MAX PLANCK
COMPUTING & DATA FACILITY

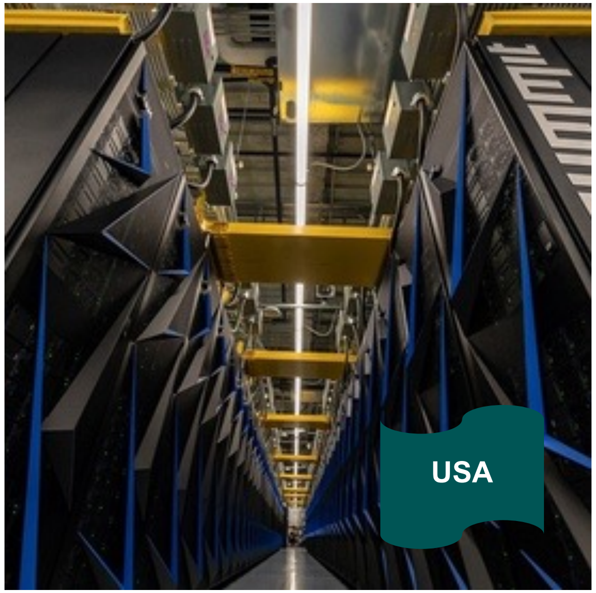
MPCDF

HPC (AND CLOUD) IN EUROPE AND AT THE MAX PLANCK SOCIETY

Erwin Laure, Director MPCDF
Markus Rampp, Deputy Director MPCDF
Salishan Conference, 2023-04-25

CRAY
AK RIDGE
ati nal Labor tory
U.S. DEPARTMENT OF
ENERGY
FRO
CRAY
AMD
USA

Japan



USA

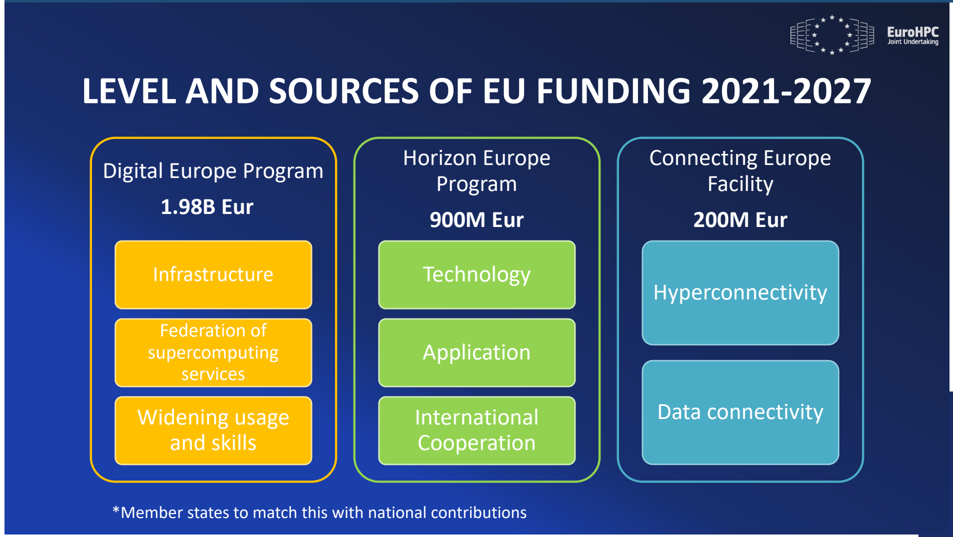
神
威
China

AURE &

China

USA

The EuroHPC Joint Undertaking (since 2018)



LEVEL AND SOURCES OF EU FUNDING 2021-2027

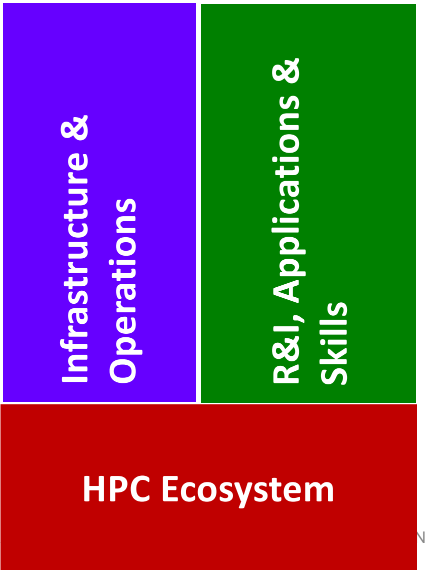
Digital Europe Program
1.98B Eur
- Infrastructure
- Federation of supercomputing services
- Widening usage and skills

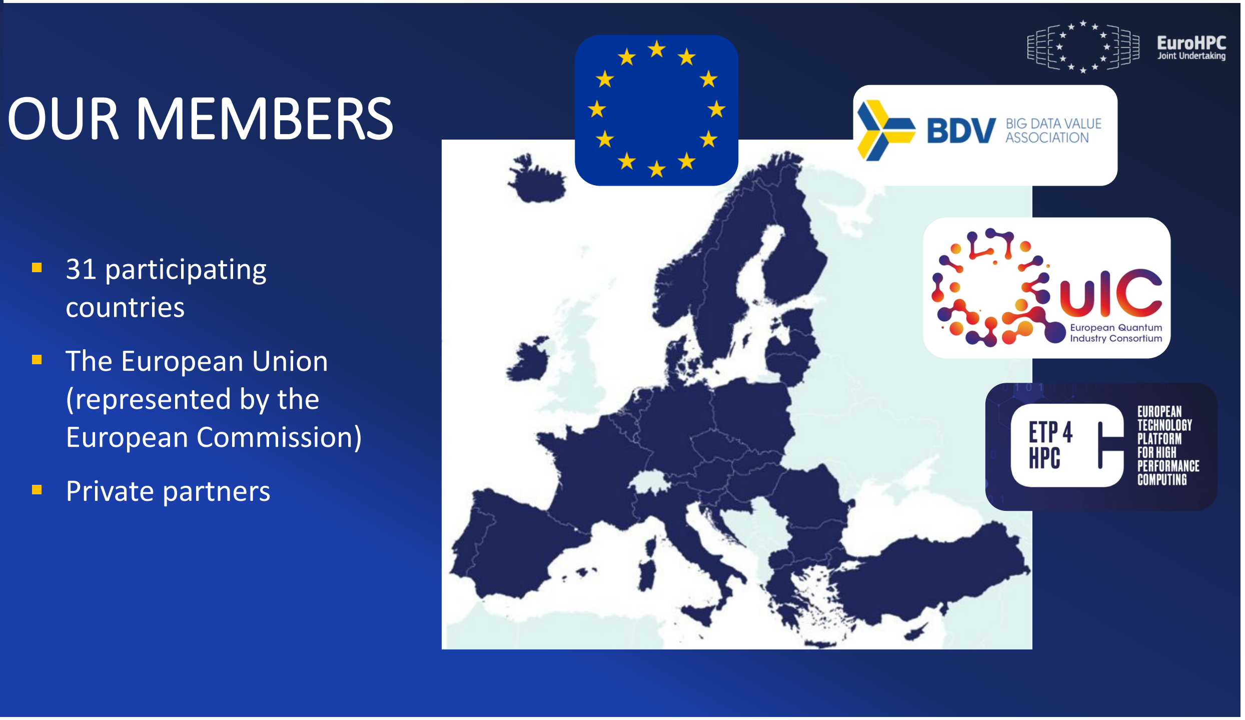
Horizon Europe Program
900M Eur
- Technology
- Application
- International Cooperation

Connecting Europe Facility
200M Eur
- Hyperconnectivity
- Data connectivity

*Member states to match this with national contributions

Mission: Establish an integrated world-class supercomputing & data infrastructure and support a highly competitive and innovative HPC and Big Data ecosystem



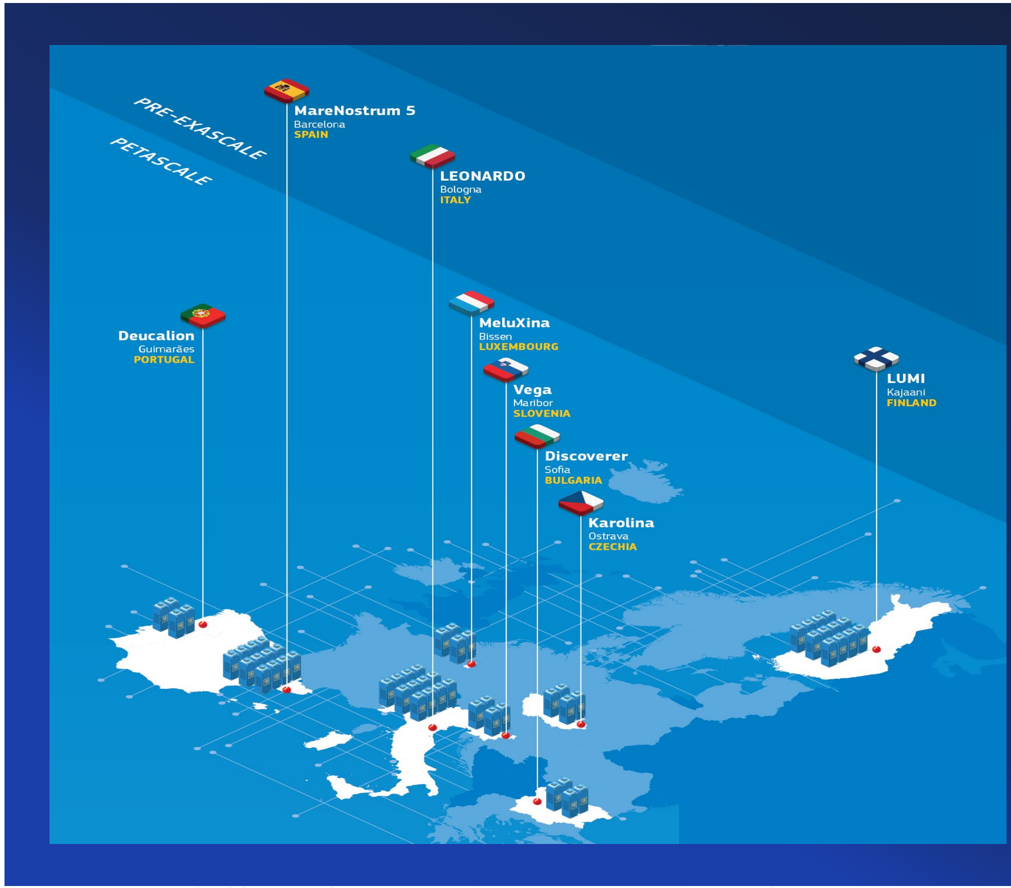

EuroHPC Joint Undertaking
BDV BIG DATA VALUE ASSOCIATION
QIC European Quantum Industry Consortium
ETP 4 HPC EUROPEAN TECHNOLOGY PLATFORM FOR HIGH PERFORMANCE COMPUTING

OUR MEMBERS

- 31 participating countries
- The European Union (represented by the European Commission)
- Private partners

EUROHPC SYSTEMS

- 1 more pre-exascale @ BSC, 2023
- “Jupiter”@Jülich will be the first European Exascale system (500 M€), 2024



EuroHPC Joint Undertaking

The EuroHPC JU has already procured seven supercomputers:

- 2 Pre-exascale
- 5 Petascale

Total contracts cost: EUR ~360M

EUROPEAN PRE-EXASCALE SYSTEMS

LUMI @ CSC (Finland): HPE/CRAY

- AMD EPYC CPUs, AMD MI250 GPUs
- 2.2 M CPU cores & GPU compute units
- 309 PFlop/s HPL, rank 3 in Top500 11/2022

LEONARDO @ CINECA (Italy): Atos

- Intel Xeon CPUs, Nvidia A100 GPUs
- 1.5 M CPU cores & GPU streaming multiprocessors)
- 175 PFlop/s HPL, rank 4 in Top500 11/2022

APPLICATION DEVELOPMENT

Centres of Excellence (CoEs)

- **Improve applications important to certain domains towards the Exascale**
  - Optimization, new algorithms, methods, etc.
- **Provide training and support**
- **Cover full workflow (including data handling)**

CoE European Excellence in HPC Applications

- **Covering a wide range of scientific domains**

https://www.hpccoe.eu/

FIRST 15 CENTRES OF EXCELLENCE

15 Centres of excellence active up to 2022: created during three calls (2015, 2018 and 2019)

bioexcel
Centre of Excellence for Computational Biomolecular Research

Computational Biomolecular Research

ChEESE

CoE for Exascale in Solid Earth

CompBioMed

Computational methods for biomedical applications

coec
Center of Excellence in Combustion

ecam

CoE of the CECAM community

EoCoE

Energy oriented CoE : toward exascale for energy

FIRST 15 CENTRES OF EXCELLENCE

esiwace

CENTRE OF EXCELLENCE IN SIMULATION OF WEATHER AND CLIMATE IN EUROPE

HiDALGO

HPC and Big Data Technologies for Global Challenges

EXCELLERAT

CoE for Engineering Applications

MAX

Materials design at the exascale

NOMAD

NOVEL MATERIALS DISCOVERY

PerMed CoE

pop

Performance Optimisation and Productivity

RAISE

Center of Excellence

TREX

Targeting Real chemical accuracy at the EXascale

COES STARTED ON JAN 1, 2023

SPACE

Astrophysics and cosmology

MaX

Materials design at the exascale

bioexcel

Centre of Excellence for Computational Biomolecular Research

CEEC

Exascale CFD

ChEESE

Solid Earth

MultiXscale

PLASMA PEPSC

esiwace

CENTRE OF EXCELLENCE IN SIMULATION OF WEATHER AND CLIMATE IN EUROPE

HiDALGO

Global Challenges

EXCELLERAT

Engineering Applications

COE AREAS OF EXPERTISE

CoEs develop strong expertise in their specific application fields as well as more transversal HPC skills needed to achieve science at exascale.

Disciplinary Expertise

- Energy production (wind, hydro, fusion,…)
- Engineering (automotive, aerospace,…)
- Combustion
- Plasma physics
- Material science
- Material for energy (batteries, PV cells,…)
- Chemistry
- Climate sciences and weather forecasts
- Global challenges (health-relevant social habits, green growth, dynamics of global urbanisation.)
- Solid earth physics
- Molecular biology
- Personalized medicine
- Biomedical applications (Cardiovascular Medicine, Neuro Musculoskeletal Medicine,…)
- Astro physics

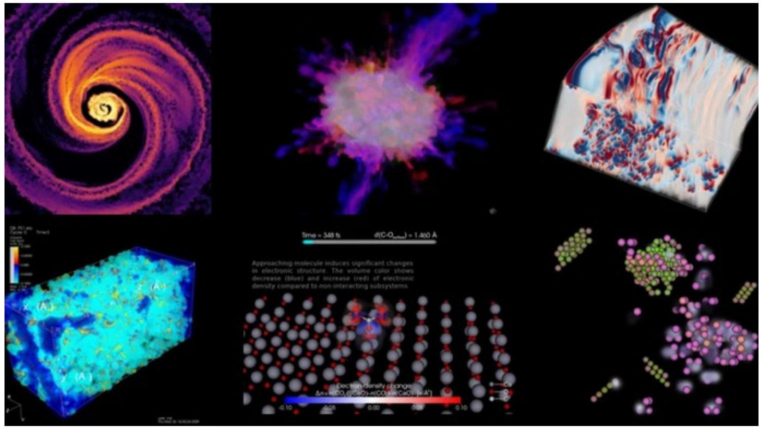
HPC expertise

- Programming models for exascale
- Performance monitoring, optimization and scalability
- Tools for HPDA in complex workflows
- Workflows
- Scalable solvers, linear algebra
- Data flow, in-situ data analysis and I/O
- Ensemble runs
- Implementing co-design and technology integration

CoE European Excellence in HPC Applications

MAX PLANCK COMPUTING AND DATA FACILITY (MPCDF):

A CROSS-INSTITUTIONAL COMPETENCE CENTRE OF THE MAX PLANCK SOCIETY TO SUPPORT COMPUTATIONAL AND DATA SCIENCES



MAX PLANCK
COMPUTING & DATA FACILITY
MPCDF

HIGH-PERFORMANCE COMPUTING AND DATA SCIENCE IN THE MAX PLANCK SOCIETY

CORE SERVICES OF THE MPCDF

MPCDF

HPC

Storage

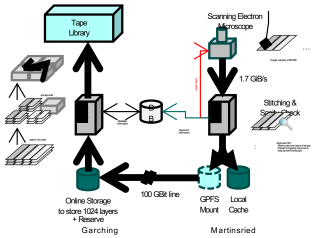
Data Management

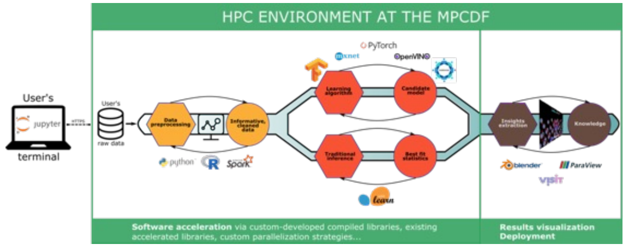
Appli-cation Support

AI

Training

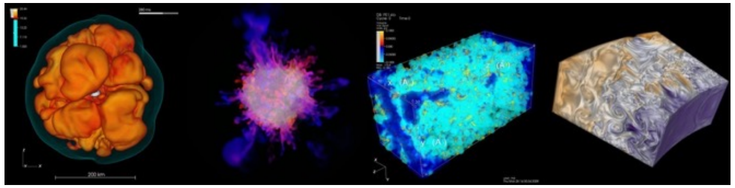
Hosting




HPC ENVIRONMENT AT THE MPCDF

Software acceleration via custom-developed compiled libraries, existing accelerated libraries, custom parallelization strategies...

Results visualization
Deployment



HPC RESOURCES FOR THE MPG

Cobra (2018 –)
~ 12 PetaFlop/s aggregated peak performance

~ 3.400 compute nodes,
~ 137.000 compute cores (Intel Xeon Gold 6148 “Skylake-SP”)
128 NVIDIA Tesla V100 GPUs; 240 Quadro RTX 5000 GPUs
OmniPath 100 Gb/s Interconnect
Rank 96 in June’22 TOP500 list

Raven & Raven GPU (2020/2021 –)
~ 5 PetaFlop/s (CPU), ~15 PetaFlop/s (GPU)

~ 1.600 compute nodes (CPU),
~ 115.000 compute cores (Intel Xeon Platinum 8360Y “Icelake-SP”)
192 GPU nodes with **768 NVIDIA A100 GPUs**
HDR Infiniband Interconnect (100 Gb/s … 400 Gb/s)
Rank 61 and 99 in June’22 TOP500 list

NEW SYSTEM 2023/24 ...

Genoa

MI300

HPC-CLOUD

Cloud Compute

- Xeon CPUs
- Nvidia A30/A100
- 512-4096 GB RAM
- SSD/NVMe

700 Cores
70 GPUs
140 TB RAM

openstack

Block & Object Storage

SSD Pool: 80 TB
HDD Pool: 6 PB

ceph

File System

HDD Pool: 3.5 PB

IBM Spectrum Scale

Nexus

HPC-CLOUD CONCEPT

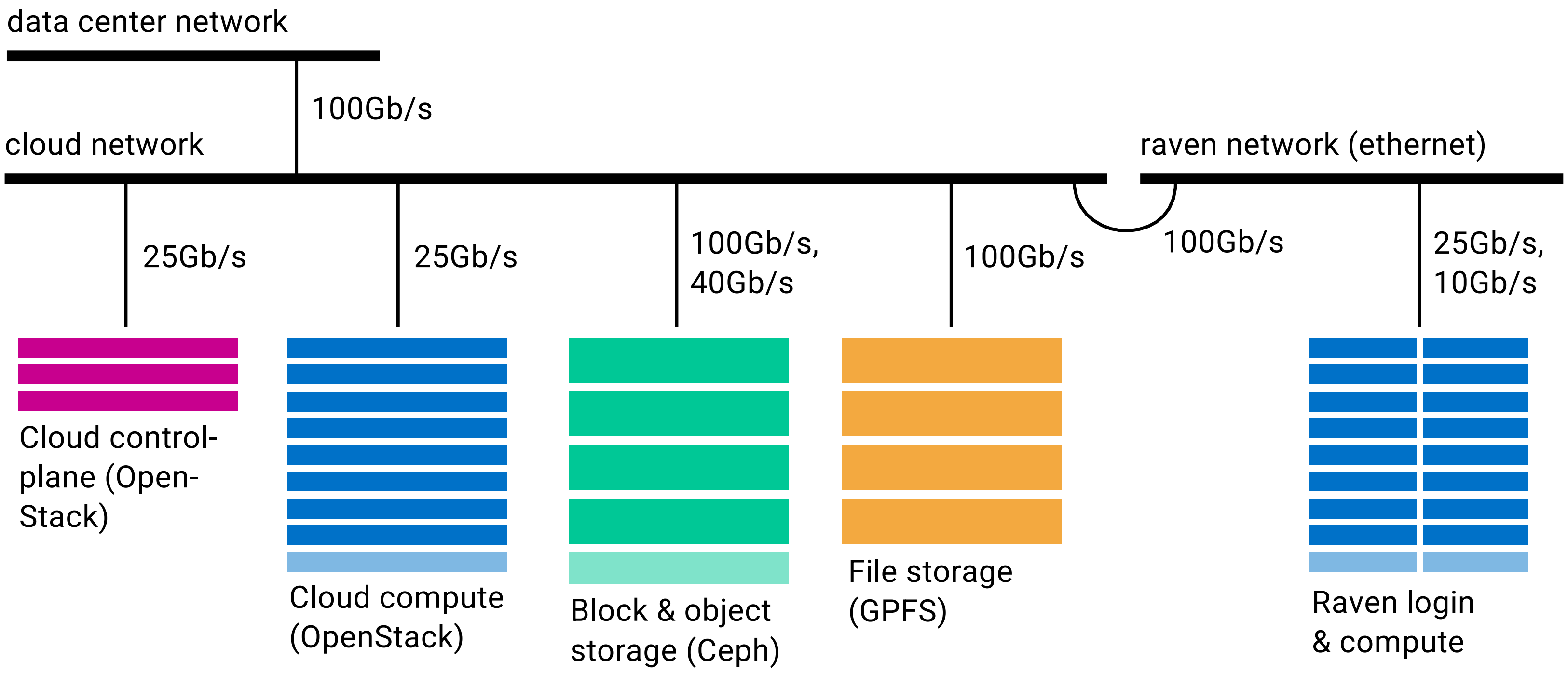
A general solution for complex workflows, complementing the HPC systems

HPC Cloud

- ✓ Logically positioned near super-computing resources, esp. Raven
- ✓ Contains significant compute power
- X Not itself a parallel compute cluster

- ✓ Flexible computing environment
- ✓ Infrastructure-as-a-Service inc. self-service dashboard, standard APIs
- X Not intended for general IT services

NETWORK ARCHITECTURE

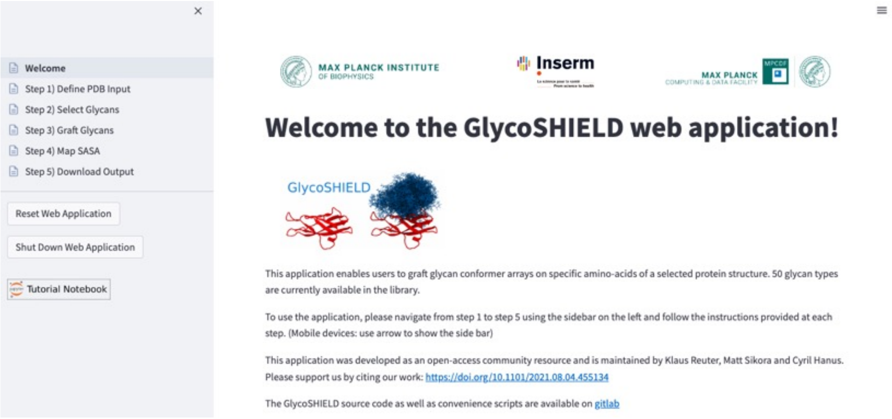


openstack. ceph IBM Spectrum Scale

https://docs.mpcdf.mpg.de/doc/cloud/index.html

HPC CLOUD USECASES

- Online Data Analysis
- Training (BinderHub)
- Data Publication
- Image Processing
- Long-running jobs



Welcome to the GlycoSHIELD web application!

This application enables users to graft glycan conformer arrays on specific amino-acids of a selected protein structure. 50 glycan types are currently available in the library.

To use the application, please navigate from step 1 to step 5 using the sidebar on the left and follow the instructions provided at each step. (Mobile devices: use arrow to show the side bar)

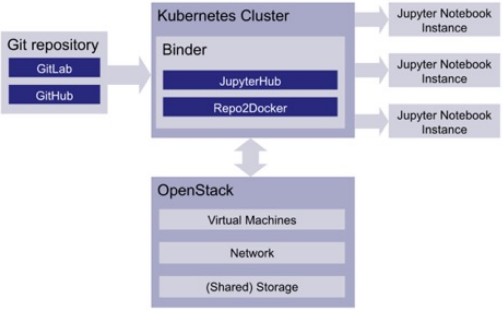
This application was developed as an open-access community resource and is maintained by Klaus Reuter, Matt Sikora and Cyril Hanus. Please support us by citing our work: https://doi.org/10.1101/2021.08.04.455134

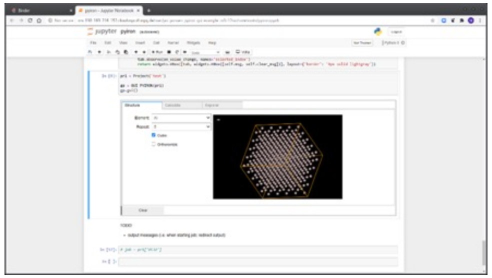
The GlycoSHIELD source code as well as convenience scripts are available on gitlab

Turn a Git repo into a collection of interactive notebooks

Have a repository full of Jupyter notebooks? With Binder, open those notebooks in an executable environment, making your code immediately reproducible by anyone, anywhere.

Build and launch a repository





HPC APPLICATION SUPPORT FOR THE MPG

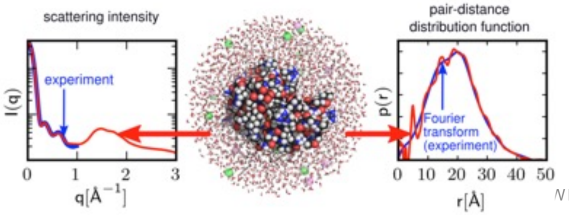
Original contributions and long-term support for

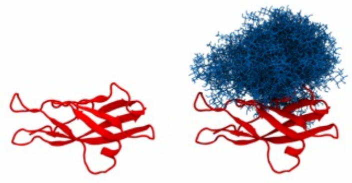
development, *optimization* and *porting* of HPC and AI applications

FHI-aims, OCTOPUS, NECI, DFTB+, ESPResSo++ (*materials science*), ELPA (*eigensolver library*)

GENE, SeLaLib, IDE, TORBEAM, GRILLIX, TurTLE, MagIC, CHIEF (*plasma & astrophysics*)

BioEM, COMPLEXES++, GlycoSHIELD, TriMEM, GROMACS, DIAMOND, ... (*biophysics, bioinformatics*)

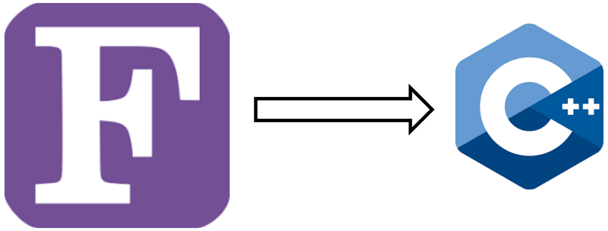




APPLICATION SUPPORT: FOCUS TOPICS

GPU porting and optimization of major production codes developed in the Max Planck Society:

- GENE, GRILLIX/PARALLAX (plasma physics / nuclear fusion research)
- OCTOPUS, FHI-aims (electronic structure theory / DFT, TD-DFT)
- ELPA (eigensolver library)
- DIAMOND (bioinformatics / fast sequence alignment)



Portable programming models and frameworks (and related SW-engineering challenges):

- OpenMP, OpenACC, HIP, SYCL, …, Kokkos

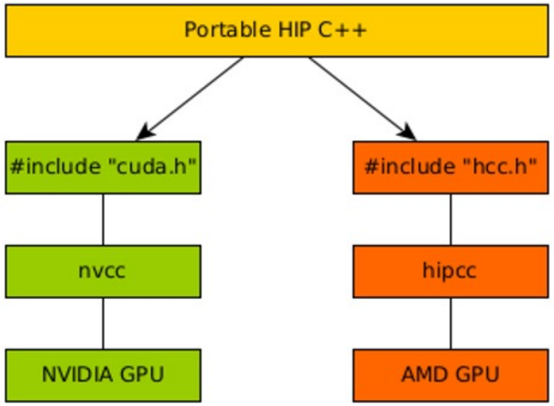
Beyond „traditional“ MPI communication and I/O

- asynchronous and GPU-aware MPI, HPX, in-situ techniques, burst-buffer filesystems

AMD HIP: FROM NVIDIA (CUDA) TO AMD (HIP) AND BACK(?)

On the AMD platform, HIP is the native programming paradigm *and portability layer* for ROCm

- **HIP enables compatibility between Nvidia and AMD GPUs**
- **HIP essentially adopts CUDA semantics**


Portable HIP C++

#include "cuda.h"

#include "hcc.h"

nvcc

hipcc

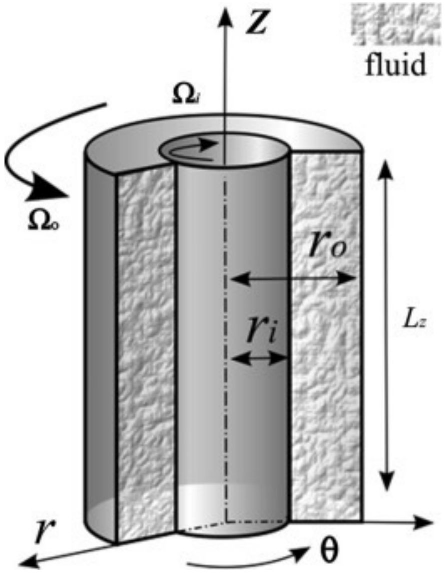
NVIDIA GPU

AMD GPU

CUDA (Nvidia)	**HIP (portable)**	**ROCm (AMD)**
cuBLAS	hipBLAS	rocBLAS
cuFFT	hipFFT	rocFFT
cuRAND	hipRAND	rocRAND
cuSPARSE	hipSPARSE	rocSPARSE
NCCL		RCCL
CUB	hipCUB	rocPRIM

https://www.lumi-supercomputer.eu/preparing-codes-for-lumi-converting-cuda-applications-to-hip/

MY FAVOURITE TEST VEHICLE: NSCOUETTE



- Pseudospectral DNS code in Taylor-Couette geometry:
  - Taylor-Couette devices, pipes, (astrophysical) discs, ...
  - FFTs, global transposes, linear systems
- Open Source:
  - https://github.com/dfeldmann/nsCouette
  - https://gitlab.mpcdf.mpg.de/mjr/nscouette
  - Jose M. Lopez, Daniel Feldmann, Markus Rampp, Alberto Vela-Martin, Liang Shi & Marc Avila, nsCouette - A high-performance code for direct numerical simulations of turbulent Taylor-Couette flow, SoftwareX, 11, 100395 (2020), preprint: arXiv:1908.00587v3
- Scalable MPI-OpenMP Fortran version for CPUs: *main workhorse*
- Single-GPU CUDA version: *development* by A. Vela-Martin

BENCHMARKS WITH NSCOUETTE-GPU (GRID: 128,512,257)

Platform	GPU	FP64 peak [Tflop/s] *(ratio)*	BW peak [GB/s] *(ratio)*	Runtime CUDA code [ms/step] *(ratio)*	Runtime HIP code [ms/step] *(ratio)*
Nvidia	A100	9.7 *(1.00)*	1555 *(1.00)*	228.3 *(1.00)*	228.2 *(1.00)*
Nvidia	V100	7.0	900 *(0.58)*	388.5 *(0.59)*	(not tested)
AMD	MI210	22.6	1638 *(1.05)*	N/A	296.5 *(0.77)*

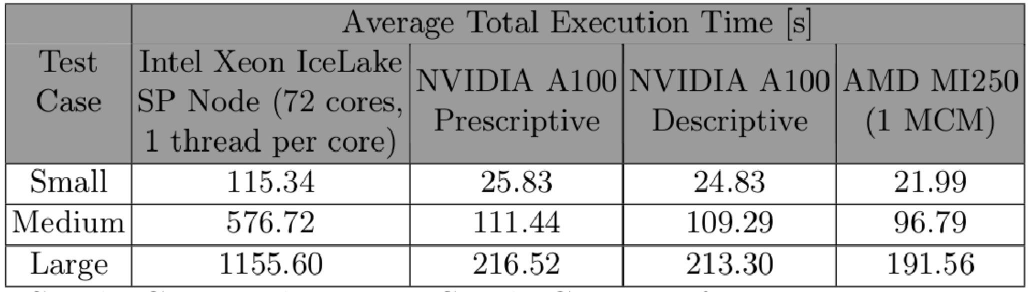
Main observations:

- **smooth and automatic conversion CUDA->HIP with hippify tool**
- **software: zero HIP overhead on A100 (no surprise: hipcc -> nvcc, hipBLAS -> cuBLAS, ...)**
- **AMD hardware/software: relative underperformance on MI210 *(0.77)***
- **AMD profiling tools are maturing**

OPENMP FOR GPUS: BS-SOLCTRA

BS-SOLCTRA: Biot-Savart Solver for Computing and Tracing Magnetic Field Lines

- C++ code developed by CeNAT (Costa-Rica) for the Stellarator of Costa Rica 1 (SCR1)
- trivially parallel CPU code (OpenMP + MPI) over particles
- deep call stack down to hotspot computations



Test Case	Average Total Execution Time [s]			
	Intel Xeon IceLake SP Node (72 cores, 1 thread per core)	NVIDIA A100 Prescriptive	NVIDIA A100 Descriptive	AMD MI250 (1 MCM)
Small	115.34	25.83	24.83	21.99
Medium	576.72	111.44	109.29	96.79
Large	1155.60	216.52	213.30	191.56

```
#pragma omp target enter data map (to:...)
for(int i=0;i<steps;i++){
#pragma omp target teams distribute parallel for
 for(int p=0; p<particle_count;p++)
   computeIteration(...)
 }
}
#pragma omp target exit data map (from:...)
```

+ demonstrated portability to Intel Xe GPU (PonteVeccio preview)

- single-source not achieved here (data structures)

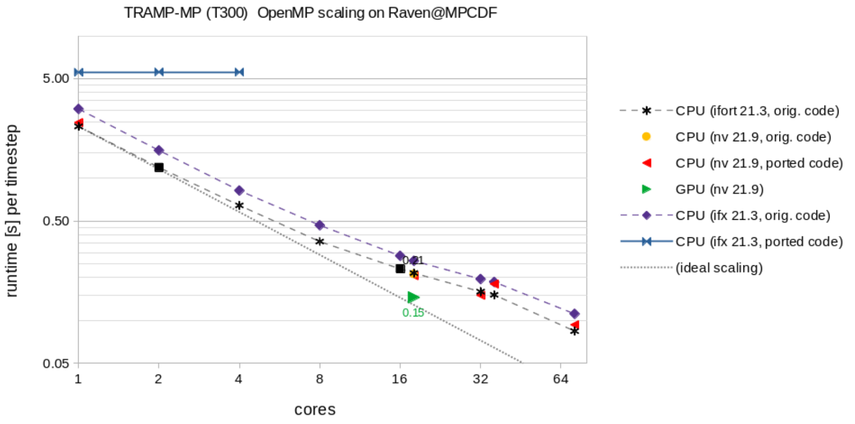
Implementing a GPU-Portable Field Line Tracing Application with OpenMP Offload, D. Jimenez, J. Herrera-Mora, M. Rampp, E. Laure, E. Menses, CARLA (2022)

OPENMP FOR GPUS: TRAMP

TRAMP: 3D-radiation-hydrodynamics (FLD) code by MPI for Astronomy

- classic (F77 / OpenMP) legacy CPU code with 3-nested loop structure

```
!$OMP TARGET TEAMS LOOP COLLAPSE(2) REDUCTION(+:RHSNORM)
      DO K=KMIN,NZ
        DO J=1,NY
!$OMP LOOP REDUCTION(+:RHSNORM)
          DO I=IMIN,NX
```



- single-source (CPU-GPU) with no performance overhead

  (baseline: original OpenMP CPU code with ifort)
- speedup on GPU within expectations of ported code parts
- further porting mostly hampered by **tedious data locality handling**

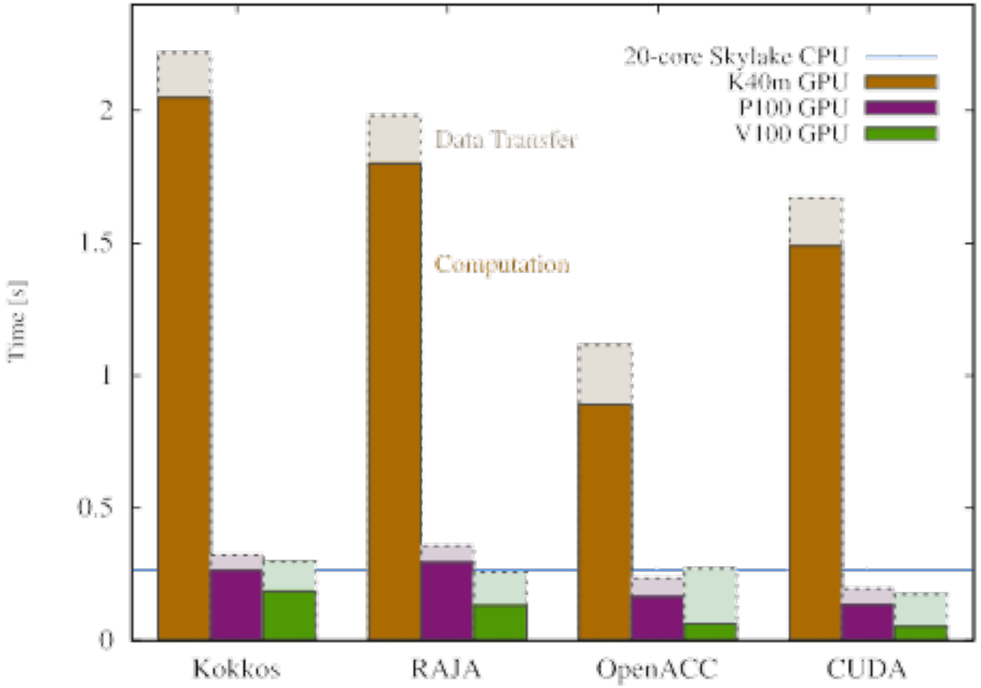
  (would greatly benefit from fast unified memory -> MI300, GRACE-HOPPER)

KOKKOS: PORTABILITY STUDY & NEW VLASOV 6D CODE

- usability and performance study of high-level frameworks Kokkos and RAJA on GPUs and CPUs (NMPP & MPCDF, 2018)
- proxy application: non-trivial PIC kernel written in C++ (SeLaLib)
- Kokkos (and RAJA) appear mature and usable for our complex proxy application

=> **Led to the development (from scratch) of a semi-lagrangian Vlasov Code in 6D with Kokkos:**

N. Schild, M. Raeth, S. Eibl, K. Hallatschek, K. Kormann. A performance portable implementation of the semi-Lagrangian algorithm in six dimensions. arXiv: 2303.05994



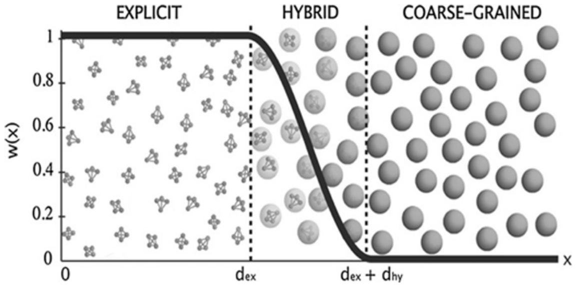
V. Artigues, K. Kormann, M. Rampp, K. Reuter. Evaluation of performance portability frameworks for the implementation of a particle-in-cell code. Concurrency and Computation: Practice and Experience 32:e5640 (2020). arXiv:1911.08394

KOKKOS: MOLECULAR DYNAMICS

In collaboration with Ch. Junghans (LANL)

ESPRESSO++ (a classical molecular-dynamics package by MPI for Polymer Research)

- **developed an adaptive-resolution MD code (H-AdResS) from scratch using Kokkos**
- **leveraging data-structures from Cabana (ECP), collaboration with LANL**
- **code is production-ready, promising GPU benchmarks:**



Binary Lennard-Jones Mixture with 36000 particles

	CPU (MPUpS) 18 IceLake Cores	GPU (MPUpS) 1 A100	speedup (CPU -> GPU)
NPT	6.5	15.9	2.4x
NVT	9.1	40.7	4.5x
SPARTIAN	3.9	23.0	5.9x

+ tested on AMD MI210 GPU

KOKKOS: GPU PORT OF A MACHINE-LEARNING APPLICATION

SISSO++ (a machine-learning code based on compressed sensing / symbolic regression by the Fritz-Haber Institute of the MPG / NOMAD CoE)

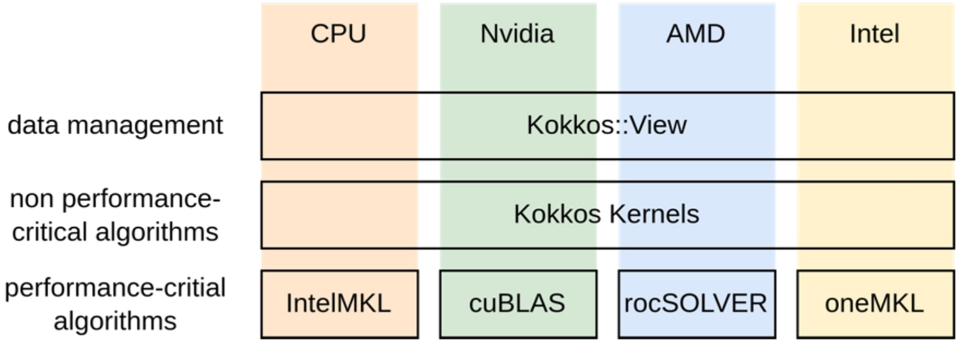
Kokkos porting based on:

- Kokkos views
- Kokkos lambdas

Hotspot: “batched” linear algebra (DGELS)

- hard to achieve with a general abstraction

  → rely on vendor specific libraries
- provide nice encapsulation
- use batched BLAS to reduce launch overhead
- developed custom solver, outperforms cuBLAS



	CPU	Nvidia	AMD	Intel
data management	Kokkos::View			
non performance-critical algorithms	Kokkos Kernels			
performance-critial algorithms	IntelMKL	cuBLAS	rocSOLVER	oneMKL

https://gitlab.mpcdf.mpg.de/nomad-lab/cpp_sisso

MY OWN VIEW: *WHAT IF I HAD TO …*

… develop a new (GPU) code ?

use C++ (+ Kokkos, …)

… port an existing CPU Fortran code to GPU ?

use OpenMP

… port an existing CUDA code to multiple GPU platforms ?

hippify for AMD GPUs (and see what happens with other GPU vendors)

… develop or port a CPU code or library with highest ambition for performance and longevity ?

use vendor-specific language (CUDA, HIP, SYCL) encapsulated by software-abstration layer

watch out for consolidation opportunities (the SYCL promise, OpenMP, …)