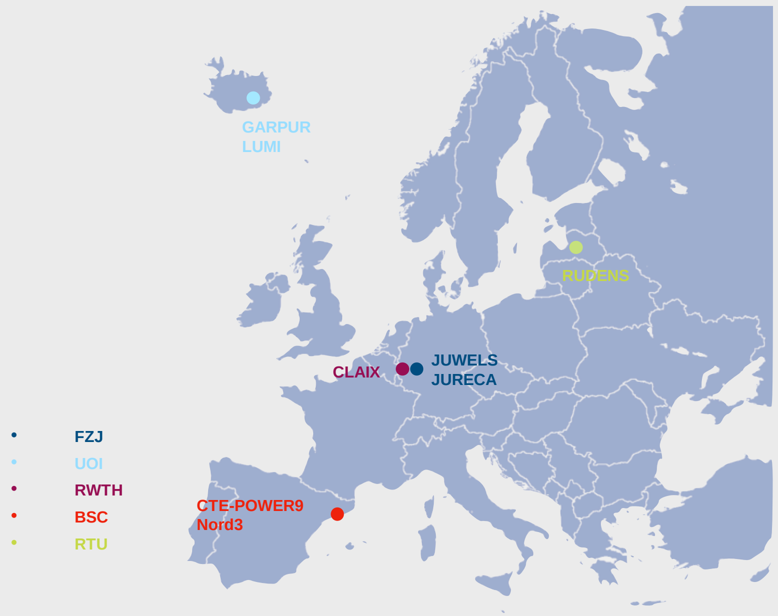
- **Where do we get supercomputing resources?**
  - Local supercomputers
    - TIER-3, TIER-2: UOI, RWTH, RTU
    - TIER-1, TIER-0: FZJ, BSC
  - PRACE specific access calls to COEs:
    - 2021-1 (1.6M), 2021-2 (0.6M), 2022-1 (0.5M)
  - Competitive access calls

Partner	Supercomputer	Hours
CERFACS	GENCI Jean Zay	20k GPU
CERFACS	GENCI Jean Zay	20k GPU
RWTH	JUWELS, GCS	6M CPU
CYI	Cyclone, acNational	1M CPU
CYI	VEGA, EuroHPC	12M CPU
FM	Cloud, regional center	35k CPU
RTU	In-kind for RAISE	550k CPU
FZJ	JURECA-DC GPU, JARA	8.6M GPU
FZJ/CERN/UOI	JURECA-DC GPU, JARA	8.3M GPU
FZJ/CERN/UOI	JUPSI, JARA	5 hours

Modular and heterogeneous supercomputing architectures

D2.1 Best practice guidelines and tutorials for the various HPC systems

- 4 countries
- 8 systems



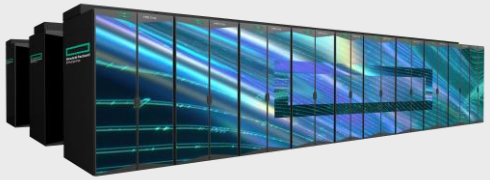
CTE-Power, Nord3

JUWELS

JURECA

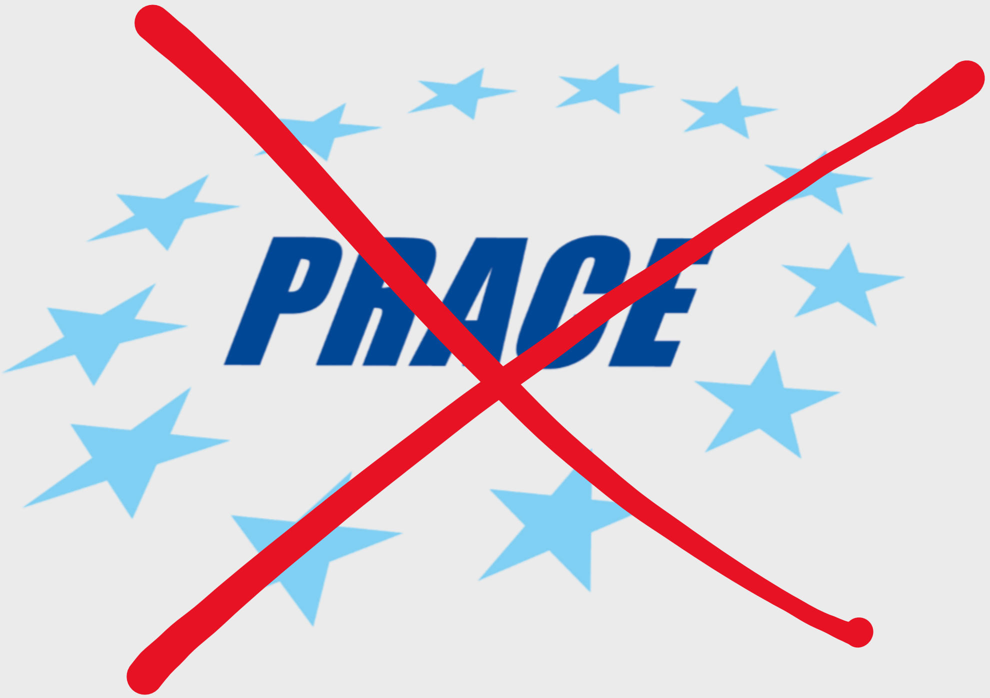
RUDENS

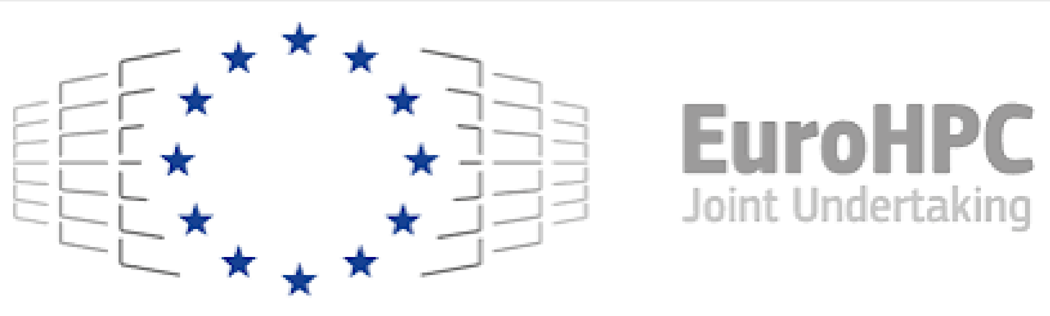
CLAIX



GARPUR & LUMI

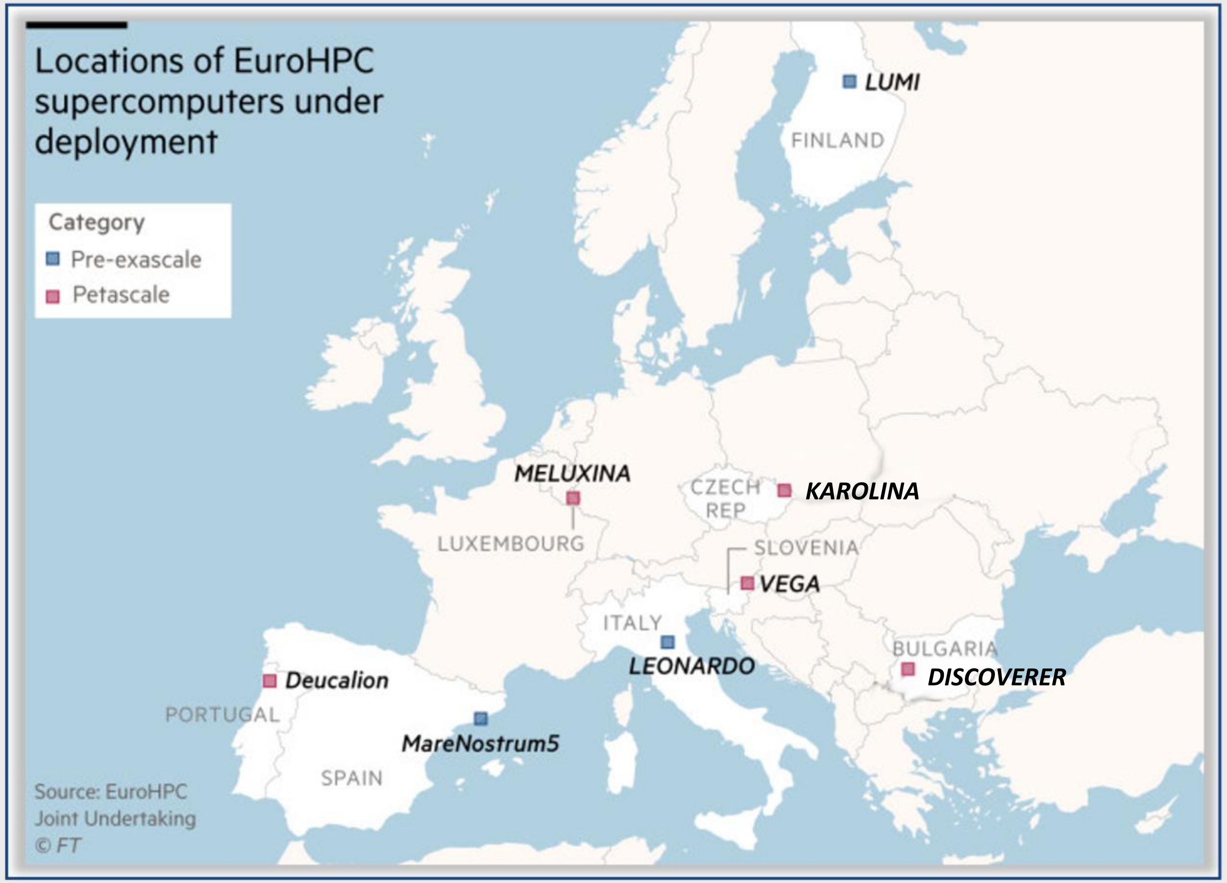
Modular and heterogeneous supercomputing architectures





Modular and heterogeneous supercomputing architectures

EURO-HPC: Towards European HPC Technologies


Locations of EuroHPC supercomputers under deployment

Source: EuroHPC Joint Undertaking
© FT

https://eurohpc-ju.europa.eu/about/our-supercomputers_en

Modular and heterogeneous supercomputing architectures

LUMI: 550 PFs

CPU: AMD EPYC. GPU: AMD Instinct

LEONARDO: 323 PFs

CPU: Intel Ice-Lake/Sapphir. GPU: NVIDIA Ampere

Marenostrum5: 314 PFs

CPU: Intel Sapphir/Emerald, NVIDIA Grace. GPU: NVIDIA Hopper, Intel Rialto Bridge

VEGA: 10 PFs

CPU: AMD Epyc. GPU: Nvidia A100

MELUXINA: 18 PFs

CPU: AMD Epyc. GPU: Nvidia A100

KAROLINA: 13 PFs

CPU: AMD, Intel Xeon. GPU: Nvidia A100

DISCOVERER: 6 PFs

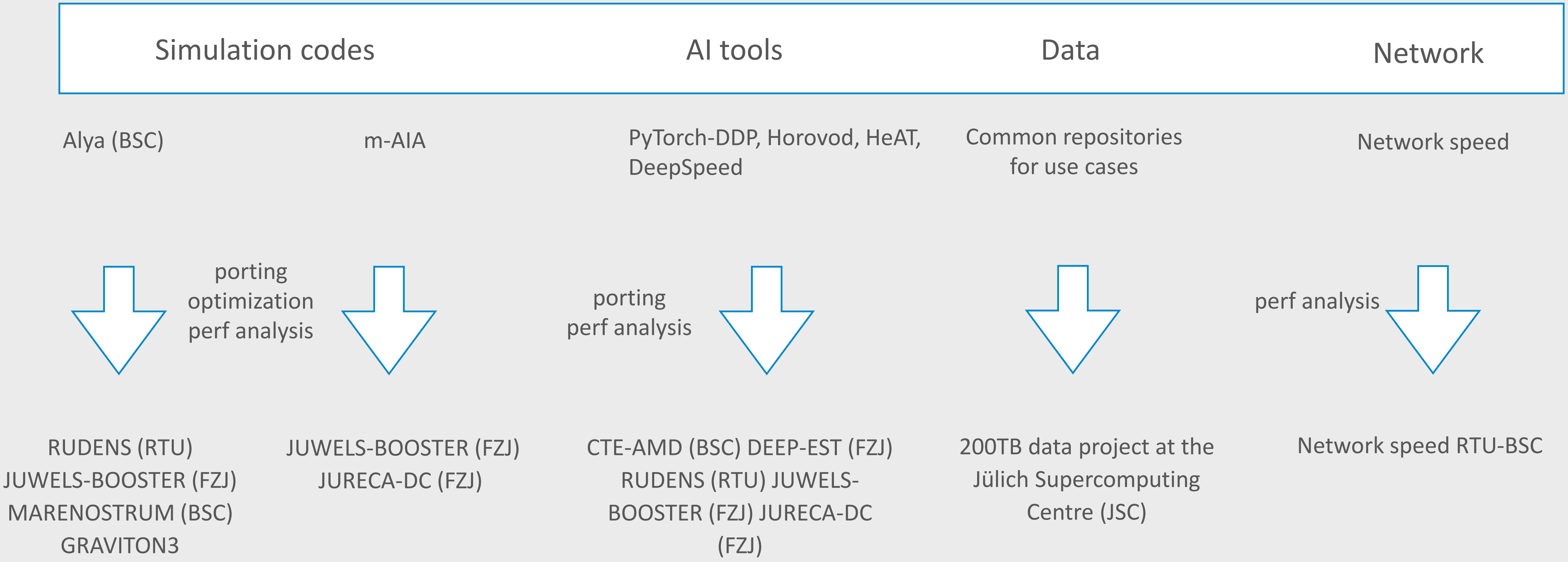
CPU: AMD Epyc.

DEUCALION: 10 PFs

CPU: ARM A64FX, AMD Epyc. GPU: NVIDIA Ampere

Modular and heterogeneous supercomputing architectures

D2.2/2.3 Report on porting and performance engineering activities



Simulation codes

Alya (BSC)

porting
optimization
perf analysis

RUDENS (RTU)
JUWELS-BOOSTER (FZJ)
MARENOSTRUM (BSC)
GRAVITON3

m-AIA

JUWELS-BOOSTER (FZJ)
JURECA-DC (FZJ)

AI tools

PyTorch-DDP, Horovod, HeAT, DeepSpeed

porting
perf analysis

CTE-AMD (BSC) DEEP-EST (FZJ) RUDENS (RTU) JUWELS-BOOSTER (FZJ) JURECA-DC (FZJ)

Data

Common repositories for use cases

200TB data project at the Jülich Supercomputing Centre (JSC)

Network

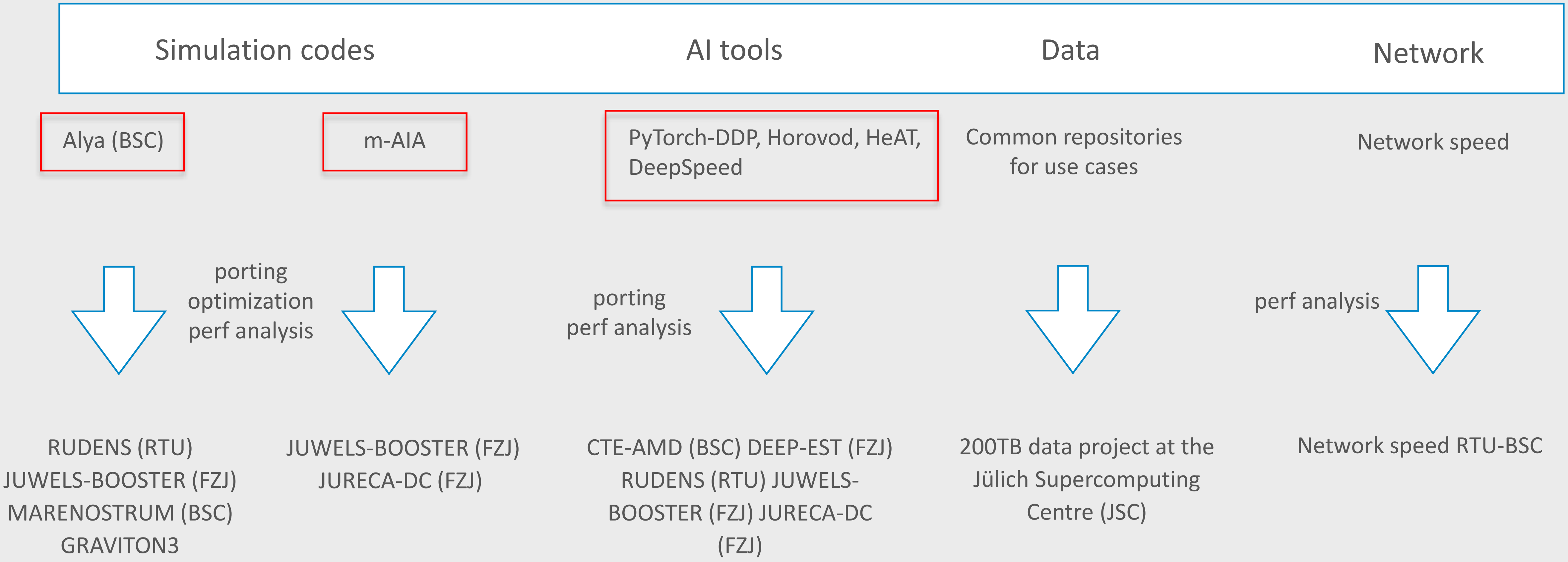
Network speed

perf analysis

Network speed RTU-BSC

Modular and heterogeneous supercomputing architectures

D2.2/2.3 Report on porting and performance engineering activities



Simulation codes

AI tools

Data

Network

Alya (BSC)

m-AIA

porting
optimization
perf analysis

RUDENS (RTU)
JUWELS-BOOSTER (FZJ)
MARENOSTRUM (BSC)
GRAVITON3

JUWELS-BOOSTER (FZJ)
JURECA-DC (FZJ)

PyTorch-DDP, Horovod, HeAT, DeepSpeed

porting
perf analysis

CTE-AMD (BSC) DEEP-EST (FZJ) RUDENS (RTU) JUWELS-BOOSTER (FZJ) JURECA-DC (FZJ)

Common repositories
for use cases

200TB data project at the Jülich Supercomputing Centre (JSC)

Network speed

perf analysis

Network speed RTU-BSC

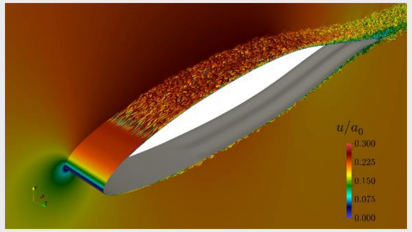
CFD codes

Porting, optimizations and performance analysis

Porting, optimization and performance analysis of RAISE software

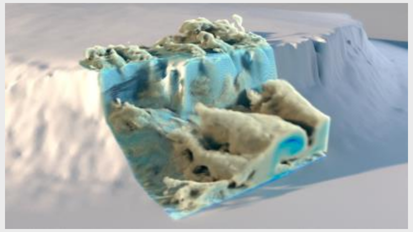
m-AIA

Task 3.1 for the simulation of Turbulent Boundary Layer (TBL) flows over an actuated surface



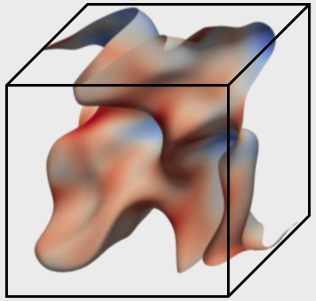
Alya

Task 3.2 for simulating the flow over wind turbines and in windfarms



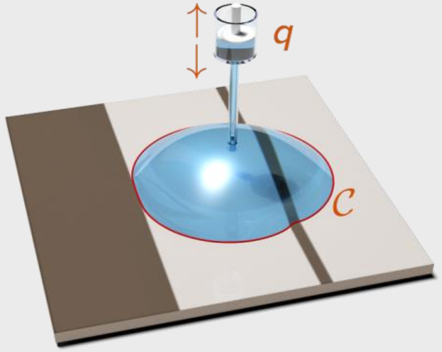
AVBP

Tasks 3.3 and 3.4 for simulating reactive flows and next generation aircraft engines



Basilisk

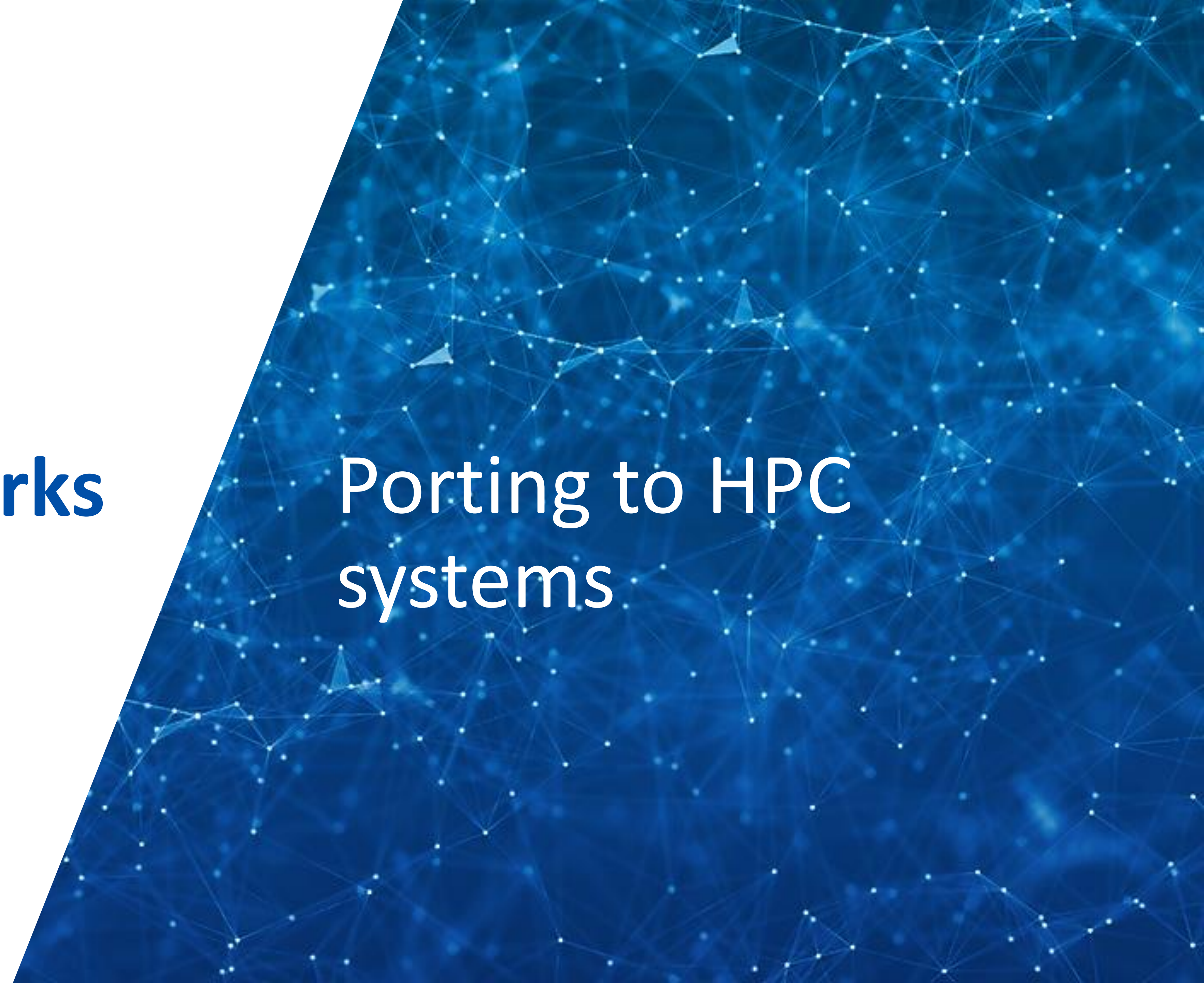
Task 3.5 for the simulation of droplet behavior on various surfaces



Conclusion & outlook

- **Alya: vectorization and sliding mesh strategy**
  - Finite element assembly speedup of 2
  - Fractional step matrices speedup of 4.3
  - Good vectorization speedup of 5 on Graviton3
  - Vectorization should be optimized for implicit scheme
  - Coupling cost is moderate
- **M-AIA: porting to GPU**
  - M-AIA with GPU support is ported to TIER-1 HPC Systems in FZJ
  - PSTL accelerates CFD solvers
  - Scaling is still a problem
  - Lattice-Boltzmann methods (instead of FVM) could better fit!
  - Work-done is documented in a Git Repo: https://git.rwth-aachen.de/aia/MAIA/Solver/-/tree/structured_pstl

AI frameworks



Porting to HPC systems

Porting existing ML frameworks

- Tier-0/1 + Prototype HPCs

Location	System name	CPU	Accelerators
CINECA	Marconi100	1,960 IBM POWER9	3,920 NVIDIA V100
FZJ	Juwels Cluster + Booster	2,560 Intel Xeon	3,774 NVIDIA A100
FZJ	Jureca DC	1,536 AMD EPYC	768 NVIDIA A100
FZJ	DEEP-EST	147 Intel Xeon	75 NVIDIA V100
FZJ	JUAWEI	11 ARM HiSilicon	
HLRS	Hawk	11,264 AMD EPYC	64 NVIDIA V100*
CSCS	Piz Daint	9,330 Intel Xeon	68,448 NVIDIA P100
BSC	Marenostrum4	6,912 Intel Xeon	
BSC	CTE-AMD	33 AMD EPYC	
BSC	CTE-ARM	192 ARM A64FX	
BSC	HUAWEI	16 ARM Kunpeng 920	
CEA	Joliot-Curie	2,484 Intel Xeon + 2,292	
LRZ	SuperMUC-NG	6,480 Intel Xeon	

Individual care:

- Software stack
- Code optimization
- Node communication

*cloud nodes

Conclusion & outlook

- Existing frameworks are ported to several HPC systems
- Good scaling performance up to 128 GPUs,
  - larger dataset needed for more GPUs
- Reduced runtimes with various optimizations (next talk for details)

Outlook:

- Test model parallelism
- Automation (container support!)
- Test intelligent processing units (IPUs)

Coupling CFD/ML with OpenFoam

Performance analysis

Conclusion & outlook

- Very preliminary results:
  - ANN is inferred at the end of each CFD timestep
  - Work group communicator for each GPU
  - Communication with GPU is carried out by each Rank 0 of each workgroup
  - MPI_Gatherv performed for ANN input
  - MPI_Scatterv performed for ANN output
  - Inference is a highly data-parallel task and thus it scales perfectly with an increasing number of accelerator devices or CPU sockets!

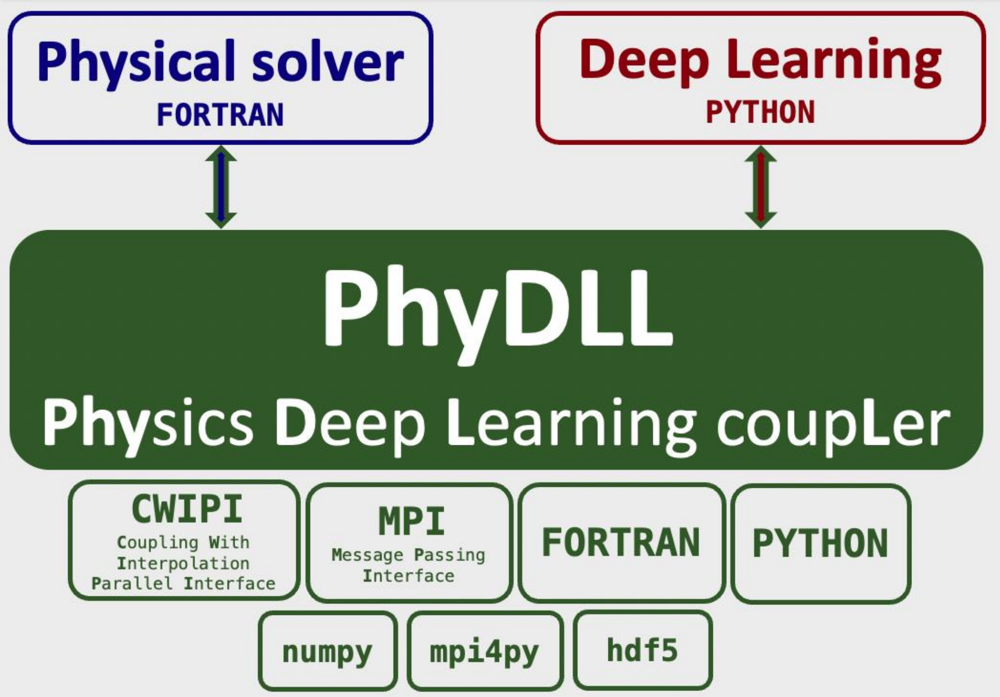
PhyDLL

Porting, and intergration with Alya

Installation of PhyDLL

- Installation of PhyDLL

PhyDLL (fidɛl) (**Phy**sics **D**eep **L**earning coup**L**er) is an open-source library to couple massively parallel physical solvers to distributed deep learning inference.


Physical solver
FORTRAN
Deep Learning
PYTHON
PhyDLL
Physics Deep Learning coupLer
CWIPI
Coupling With Interpolation Parallel Interface
MPI
Message Passing Interface
FORTRAN
PYTHON
numpy
mpi4py
hdf5

https://phydll.readthedocs.io/en/latest/

Installation of PhyDLL

- Installation of PhyDLL
  - Test to communicate a python and Fortran using MPI in supercomputers with GPU and CPU support
    - Writing simple codes for testing
    - Testing on CTE-AMD (we did not succeed)
    - Testing on CTE-Power
  - Finding or installing a version of phyton on CTE-Power with a mpi4py library that have the same MPI Fortran compiler
    - Testing with different versions of python, 3.7.4 finally works
  - Installing PhyDLL on CTE-Power

PhyDLL+Alya

- Installation of PhyDLL+Alya
  - Installing Alya on CTE-Power
  - Loading right modules on CTE-Power for making Alya work together with PhyDLL
  - Test to communicate Alya and pyhton with PhyDLL
    - Finding lines of code in Alya where call PhyDLL subroutines
    - Writing simple test to testing
  - Using a real example for assign porosity
    - Sending right velocity data from Alya to python using PhyDLL
    - Sending calculated porosity from python to Alya using PhyDLL
    - Receiving and sending data dynamically in execution time.

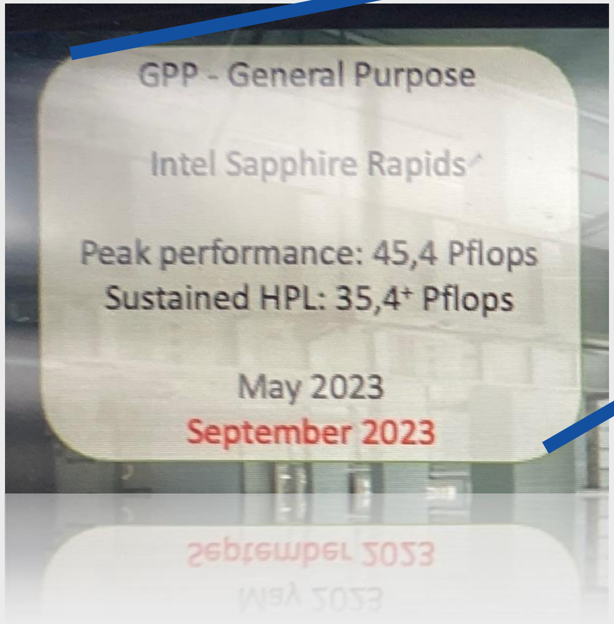
What’s next?

What’s next?

Porting to LUMI?

What’s next?

Marenostrum V

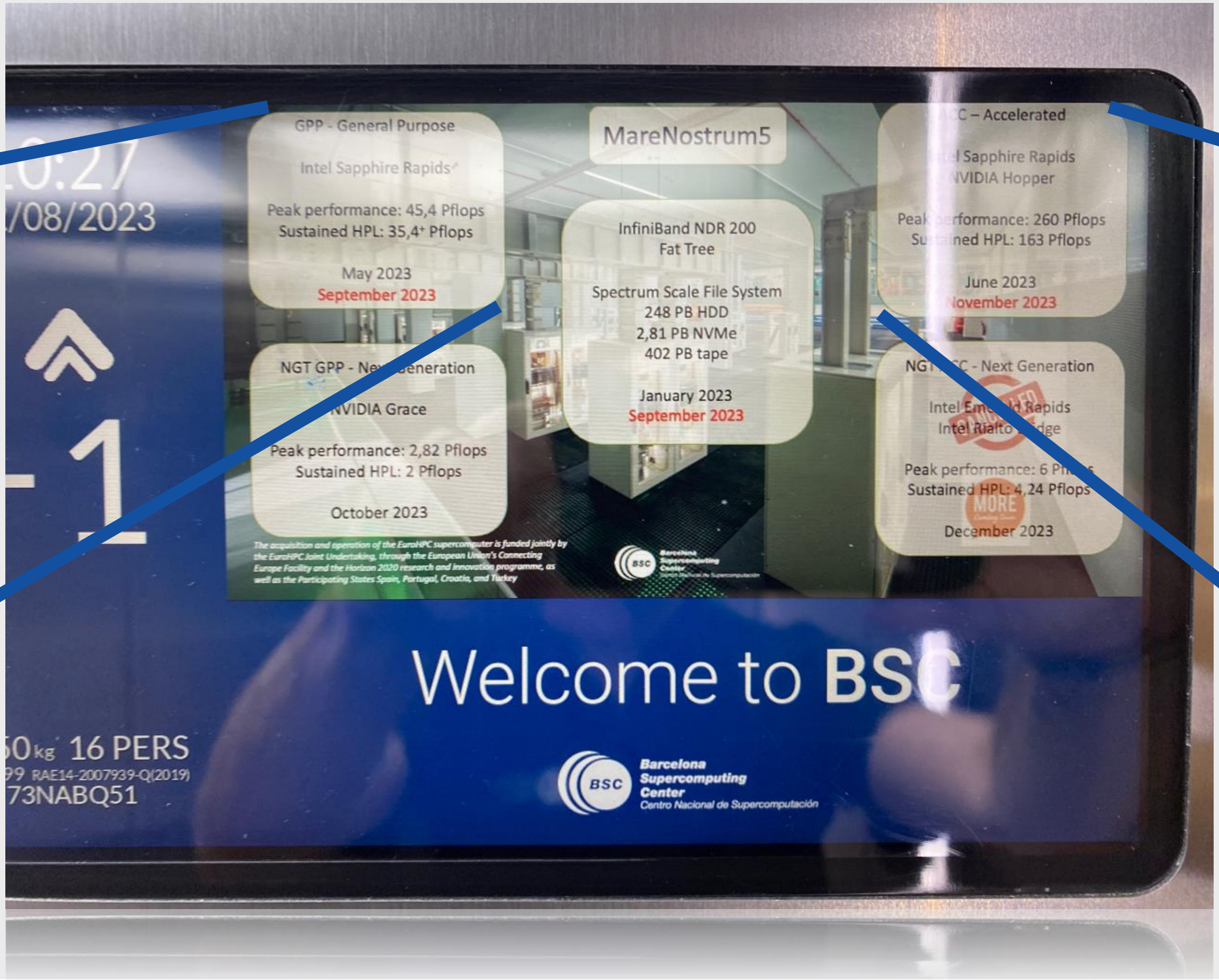


GPP - General Purpose

Intel Sapphire Rapids

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

May 2023
September 2023



0:27
/08/2023
-1
0kg 16 PERS
99 RAE14-2007939-Q(2019)
73NABQ51

MareNostrum5

GPP - General Purpose

Intel Sapphire Rapids

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

May 2023
September 2023

NGT GPP - Ne Generation

NVIDIA Grace

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

October 2023

InfiniBand NDR 200
Fat Tree

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

January 2023
September 2023

C – Accelerated

l Sapphire Rapids
NVIDIA Hopper

Peak erformance: 260 Pflops
Su ned HPL: 163 Pflops

June 2023
November 2023

NGT C - Next Generation

Intel Em ld Rapids
Intel Rialto idge

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

December 2023

CANCELLED

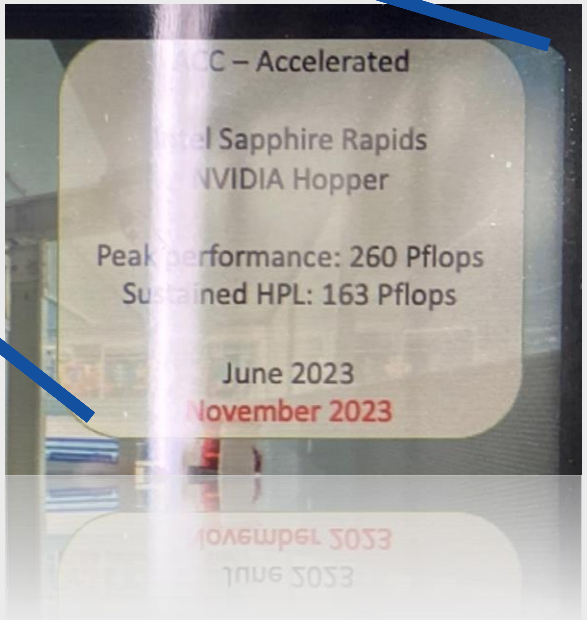
MORE

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, Croatia, and Turkey

BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación

Welcome to BSC

BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación



CC – Accelerated

tel Sapphire Rapids
NVIDIA Hopper

Peak erformance: 260 Pflops
Su ined HPL: 163 Pflops

June 2023
November 2023

What’s next?

Marenostrum V

- **CPU: Sapphire Rapids** is a codename for Intel's server (fourth generation Xeon Scalable) and workstation processors based on Intel 7

Super-Charged CPU

Sapphire Rapids HBM

Sampling Today
Production 2H ’22

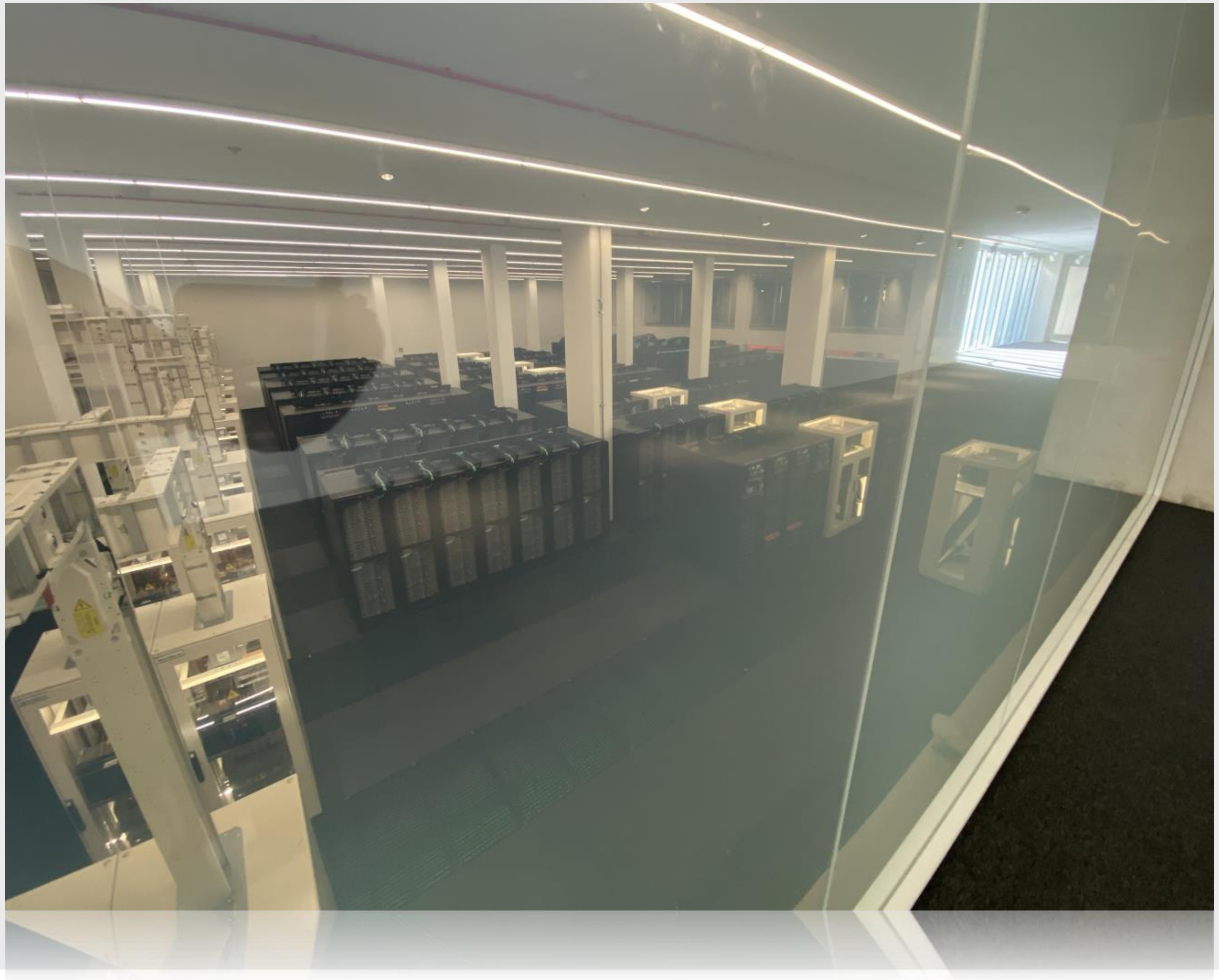
Up to 2.8x Perf
over 3rd Gen Xeon[1]

[1] Based on pre-production measurements.
Learn more at www.intel.com/PerformanceIndex. Results may vary.

Investor Meeting 2022

- **GPU: NVIDIA Hopper** is a graphics processing unit (GPU) microarchitecture developed by Nvidia. It is designed for datacenters and is parallel to Ada Lovelace.

What's next?



drive. enable. innovate.

The CoE RAISE project has received funding from
the European Union's Horizon 2020 –
Research and Innovation Framework Programme
H2020-INFRAEDI-2019-1 under grant agreement no. 951733

Follow us: