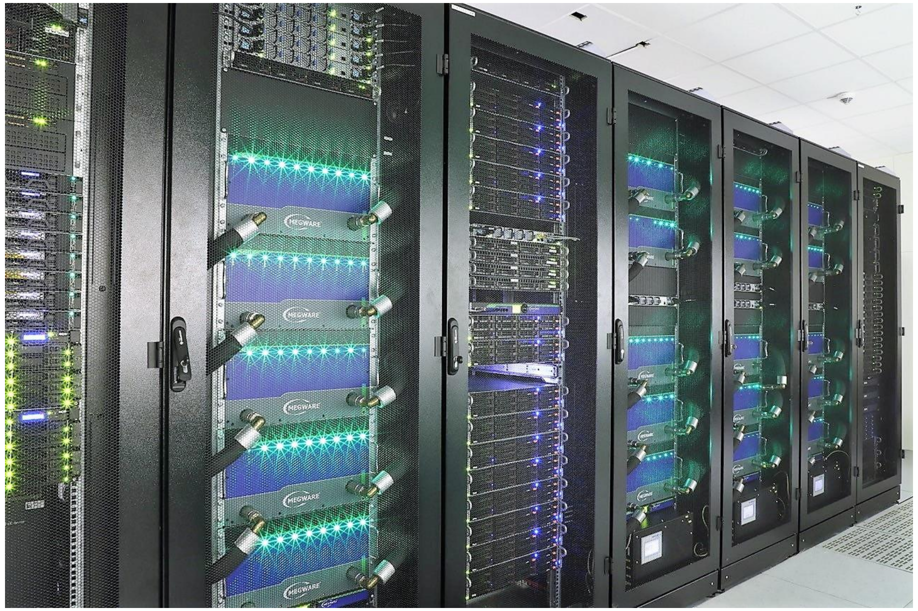
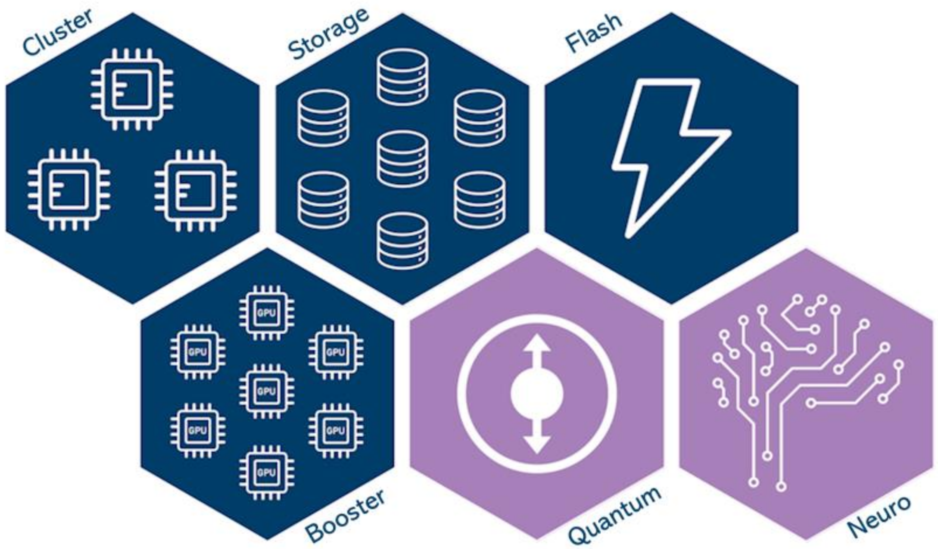
DEEP Projects



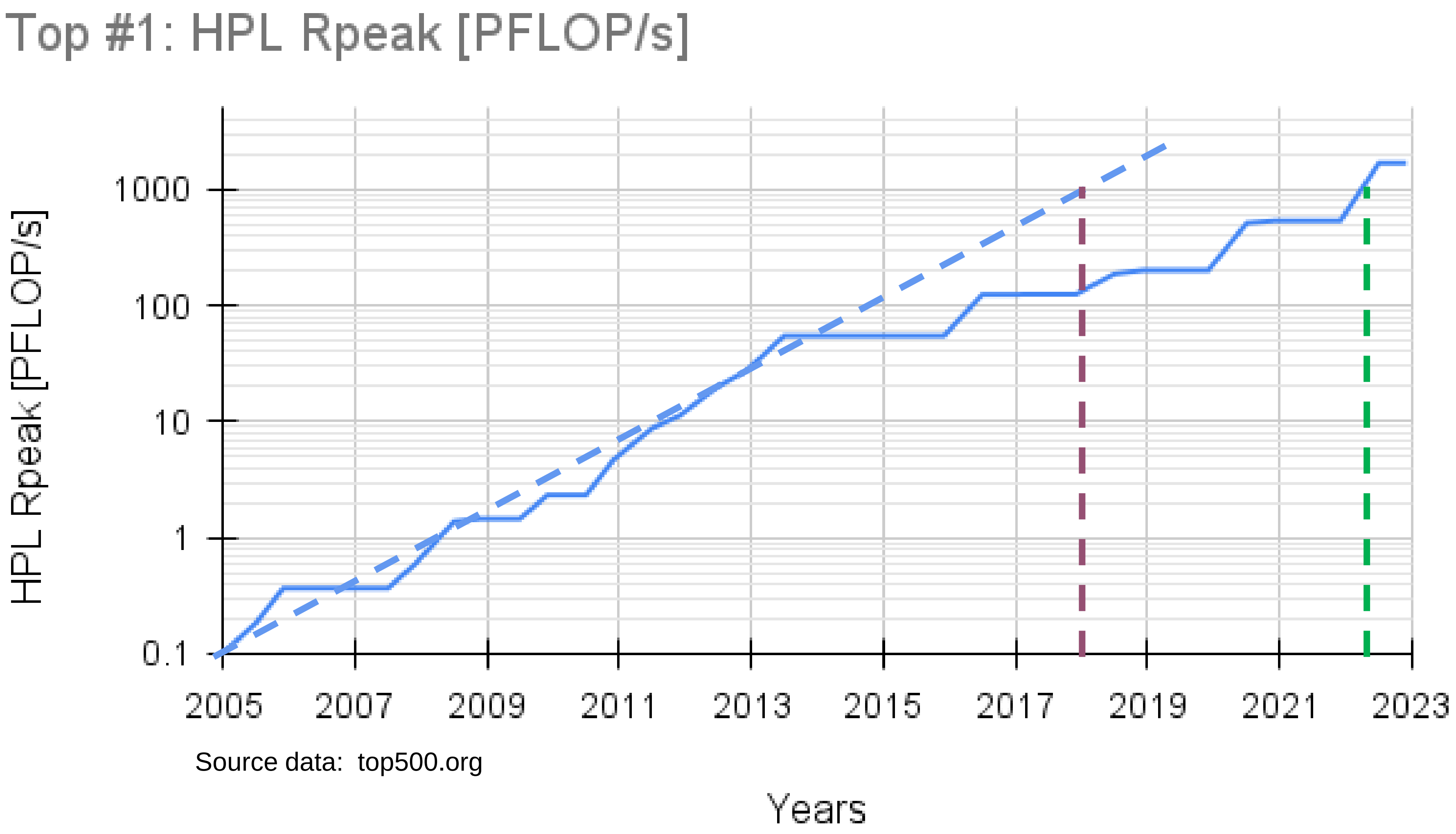


The DEEP-SEA Project

Hans-Christian Hoppe, Jülich Supercomputing Centre
CONCERTO Workshop at HiPEAC 2024 Conference

EVIDEN an atos business | BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación | cea | ECMWF | ETH zürich | FORTH | Fraunhofer ITWM | JÜLICH Forschungszentrum | KTH | KATHOLIEKE UNIVERSITEIT LEUVEN | ParTec MODULAR SUPERCOMPUTING | lrz Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften | TECHNISCHE UNIVERSITÄT DARMSTADT | Technical University of Munich

Road to Exascale – Slower than Expected



Source data: top500.org

1997: First **1TFlop/s** computer:
(ASCI Red/9152)

2008: First **1 PFlop/s** computer: (*Roadrunner*)

So…. First **1 EFlop/s** computer: **2018 !!**

– Well… not really

It took 4 years longer….
2022
for *Frontier* to appear

Exascale Challenges

Application parallelism
- Applications must support billions of individual threads
- Lower-scaling applications / parts of applications should not run on a full Exascale system

DEEP-SEA

Truly scalable systems
- Huge numbers of devices need to exchange data with each other
- Collective communication operations are “slowing down” due to larger system sizes
- Network contention and reliability become worries

Energy efficiency
- Accelerators clearly beat CPUs for many (most?) codes
- System heterogeneity is a must
- Yet – portable accelerator programming is hard

DEEP-SEA

Memory and storage
- Ever growing gap between compute throughput and memory bandwidth
- New technologies like HBM suffer from capacity limitations & high energy consumption

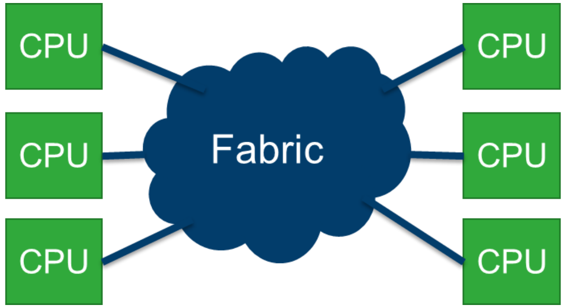
DEEP-SEA

Workload diversity
- Exascale centers must run a wide variety of HPC, AI and data analytics workloads with highest energy efficiency
- One size does not fit all

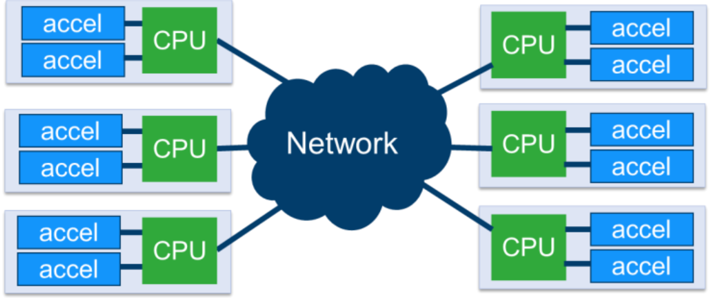
DEEP-SEA

Approaches to System Heterogeneity

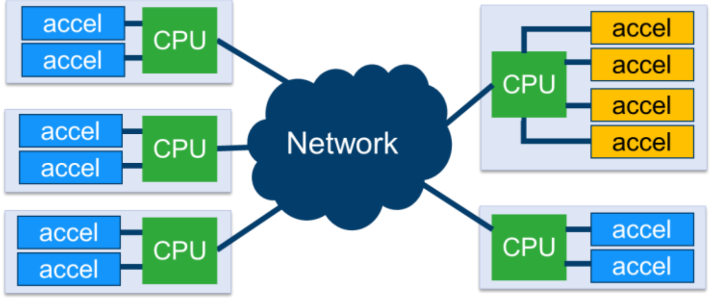
Homogeneous



Accelerated Nodes



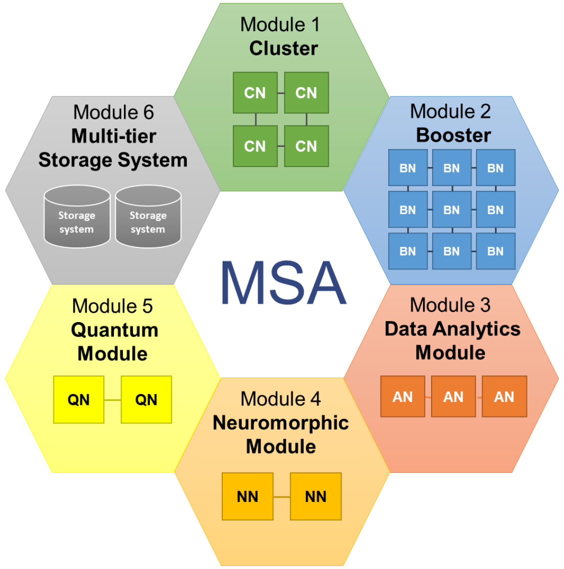
Accelerated Nodes + Special Nodes



Homogeneous systems lack efficiency*

Accelerated nodes fix the ratio of CPUs vs. accelerators, complicate sharing resources across nodes

Adding „special nodes" for certain tasks



*: certainly for AI and dense linear algebra applications

Modular Supercomputing Architecture

Composability of heterogeneous resources

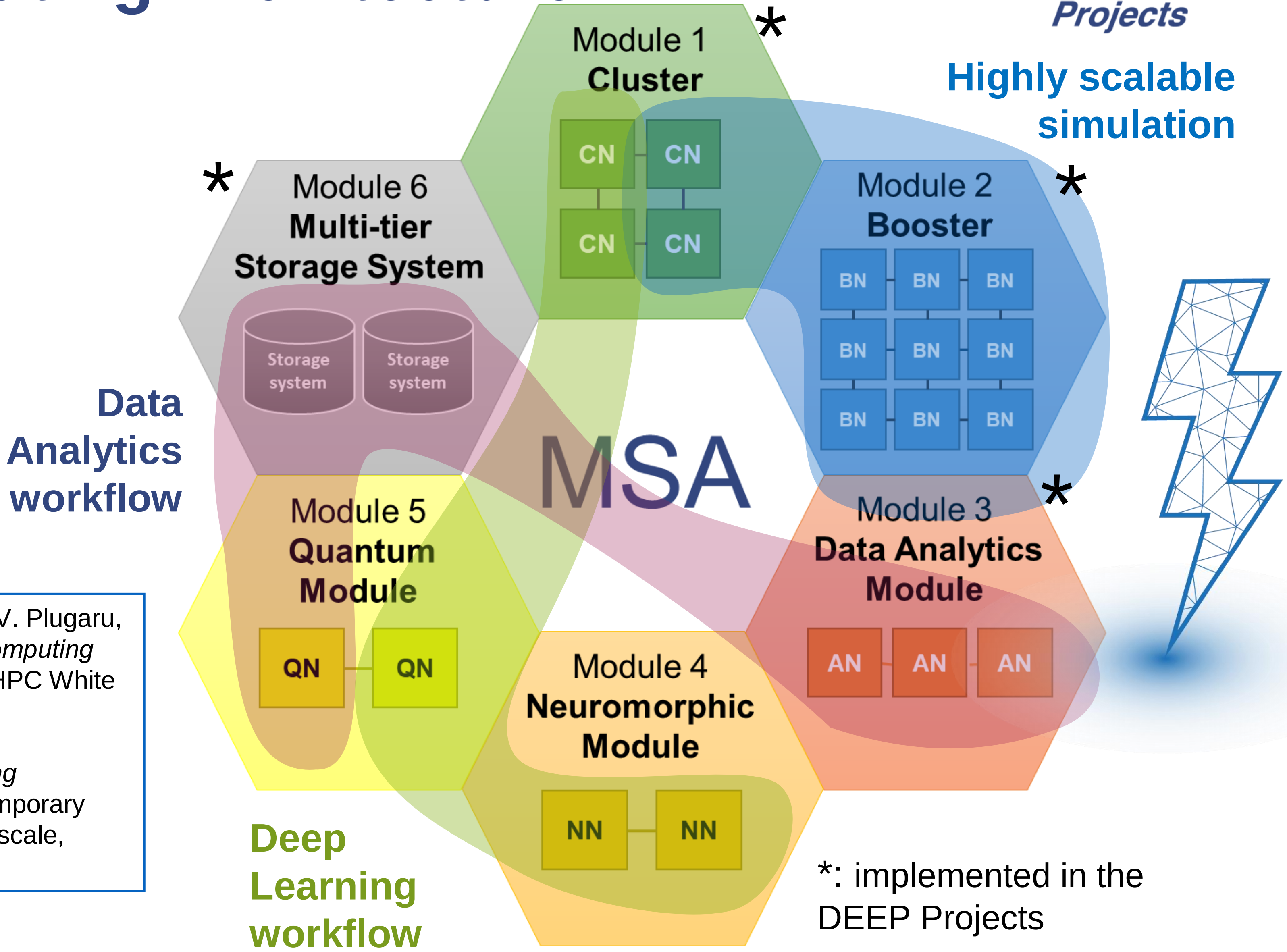
Cost-effective scaling

Effective resource-sharing

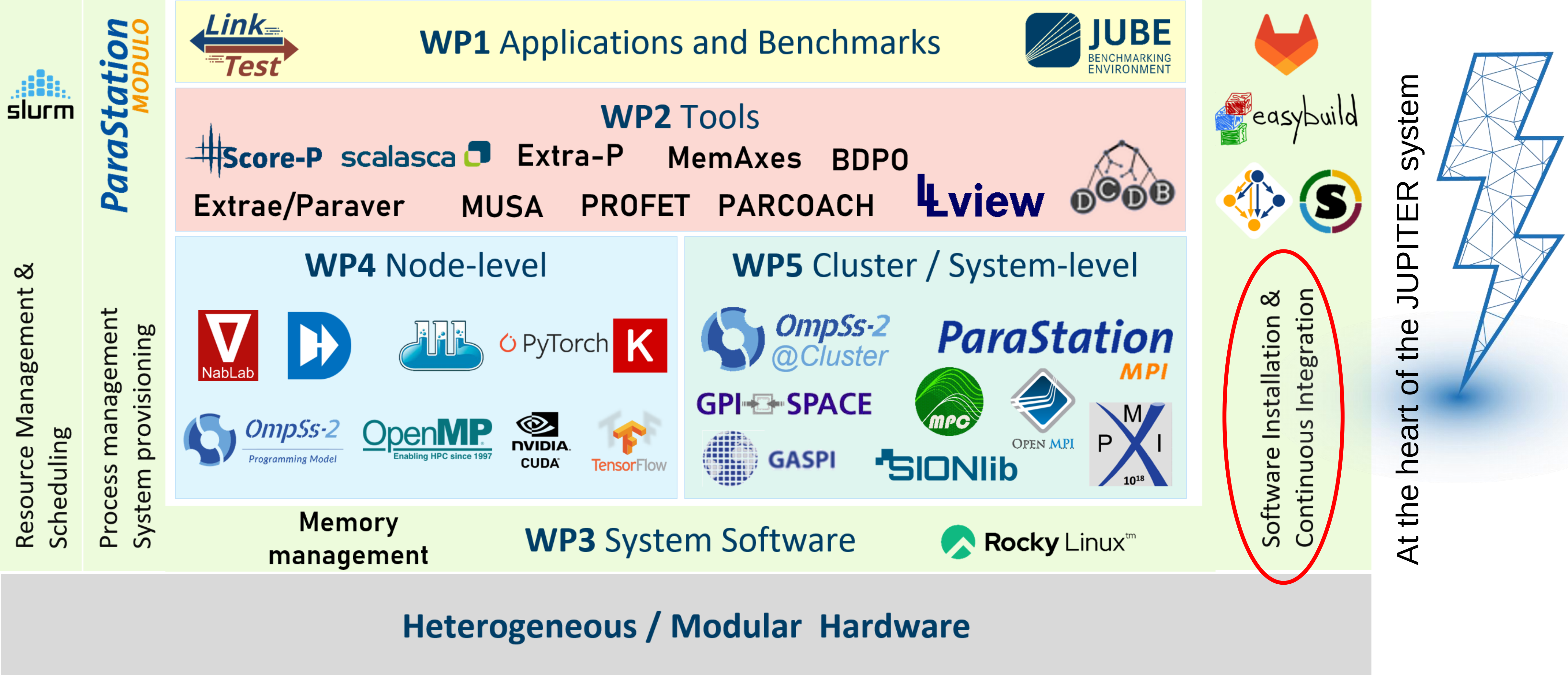
Match workload diversity

- Data analytics
- Machine- and Deep Learning
- Artificial Intelligence

- E. Suarez, N. Eicker, T. Moschny, S. Pickartz, C. Clauss, V. Plugaru, A. Herten, Kristel Michielsen, T. Lippert, *“Modular Supercomputing Architecture – A Success Story of European R&D”,* ETP4HPC White Paper. (2022) Available at https://www.etp4hpc.eu/white-papers.html#msa.
- E. Suarez, N. Eicker, Th. Lippert, "*Modular Supercomputing Architecture: from idea to production*", Chapter 9 in Contemporary High Performance Computing: from Petascale toward Exascale, Volume 3, p 223-251, CRC Press. (2019)



Integrated Exascale-Ready SW Stack



Public release at https://gitlab.jsc.fz-juelich.de/deep-sea/wp3/software/easybuild-repository-deep-sea

Optimisation Cycles

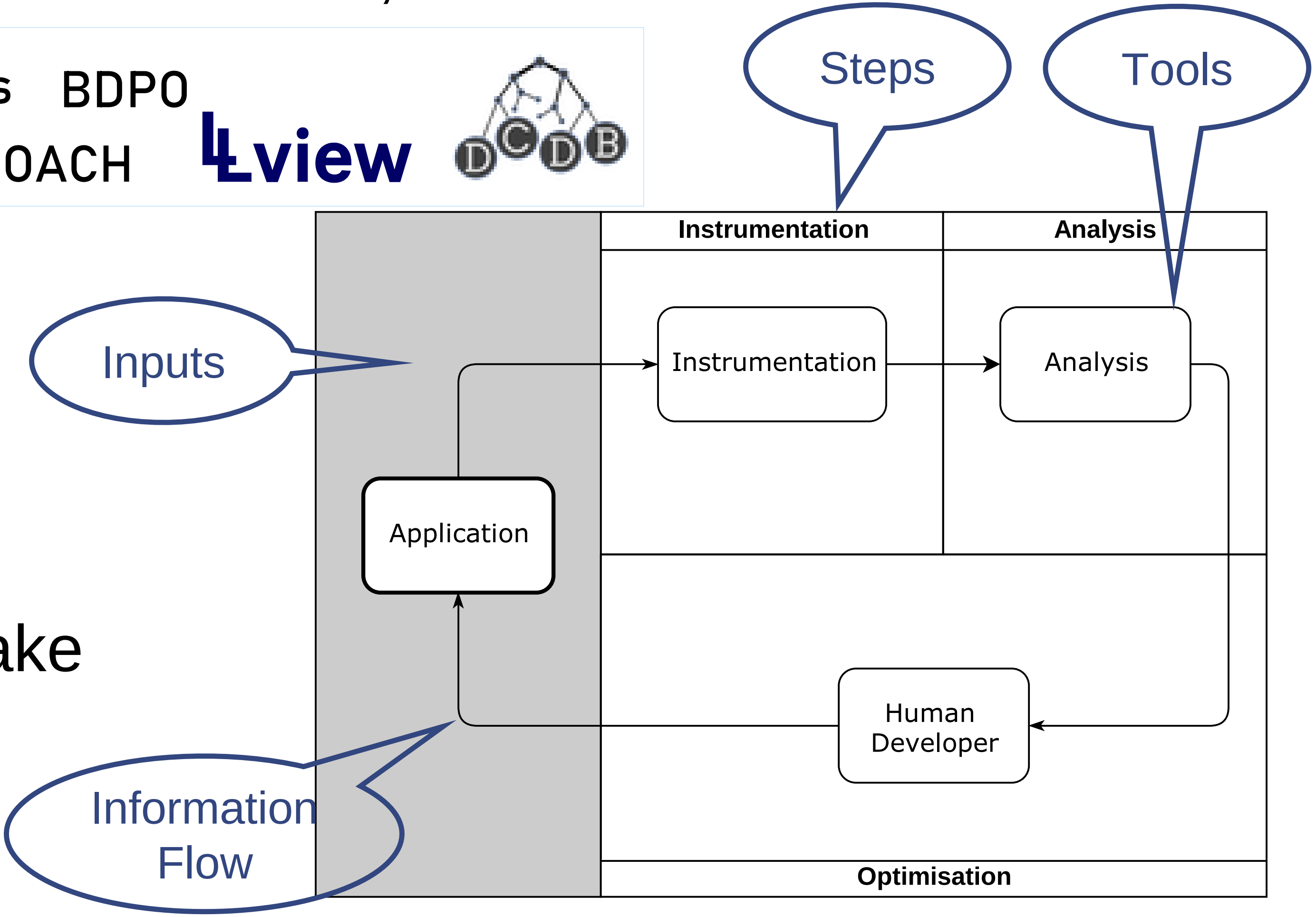
Bewildering variety of SW tools available to HPC SW developers for analysis and optimisation – in DEEP-SEA alone, these:

Score-P scalasca Extra-P MemAxes BDPO
Extrae/Paraver MUSA PROFET PARCOACH LLview DCDB

Optimisation cycles encapsulate (complex) tool workflows for *specific purposes*

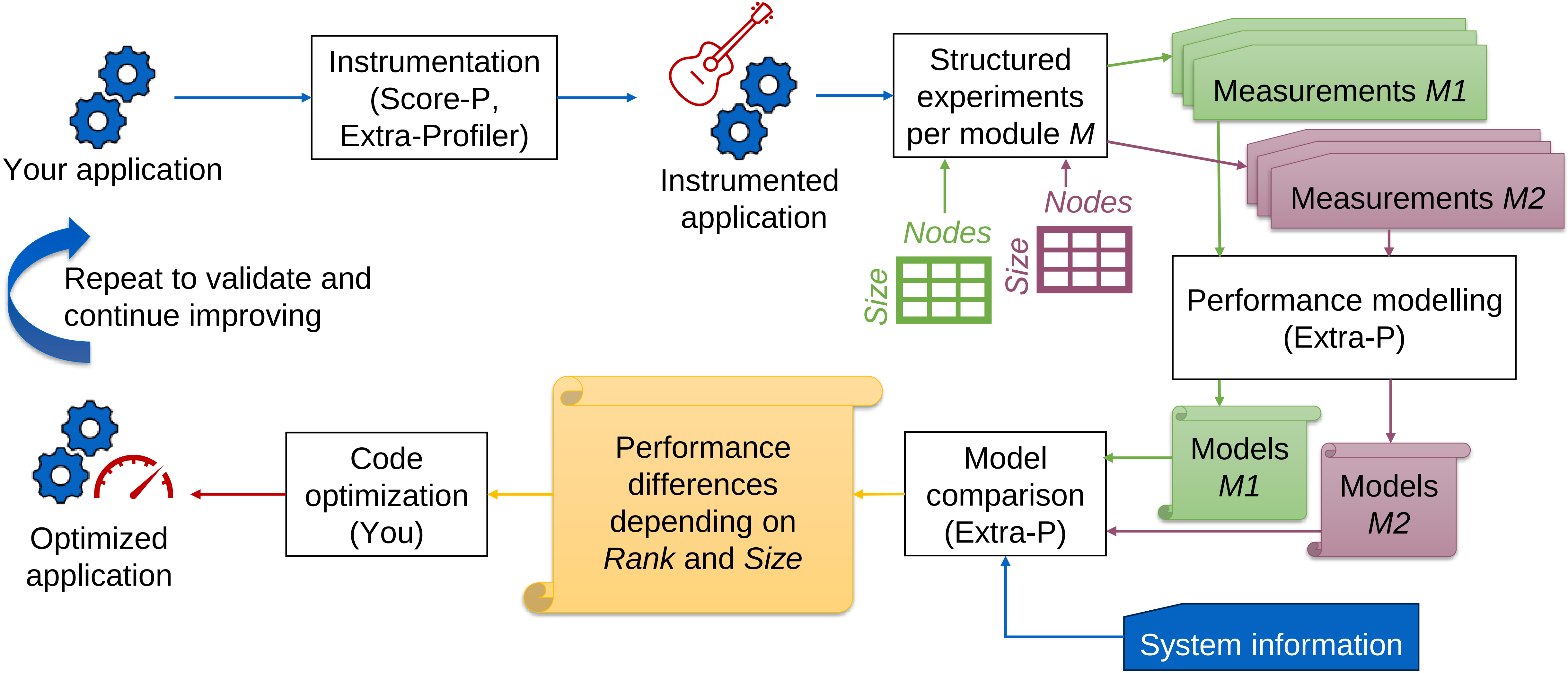
- Like assessing load balance or optimising energy use

They guide SW developers and make it easier to achieve specific goals



Based on material by A. Geiß, TU Darmstadt

Application Mapping Optimisation Cycle



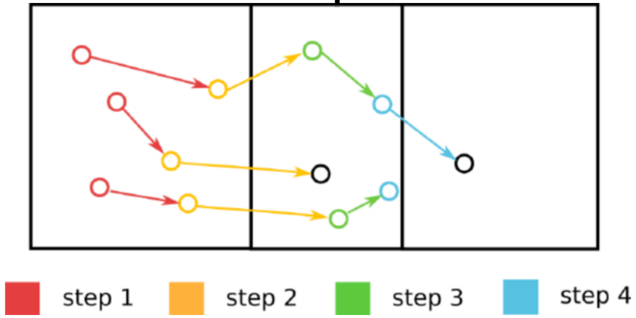
Based on material by A. Geiß, TU Darmstadt

Use Case: PATMOS

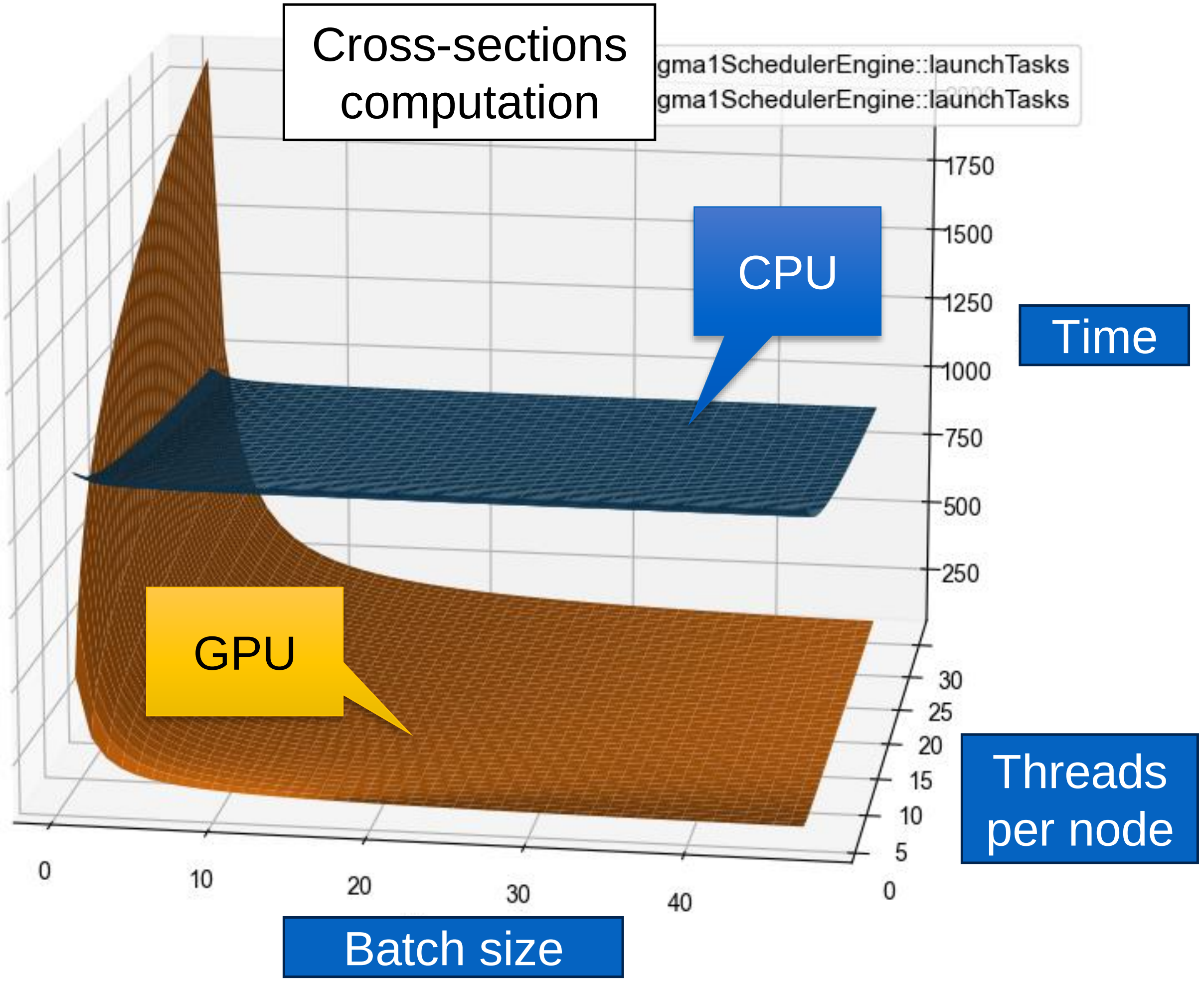
Solves the neutron transport equations to simulate evolution of physical quantities for complex systems

Cross-sections computation represents 60% to 90% of total runtime

- Porting cross section computation to GPU
- Offload batch-size particles at a time



Split of application depends on batch size

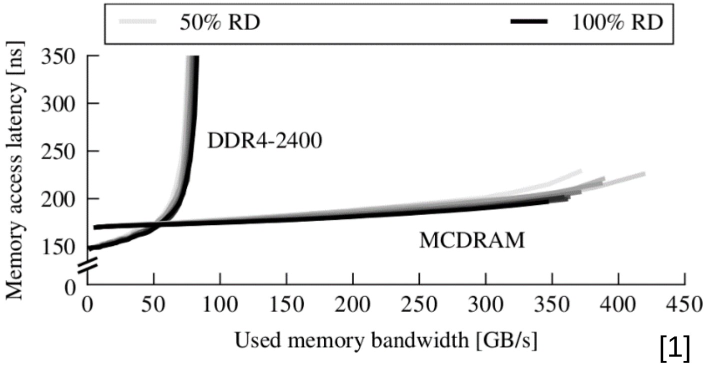


Based on material by T. Hovi, CEA

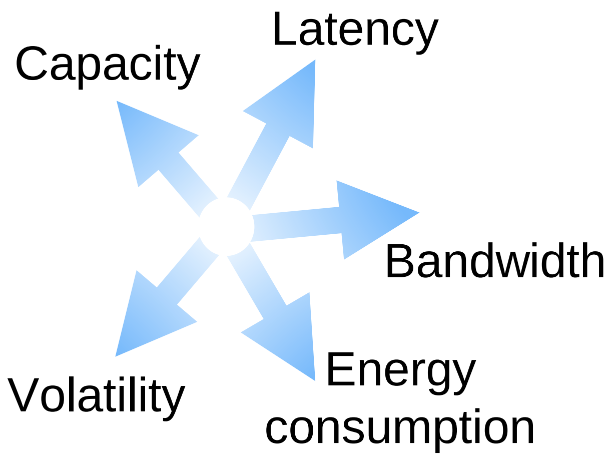
Heterogeneous/Hierarchical Memory

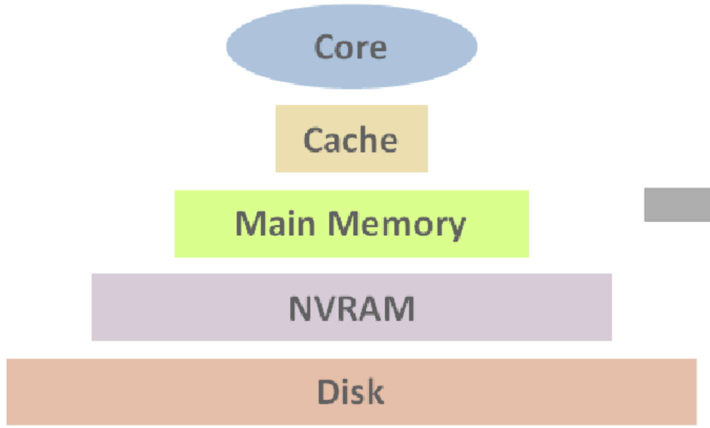
Examples…

- DDR DRAM
- Scratchpad (Embedded systems-on-chip, GPUs)
- High bandwidth memory (Intel Xeon Phi, GPUs)
- Byte addressable non-volatile memory (HP's Machine, Intel Optane)
- Compute Express Link (CXL): high-speed interface to accelerators and memory modules

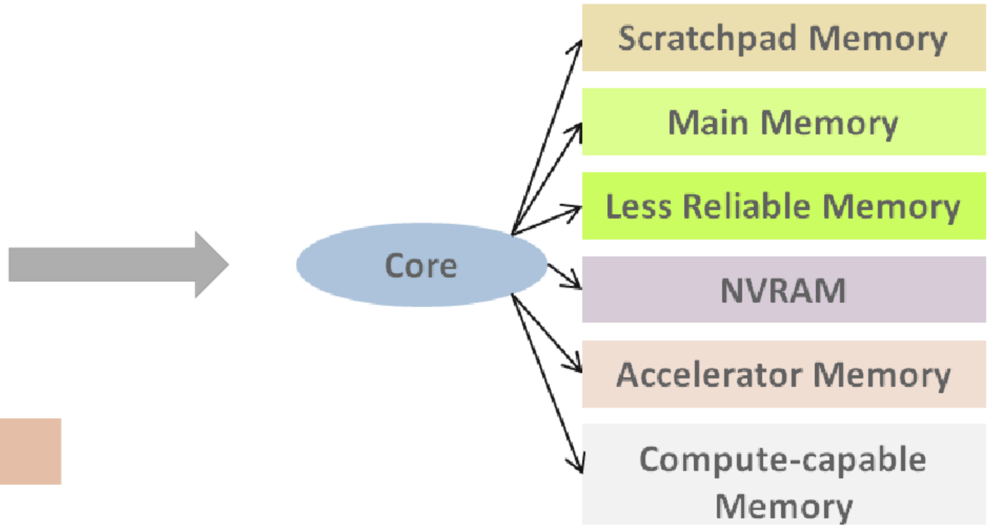


[1]





Memory hierarchy



Explicitly managed

[1] Milan Radulovic et al. PROFET: Modeling System Performance and Energy Without Simulating the CPU. ACM SIGMETRICS 2019

Heterogeneous/Hierarchical Memory Tools

- To which degree do the applications need to be modified?
- Which layer manages the memory? When?
- How much can the applications benefit?

Modifying the application

Standard interfaces
OpenMP 5.0, SICM
OpenMP

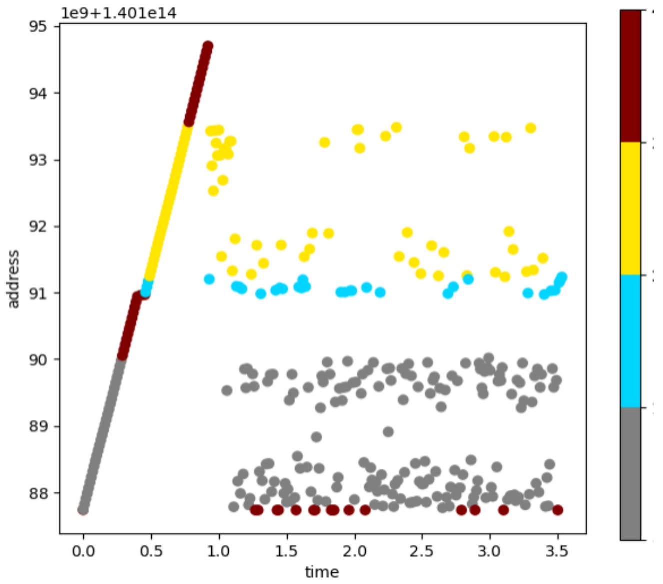
Transparent to the application

Static
(compile time)

ecoHMEM
object level mapping by source-to-source compiler based on profile run

Dynamic
(run time)

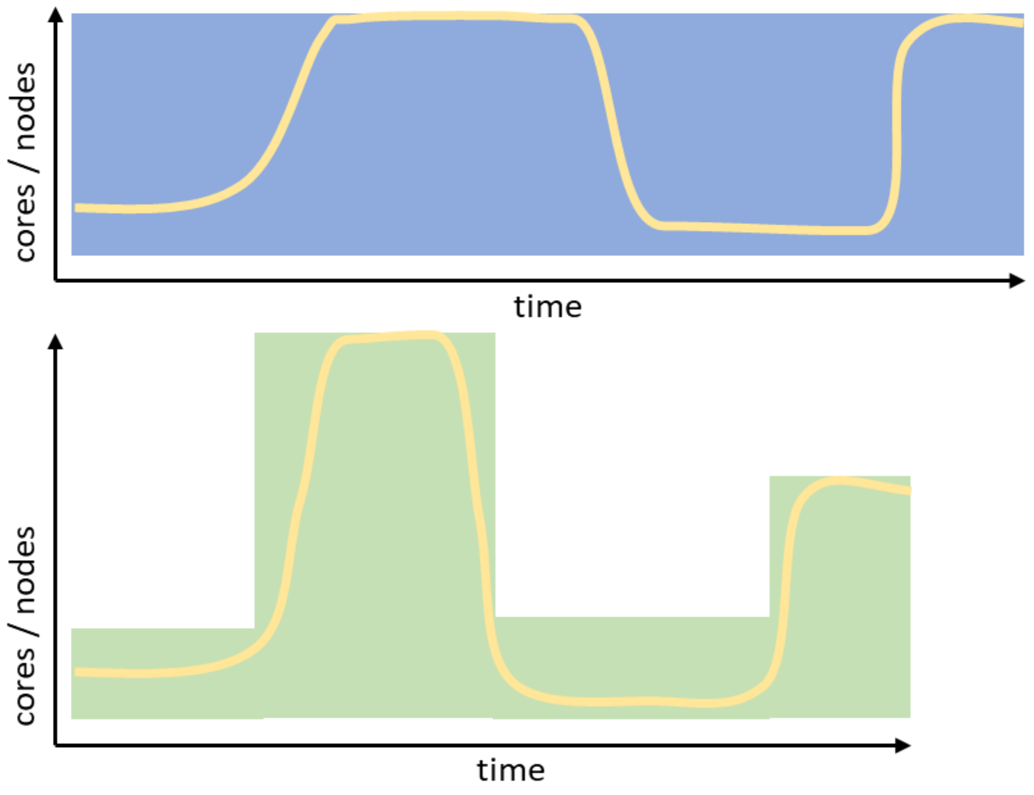
SHAMBLES
Run time page-level migration by Operating System



SHAMBLES scatter plot example for sparse kernel

Malleability

- Usual HPC workload resource reservation (constant # cores or nodes over time)
- Actual use of resources varies over time (yellow curve)
- Workload is able to use more resources in certain phases (arrow)
- Ideal resource allocation for the workload in green
- Malleable applications
  - Release resources not required
  - Acquire more resources if advantageous
- Change in # of nodes do require data redistribution in the workload
- DEEP-SEA provides MPI & Slurm prototypes for enabling application-driven (active) malleability

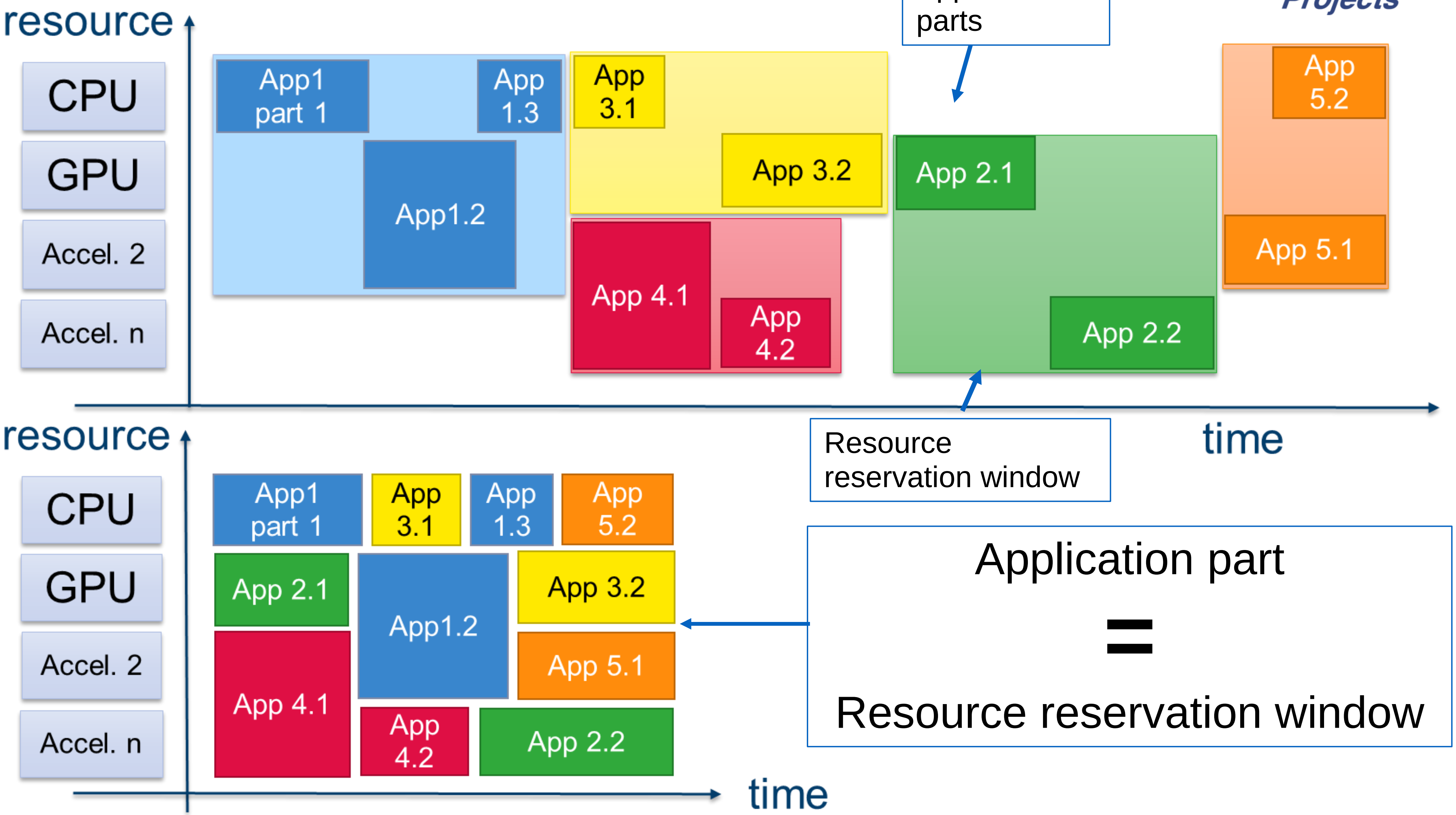


Scheduling

Current behaviour

WIP

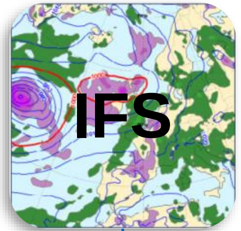
Ideal behaviour



Application part
=
Resource reservation window

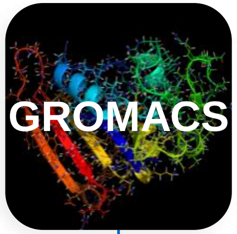
Seven Co-Design Applications

Space Weather
ML, Python, C++, OpenMP, MPI, CUDA



Weather and Climate
ML, Fortran, OpenMP, MPI

Seismic Imaging
C++, OpenMP, MPI, CUDA, GASPI



Molecular Dynamics
C++, OpenMP, MPI, CUDA

Computational Fluid Dynamics
Fortran, MPI

Neutron Monte-Carlo Transport
C++, OpenMP, MPI

Earth Systems Modelling
C, C++, Fortran, MPI

Funding Acknowledgement

DEEP Projects

EuroHPC Joint Undertaking

SPONSORED BY THE Federal Ministry of Education and Research

fwo

bpifrance

ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
ΥΠΟΥΡΓΕΙΟ ΑΝΑΠΤΥΞΗΣ ΚΑΙ ΕΠΕΝΔΥΣΕΩΝ

ΓΓΕΚ
ΓΕΝΙΚΗ ΓΡΑΜΜΑΤΕΙΑ ΕΡΕΥΝΑΣ ΚΑΙ ΚΑΙΝΟΤΟΜΙΑΣ

MINISTERIO DE CIENCIA E INNOVACIÓN

Financiado por la Unión Europea
NextGenerationEU

Plan de Recuperación, Transformación y Resiliencia

AGENCIA ESTATAL DE INVESTIGACIÓN

Swedish Research Council

euresearch
Swiss guide to European research and innovation

The DEEP Projects have received funding from the European Commission's FP7, H2020, and EuroHPC JU Programmes, under Grant Agreements n° 287530, 610476, 754304, and 955606. The DEEP-SEA project receives also support from Belgium, France, Germany, Greece, Spain, Sweden, and Switzerland

www.deep-projects.eu
@DEEPprojects