Barcelona Supercomputing Center
Centro Nacional de Supercomputación

HIGH-PERFORMANCE ELECTRONIC-STRUCTURE CALCULATIONS IN THE EXASCALE ERA

Julio Gutiérrez Moreno

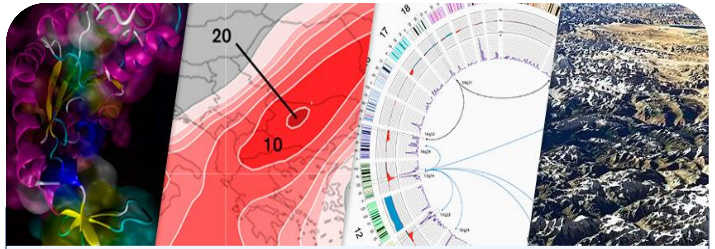
julio.gutierrez@bsc.es

05/08/2023

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

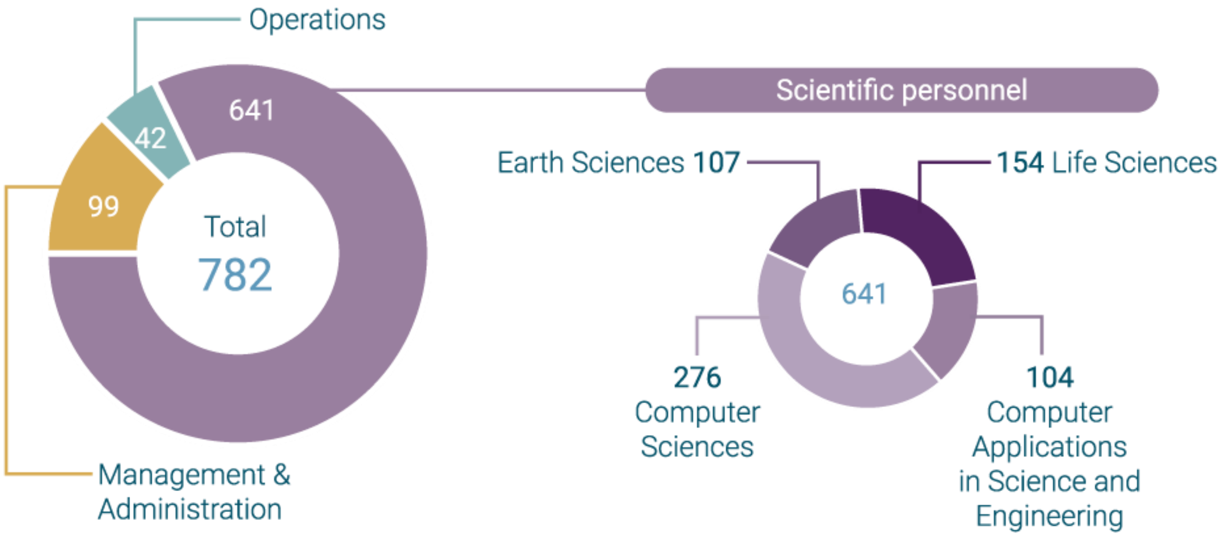
BSC-CNS objectives

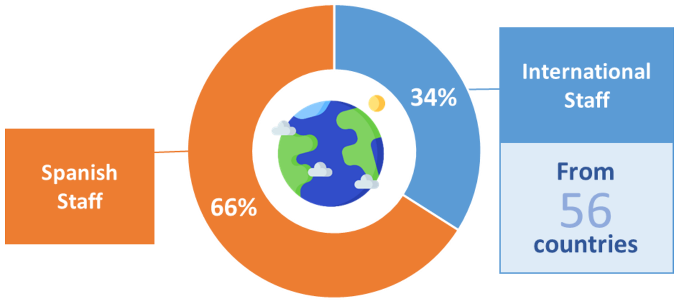
Supercomputing services to Spanish and EU researchers



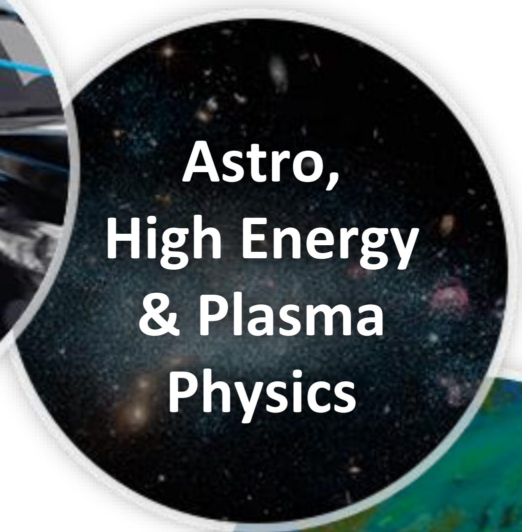
R&D in Computer, Life, Earth and Engineering Sciences

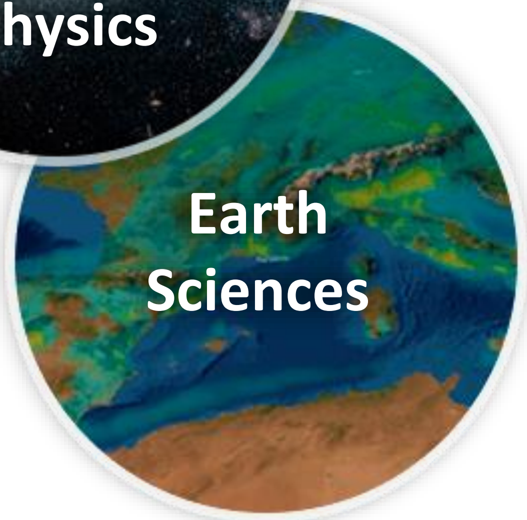
PhD programme, technology transfer, public engagement





HPC: An enabler for all scientific fields





Advances leading to:

- Improved Healthcare
- Better Climate Forecasting
- Superior Materials
- More Competitive Industry

MareNostrum 4

Total peak performance: **13,9** Pflops	
General Purpose Cluster:	11.15 Pflops
MN4 CTE-Power:	1.57 Pflops
MN4 CTE-ARM:	0.65 Pflops
MN4 CTE-AMD:	0.52 Pflops

PRACE

Access: prace-ri.eu/hpc_acces

RES
RED ESPAÑOLA DE SUPERCOMPUTACIÓN

Access: bsc.es/res-intranet

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

MareNostrum 1
2004 – 42,3 Tflops
1st Europe / 4th World

MareNostrum 2
2006 – 94,2 Tflops
1st Europe / 5th World

MareNostrum 3
2012 – 1,1 Pflops
12th Europe / 36th World

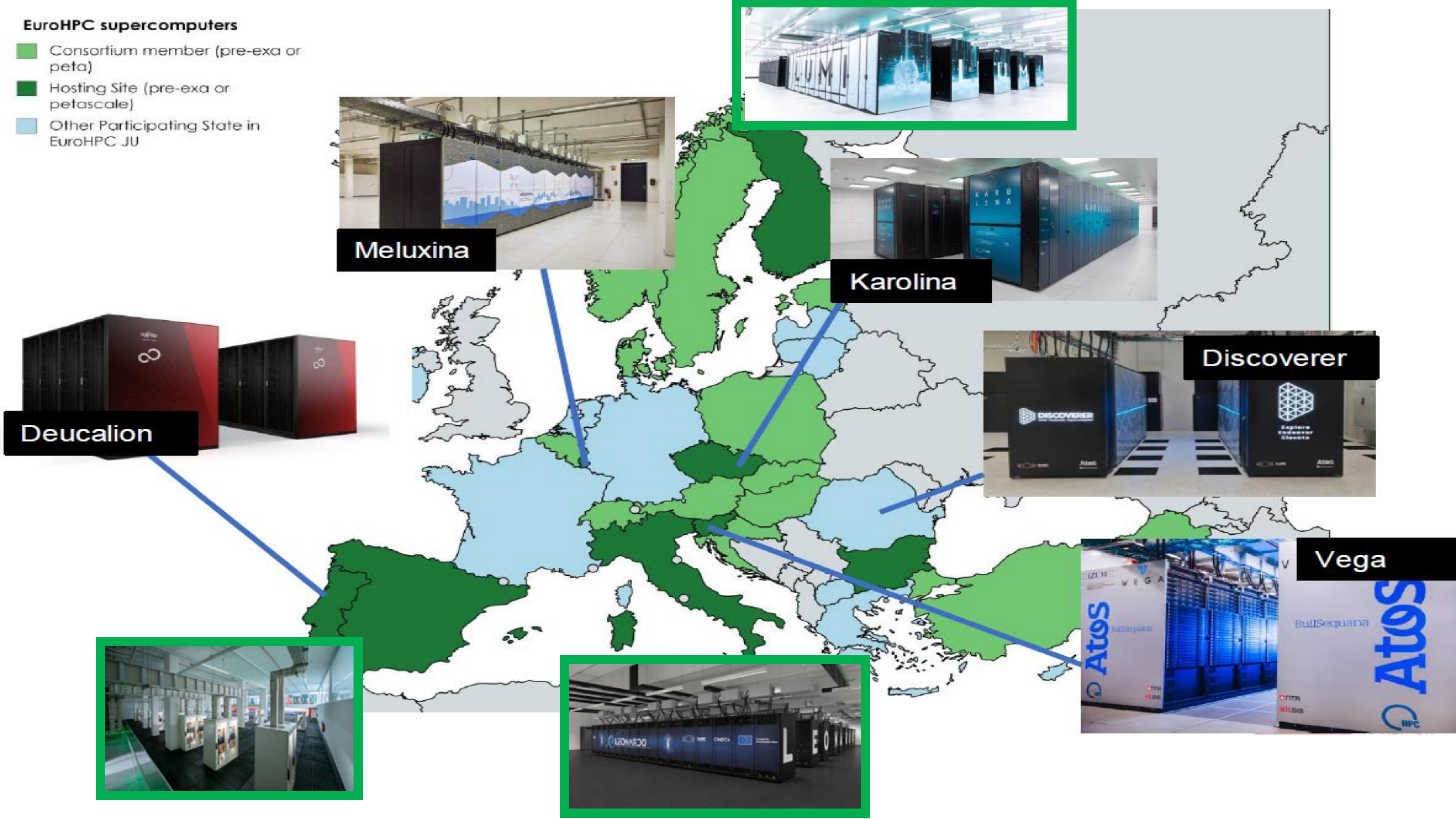
MareNostrum 4
2017 – 11,1 Pflops
2nd Europe / 13th World

Exascale is (almost) here

- Frontier @ Oak Ridge National Lab
- 1.102.000.000.000.000.000 FLOP/S
- 8.730.112 cores (8.138.240 accelerated)
- 21,1 MW

EuroHPC pre-exascale supercomputers :
LUMI, LEONARDO and MARENOSTRUM 5 …

… and **petascale**: VEGA, MELUXINA, KAROLINA, DISCOVERER and DEUCALION


EuroHPC supercomputers
Consortium member (pre-exa or peta)
Hosting Site (pre-exa or petascale)
Other Participating State in EuroHPC JU
Meluxina
Karolina
Discoverer
Deucalion
Vega
LUMI
LEONARDO

MareNostrum 5: European pre-exascale HPC

- **314 Petaflops** peak performance (314 x 10^{15})
- Will facilitate world-changing scientific breakthroughs like the creation of digital twins and the advancement of precision medicine
- Total investment: **>200 M€**

Hosting Consortium:

Spain **Portugal** **Turkey**

European Commission

Generalitat de Catalunya
Departament de Recerca i Universitats

GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

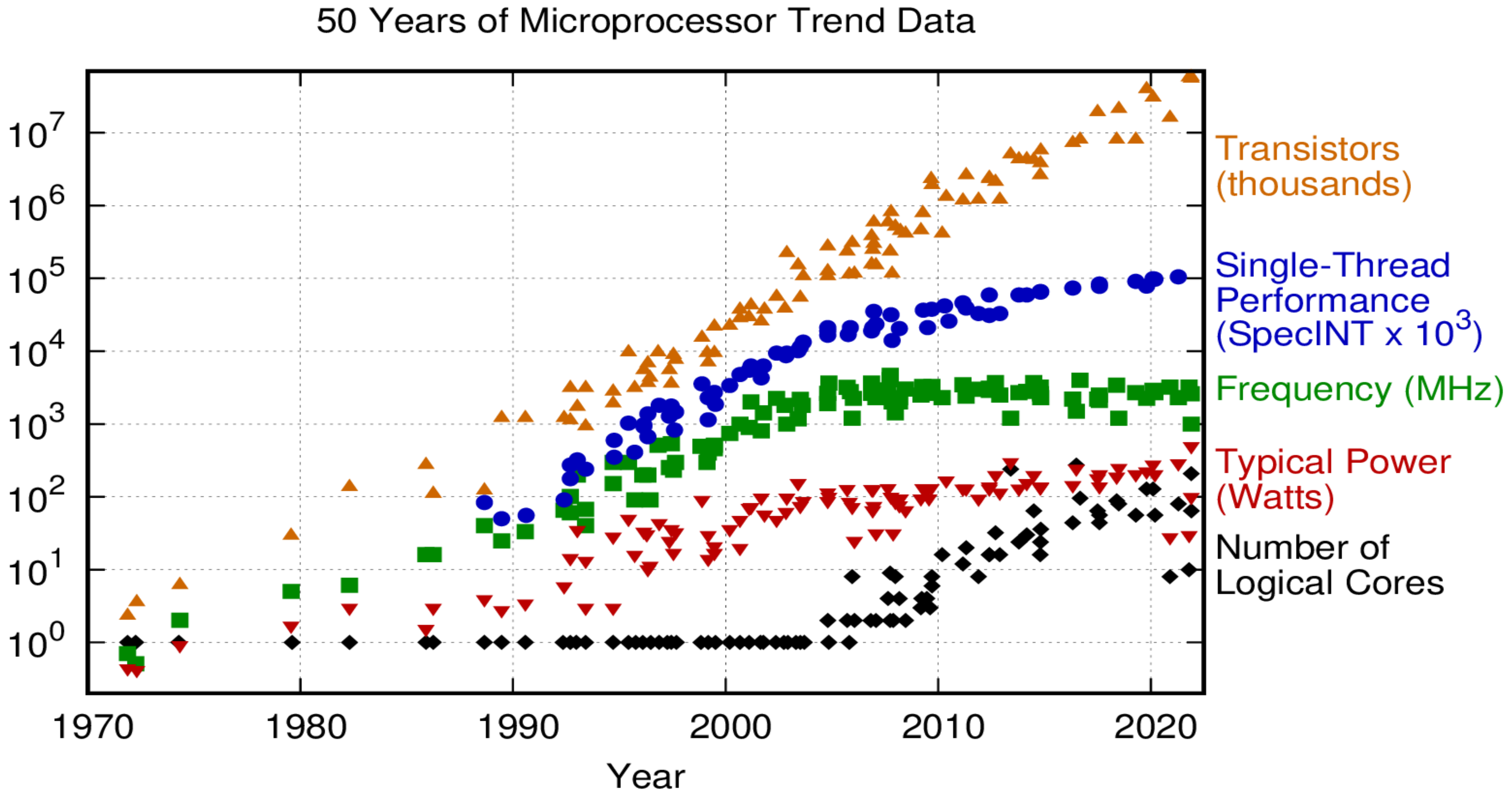
UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

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 Turkey

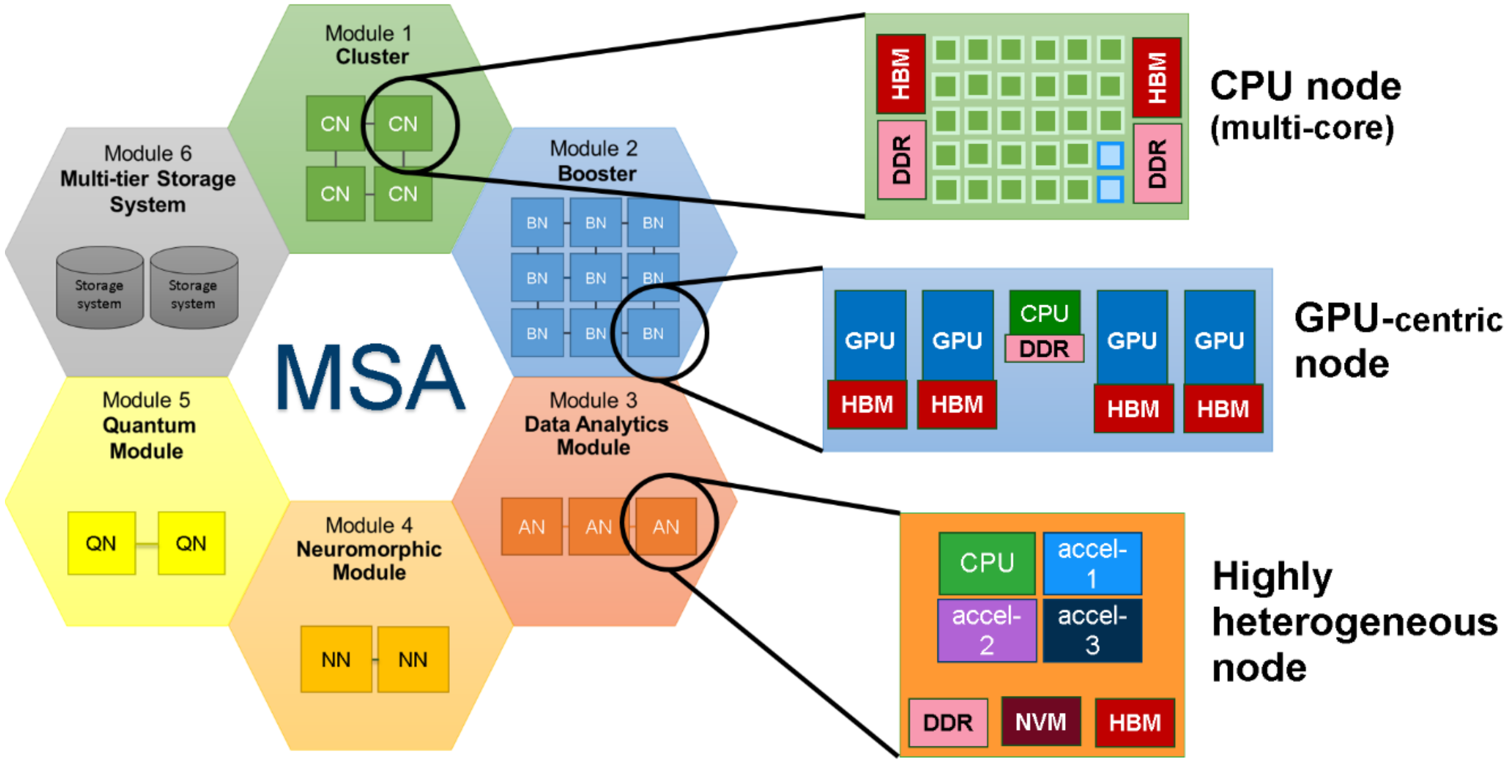
EuroHPC Joint Undertaking

The end of Dennard's & Moore's scaling?



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2021 by K. Rupp

JUPITER: The Arrival of Exascale in Europe



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

DFT & beyond: The NOMAD CoE

CHALLENGES IN MATERIALS SIMULATIONS

So far, Density Functional Theory has been the workhorse of ab initio computational materials

Many materials and/or properties require methods better than DFT

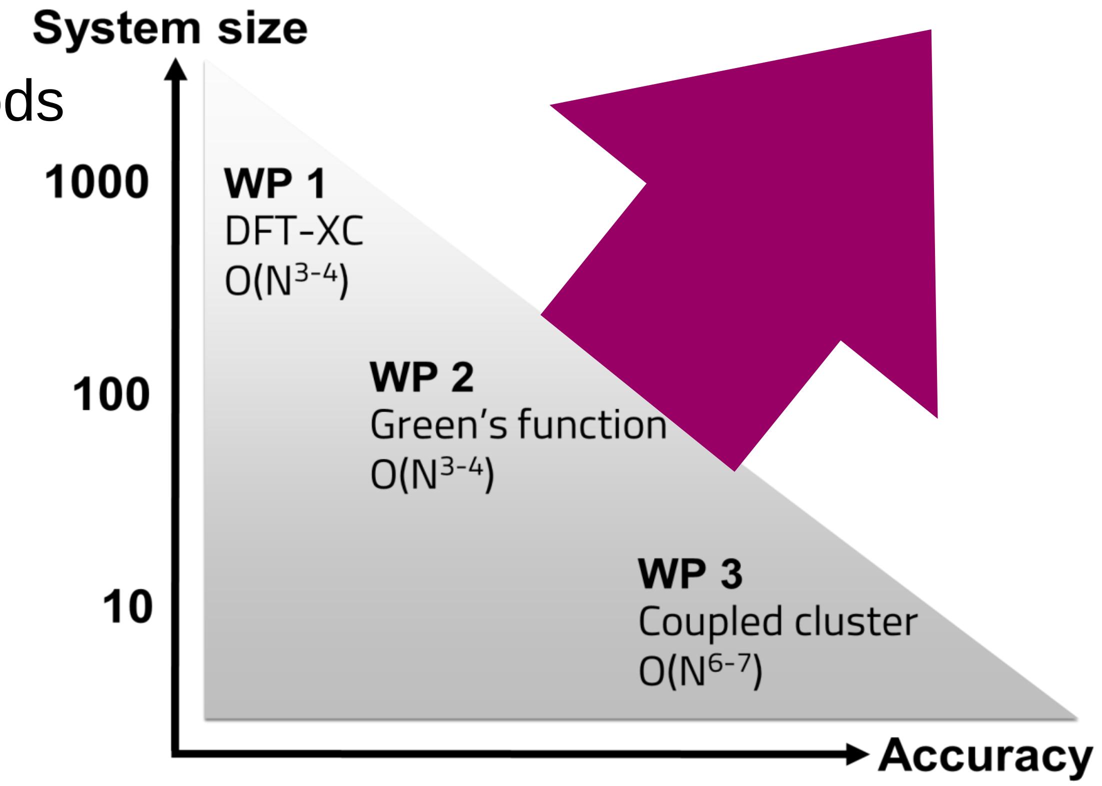
- Energy research, optics...

Complex systems require larger simulation cells

… and/or better methodologies

- Green´s function (GW), Coupled cluster theory
- Larger computational cost than DFT

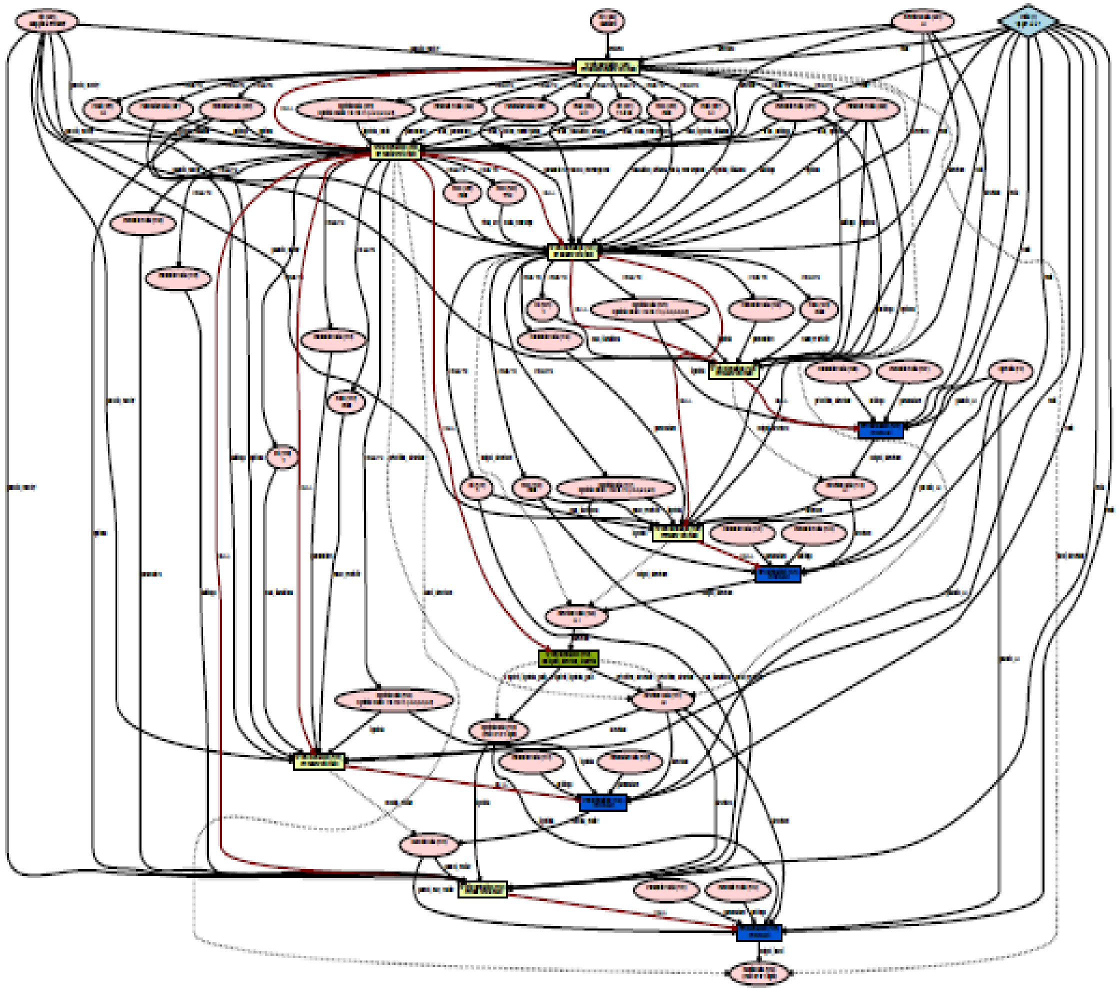
Exascale is required



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Source: Matthias Scheffler (FHI) NOMAD-CoE

Workflows & extreme data



Materials science workflow

- Multi-tiered / multi-scale interfaces
- Large-scale calculation

- Modular codes
- Interoperability traits
- Resilience & provenance

- High throughput computations

Goal of exascale (some numbers)

Exascale-enabled codes:

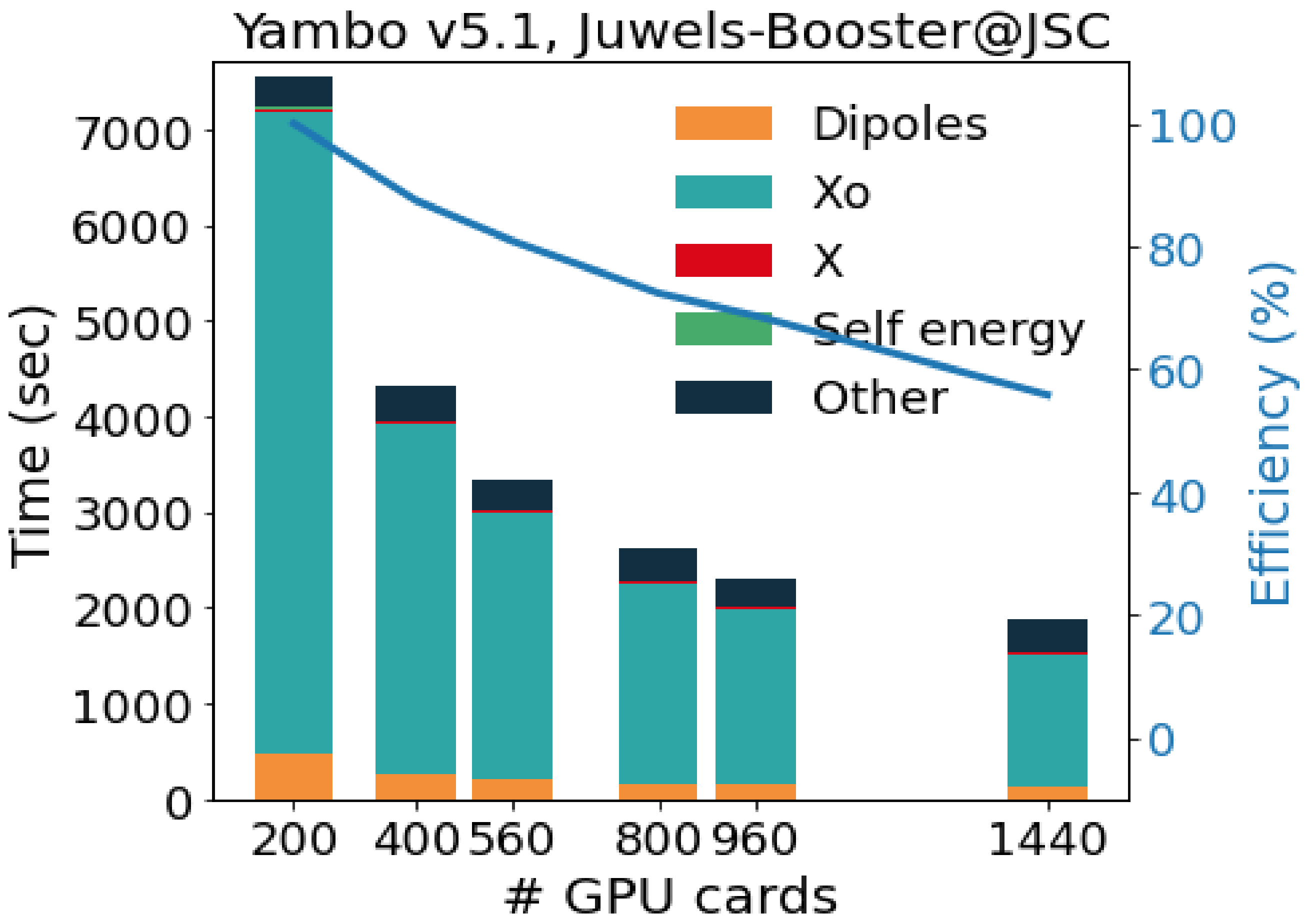
large scale MPI parallelism (~10K tasks)

+ **GPU**

<u>Accelerators for exascale:</u>

- NVIDIA A100 / H100
- AMD Instinct MI200 / MI300

1 card ~25TFLOPS
Goal ~10K cards!



Quasi-particle corrections on 4-layer GrCo. The test involves the evaluation of a response function, of the HF self-energy and of the correlation part.

Source: www.max-centre.eu/codes-max/yambo

Code optimization: SIESTA

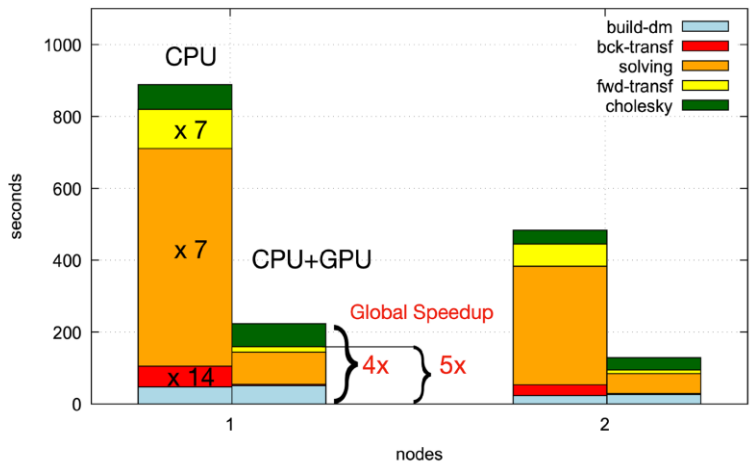
Identify expensive section(s):
solver typically takes >90% the CPU time

- Use of high-performing libraries
- Portable to (pre)-exascale

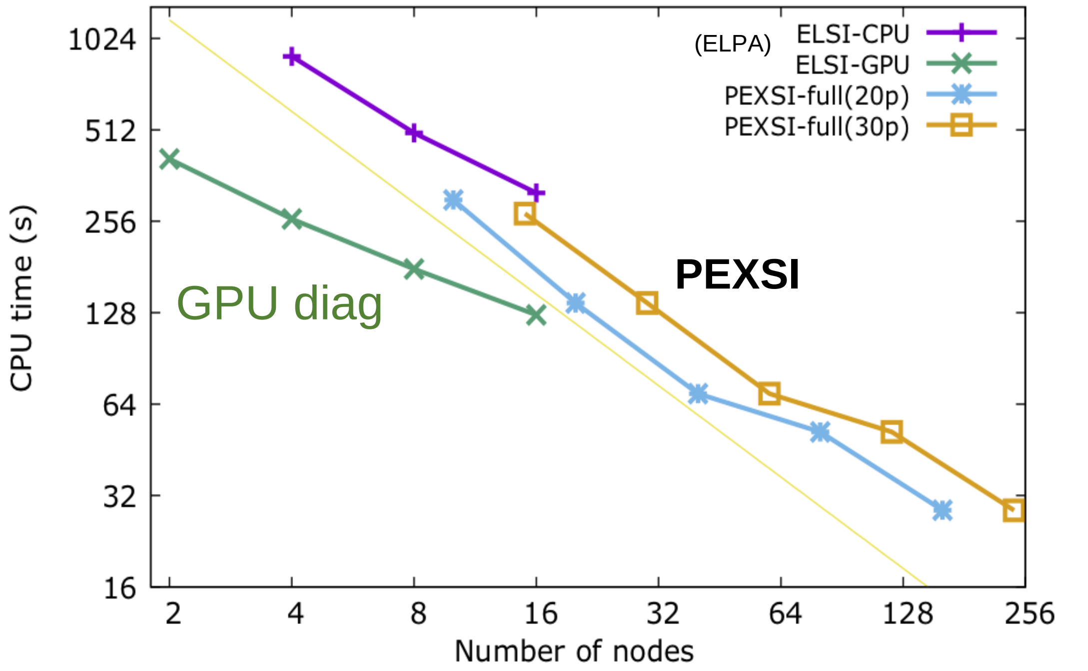
ELSI library: ELPA, PEXSI…

Support new architectures: AMD GPUs

GPU acceleration with ELSI-ELPA in Marconi-100



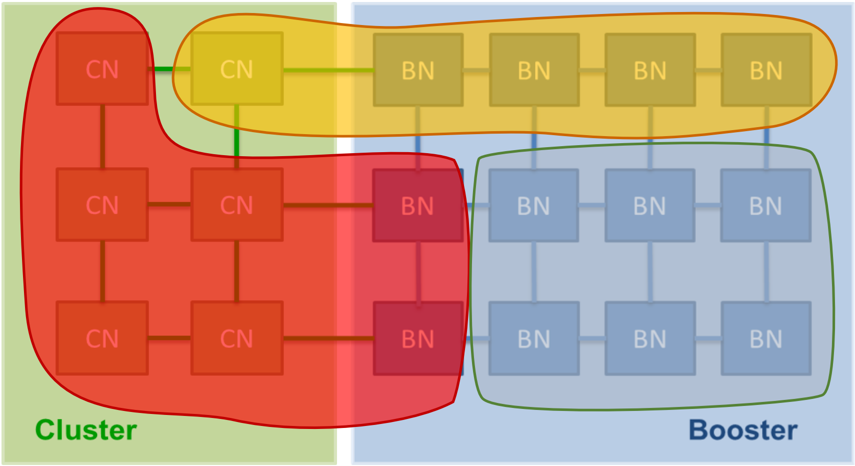
Performance and scalability for sars-cov-2 protein (8800 atoms)



Source: Alberto García (CSIC-ICMAB) MaX-CoE

(Some) Challenges at the Exascale

- Abrupt technology changes
- Hardware heterogeneity
  - CPUs, GPUs, APUs…
  - HBM, SSD…
- Level of parallelism
  - O(10^{18}) flops/s, Bytes



N. Eicker et al. Concurrency and comp. 28 (2016) 2394

- Novel numerical approaches
  - Low scaling algorithms, highly scalable computations, mixed precision
- Modularity
  - Single simulation → integrated workflows
  - Simulation + AI + Data Analytics + …

- Resilience
  - Technical support for codes on exascale systems
  - Deployment and tuning codes and workflows
- Programming – Performance Portability
  - MPI, OpenMP, CUDA, OpenACC, HIP, OneAPI …

The Babel Tower of Programming Languages

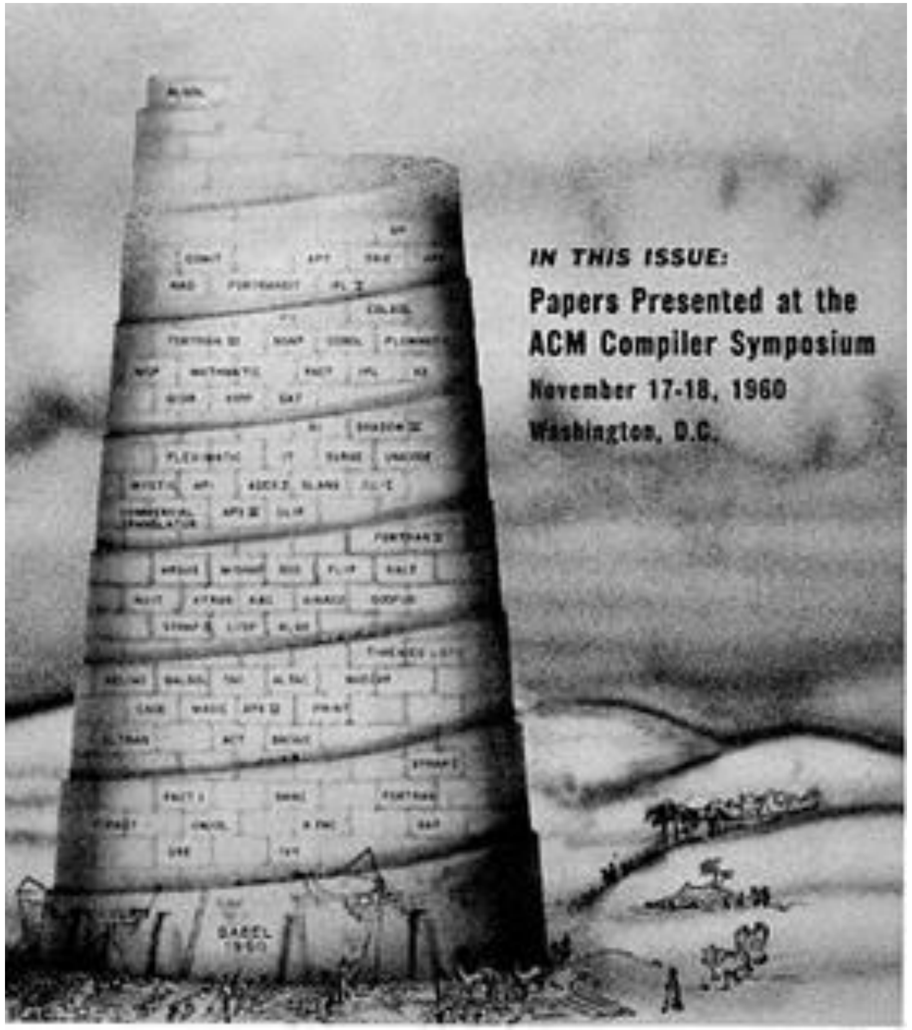
In the absence of suitable libraries… which **programming model**?

https://x-dev.pages.jsc.fz-juelich.de/2022/11/02/gpu-vendor-model-compat.html

- ● Full vendor support
- ◡ Indirect, but comprehensive support, by vendor
- ■ Vendor support, but not (yet) entirely comprehensive
- ▲ Comprehensive support, but not by vendor
- ★ Limited, probably indirect support -- but at least some
- ⟋ No direct support available, but of course one could ISO-C-bind your way through it or directly link the libraries
- C C++ (sometimes also C)
- F Fortran

	CUDA		HIP		SYCL		OpenACC		OpenMP		Standard		Kokkos		ALPAKA		*etc*
	C	F	C	F	C	F	C	F	C	F	C	F	C	F	C	F	Python
NVIDIA	● [1]	● [2]	◡ [3]	⟋ [4]	▲ [5]	⟋ [6]	● [7]	● [8]	■ ● [9]	■ ● [10]	● [11]	● [12]	▲ [13]	★ [14]	▲ [15]	⟋ [16]	■ ▲ [17]
AMD	◡ [18]	★ [19]	● [20]	⟋ [4]	◡ [21]	⟋ [6]	▲ [22]	▲ ★ [23]	● [24]	● [24]	★ [25]	⟋ [26]	▲ [27]	★ [14]	▲ [28]	⟋ [16]	★ [29]
Intel	◡ [30]	⟋ [31]	★ [32]	⟋ [33]	● [34]	⟋ [6]	★ [35]	★ [35]	● [36]	● [36]	■ [37]	■ [38]	▲ [39]	★ [14]	▲ [40]	⟋ [16]	★ [41]

- Vendor-locked & architecture-dependent models!
- Data movements? CPU-GPU comms?
- New architectures? (e.g. NVIDIA Grace-Hopper)

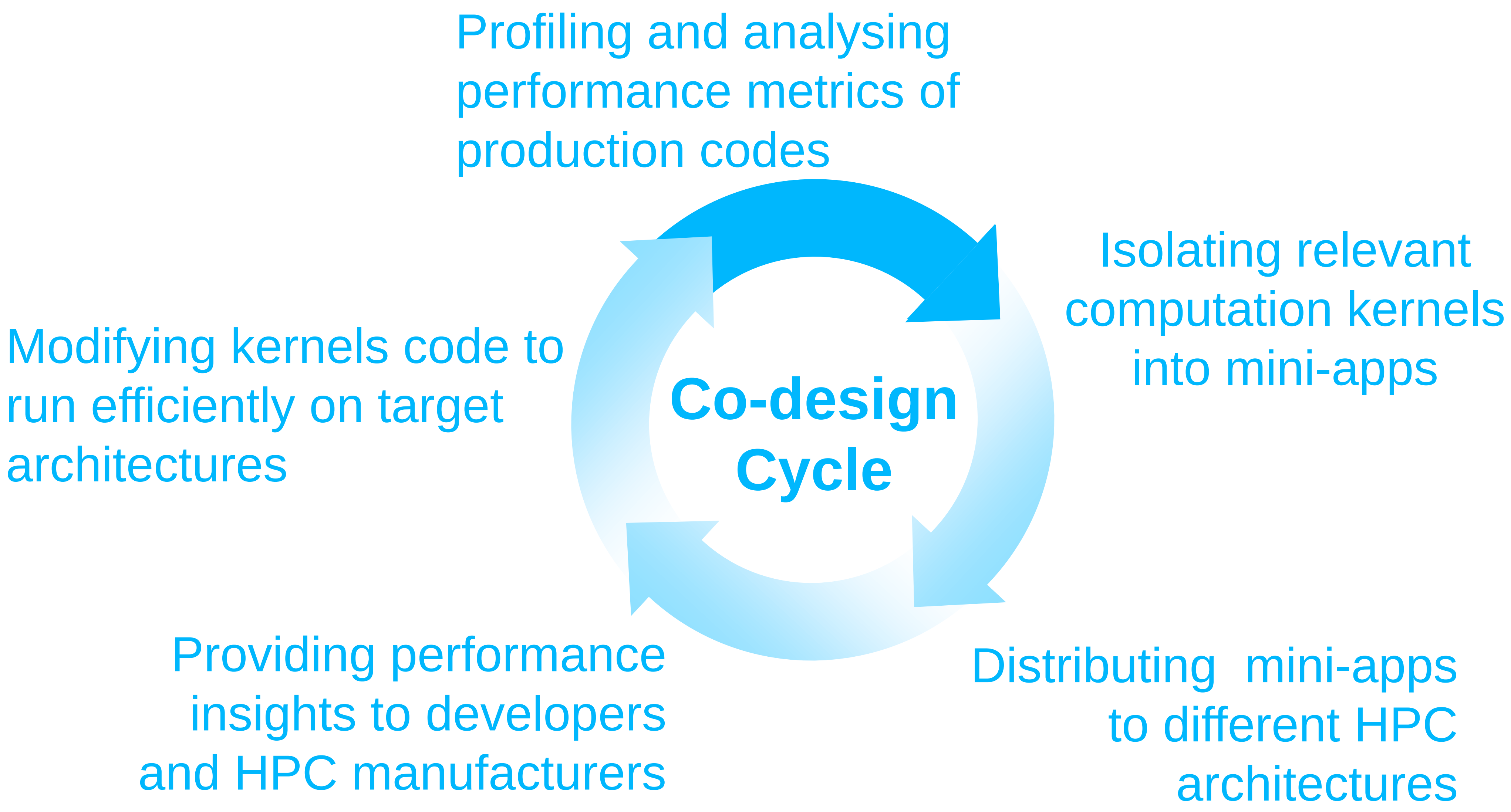


Co-design: developer objectives

- **Prepare codes for (post)exascale and modular HPC computing**
  - Processor level (CPU & GPU), memory characteristics (heterogeneity, bandwidth), vector length…
  - Node level: # sockets/accelerators per node
- **Provide technology developers with realistic data**
  - European R&D projects (EUPEX, EUPilot): Co-design is becoming more accessible
  - Developers of system software (DEEP), processors & platforms (EPI)
- **Get ready for advance hardware platforms**
  - Arm (Nvidia, SiPearl RHEA), EU RISC-V ecosystem (co-design at the Instruction Set Architecture level)
  - New accelerators, neuromorphic, quantum…

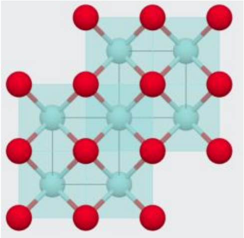
Co-design: applications for materials science

- **Materials science is a very strong and relevant use-case**
- **Ab initio materials science codes should influence system design/procurements**
- **Influencing hardware design is difficult, especially for the HPC community**
  - Examples of good practice: EPI, RISC-V, Fugaku
- **Potential as a co-design vehicle for next-gen hardware**
  - Market is more malleable
  - Interaction with EU developed exascale prototypes
  - Prepare codes for future hardware

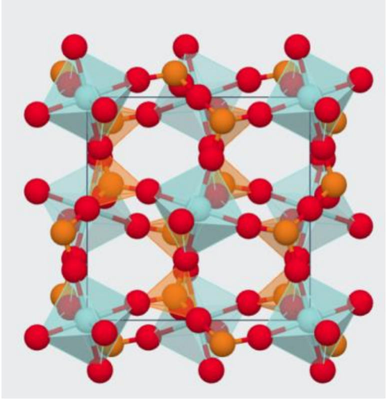

Profiling and analysing performance metrics of production codes
Isolating relevant computation kernels into mini-apps
Distributing mini-apps to different HPC architectures
Providing performance insights to developers and HPC manufacturers
Modifying kernels code to run efficiently on target architectures
Co-design Cycle

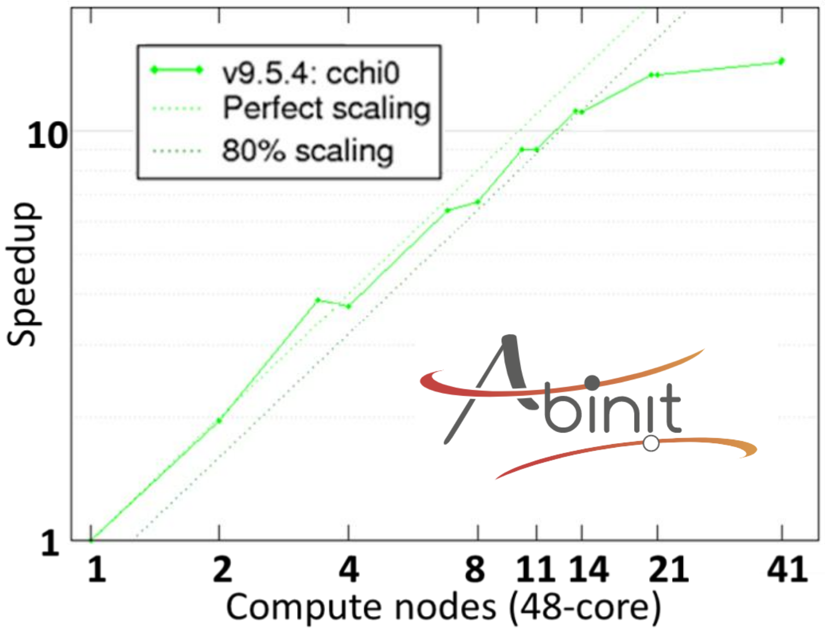
Performance analysis

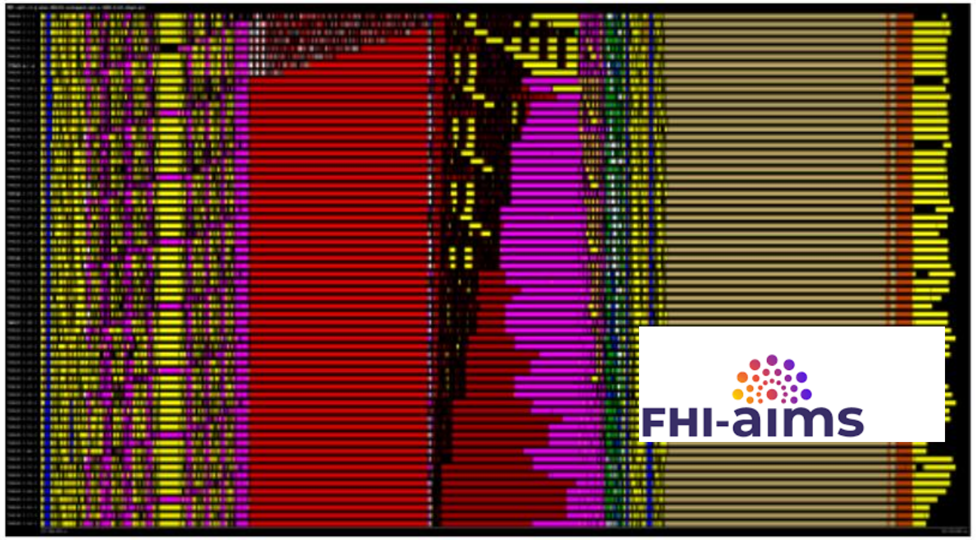
3-atom FCC ZrO_2

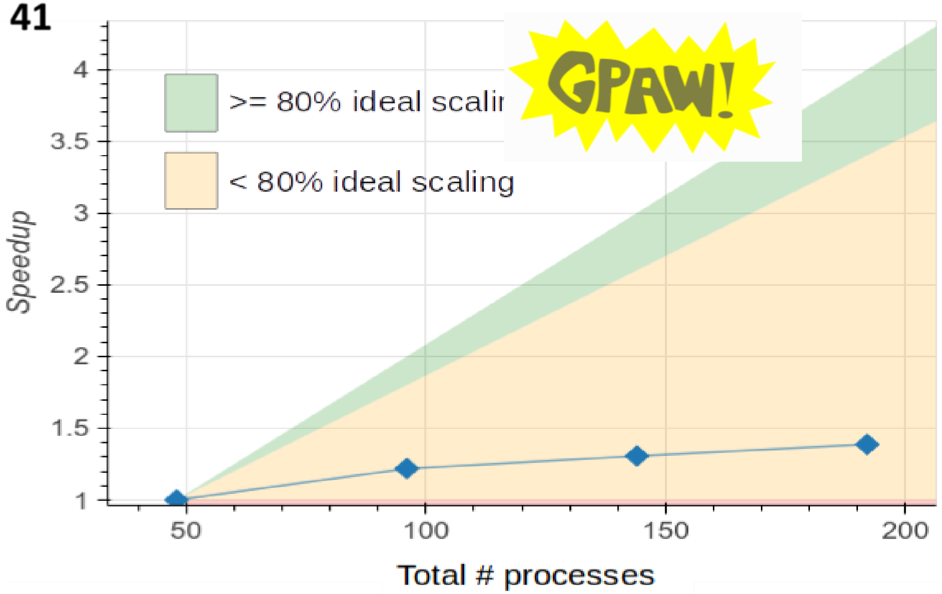


11-atom $Zr_2Y_2O_7$





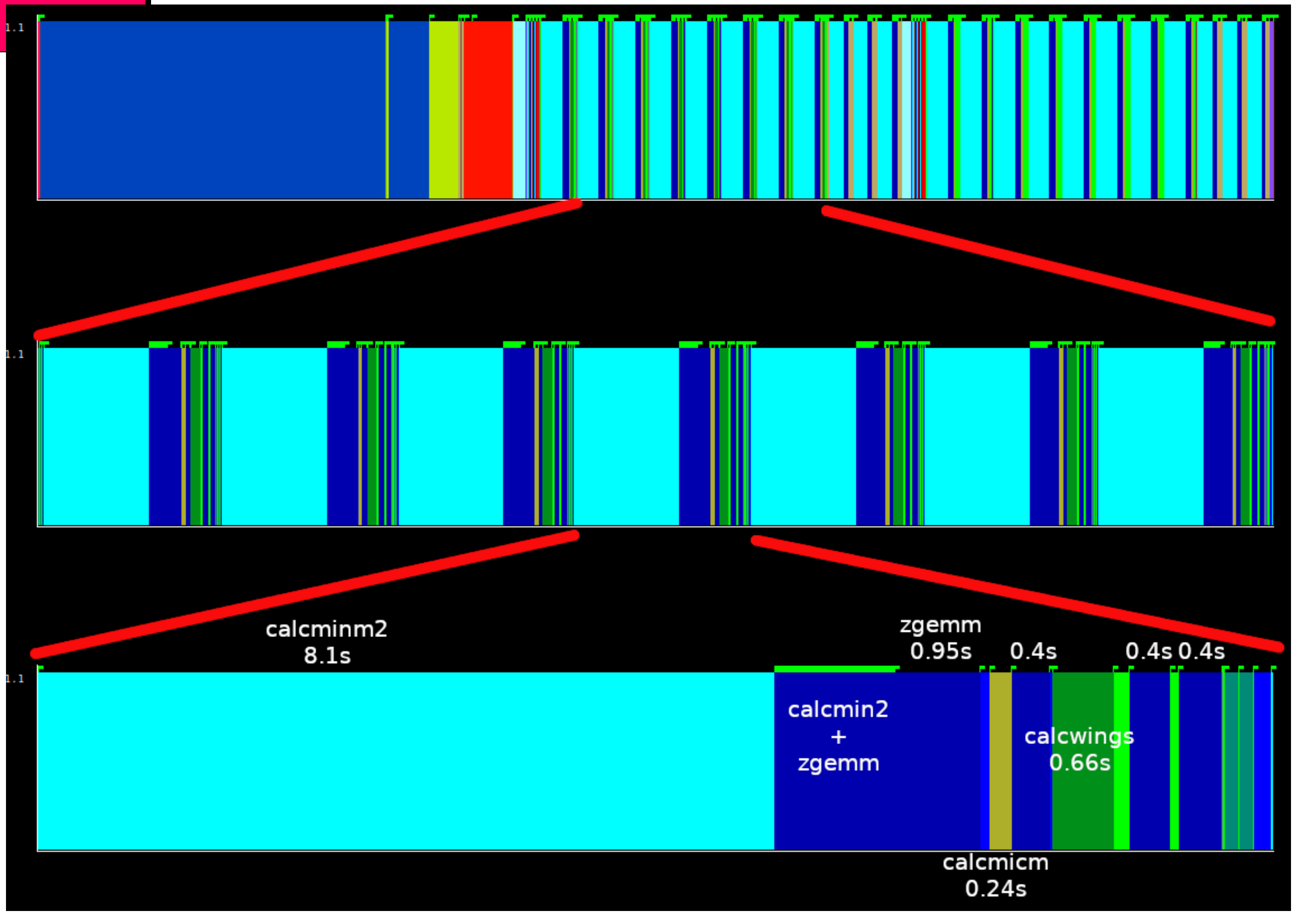




Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Mini-app development

Trace of the first thread - first process

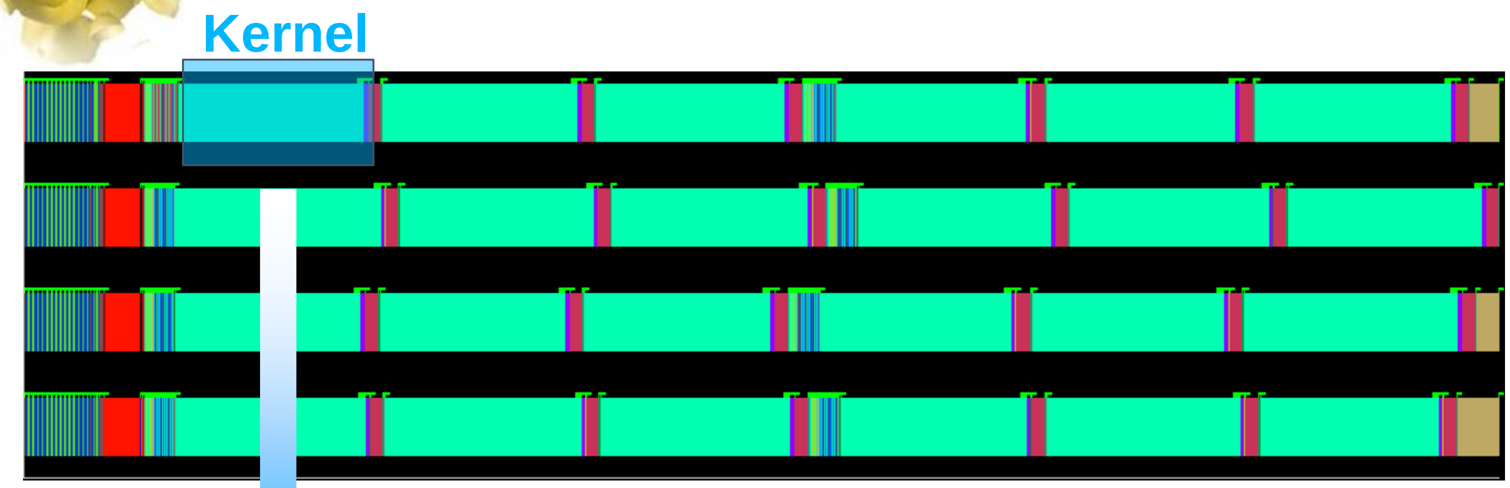


Look deeper: ***calcepsilon*** & ***calcselfc*** → iteratively calls to ***calcminm2*** + ***zgemm***

zgemm is called for big matrices (from ***calcepsilon*** and ***calcselfc***)

many times by **calcminm2** with a small size

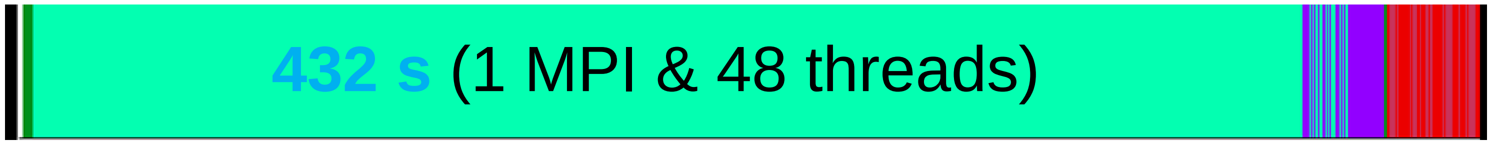
Mini-app development



3214 s (4 MPI processes & 48 threads)

isolates subroutines: **expand_products** (including calcminm2, ~80% runtime) and **calcmwm**

Developed a new checkpointing based on HDF5 & binary



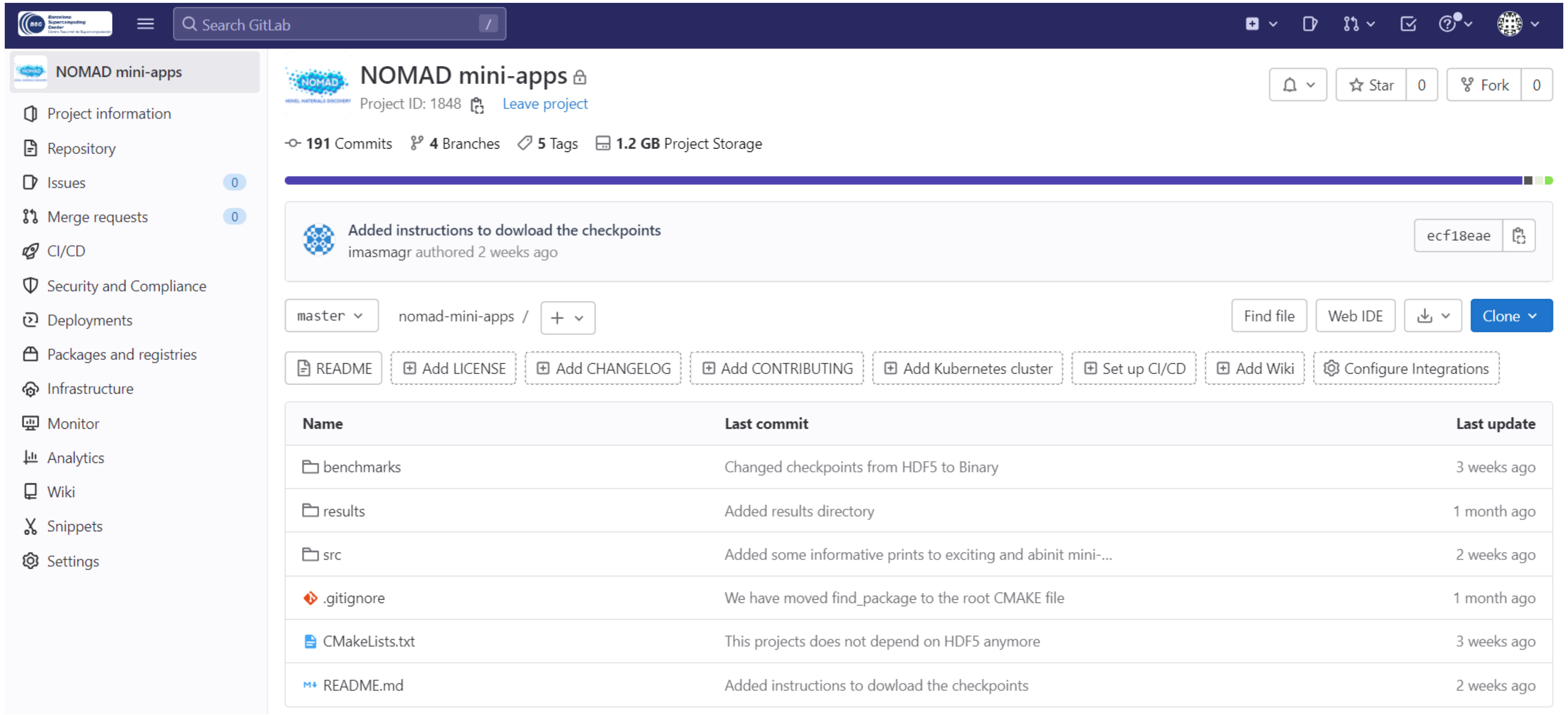
calcmwm_	102,264,974 ns
calcminc_	354,683,952 ns
expand_products_	4,319,179,072 ns
mod_mini..emm_tmp_	13,634,904,347 ns
mod_mini..emv_tmp_	26,600,842,301 ns
calcminm2_	394,802,263,730 ns
Total	439,814,138,376 ns
Average	73,302,356,396 ns
Maximum	394,802,263,730 ns
Minimum	102,264,974 ns
StDev	144,072,737,255 ns
Avg/Max	0

The NOMAD mini-apps suite

GW implementation

- exciting
- Abinit
- FHI-AIMS

ELPA lib eigensolver



cmake compilation

Tested dependencies on:

- MN4: Intel, GNU
- CTE POWER (IBM Power-9) : GNU, XLF, PGI
- KAROL1NA: Intel, AMD, NVIDIA

open to contributions

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Take-home messages

- The (post)exascale era is **full of technical challenges, but also opportunities**
- Work on **performance portability**
- Profit performance from: **accelerators**, network… but be careful with GPU-to-GPU communications
- Implement modular codes (interoperability) & integrate on workflows
- Enforce the application’s robustness & resilience (check-pointing)
- Machine learning

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Thank you!

EIG CONCERT JAPAN
Connecting and Coordinating European Research and Technology Development with Japan

PCI2023-143426 funded by MCIN/AEI/10.13039/501100011033

NOMAD
NOVEL MATERIALS DISCOVERY

MAX DRIVING THE EXASCALE TRANSITION

PRACE Summer of HPC

FUSION CAT

Generalitat de Catalunya
Departament de Recerca i Universitats

Unió Europea
Fons Europeu de Desenvolupament Regional

julio.gutierrez@bsc.es

materialsmodelling.wordpress.com