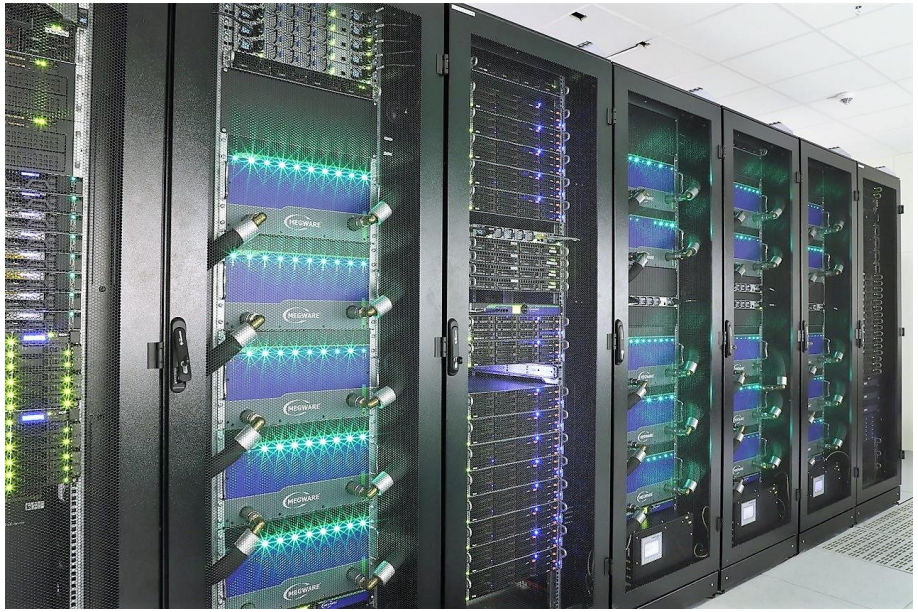
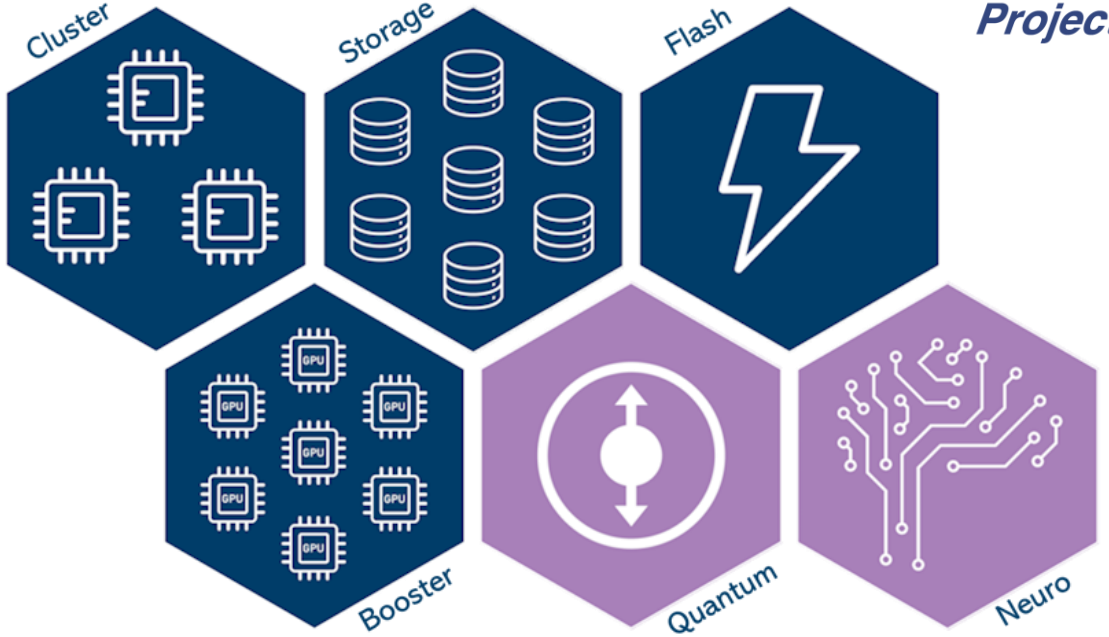




Towards Highly Efficient Heterogeneous Supercomputers - the DEEP Approach

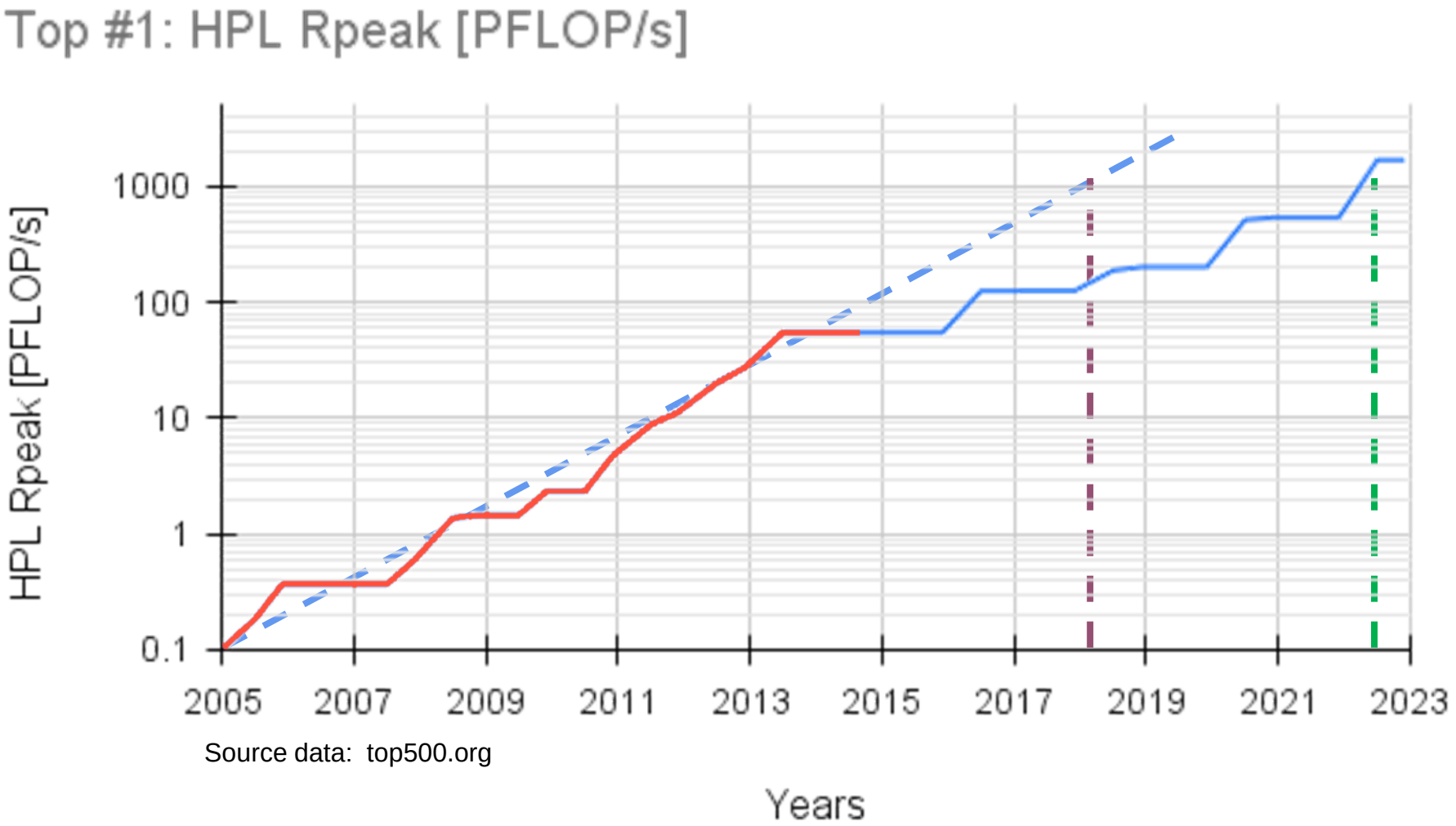
Hans-Christian Hoppe, Jülich Supercomputing Centre

Plasma Physics Towards the ExaScale Era WS at HiPEAC 2024

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

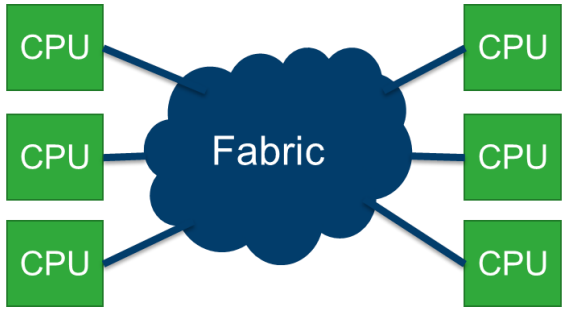
DEEP-SEA

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

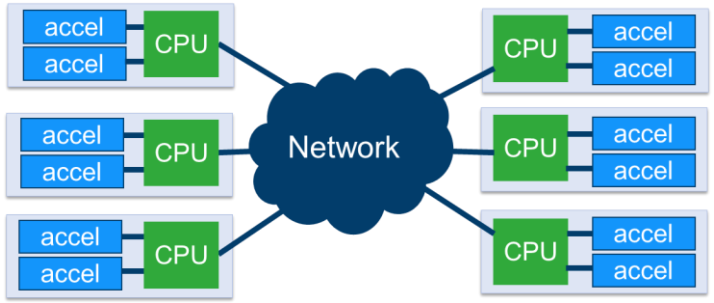
DEEP-SEA

Heterogenous Systems – HPC Centre View

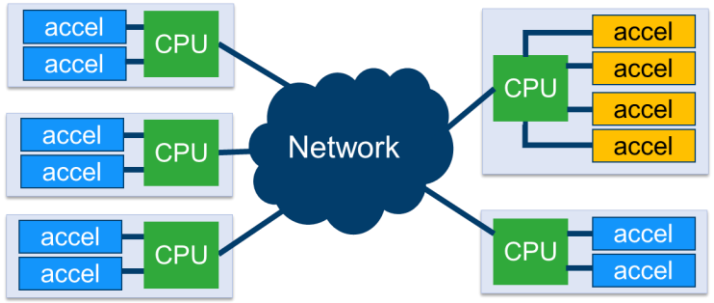
Homogeneous



Accelerated Nodes



Accelerated Nodes + Special Nodes



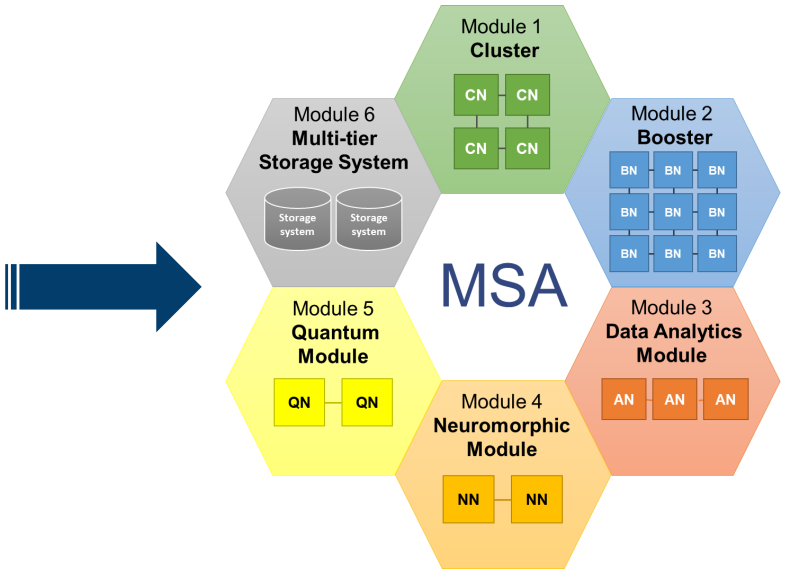
Different workloads need different CPU vs. accelerator ratios

- Statically configured systems are always a compromise
- “Dark” silicon eating energy for nothing for some WLs

Restriction of achievable performance on other WLs

Adding “special” nodes only helps so much …

Really want to be able to compose arbitrary mixes of CPUs plus accelerators



Modular Supercomputing Architecture

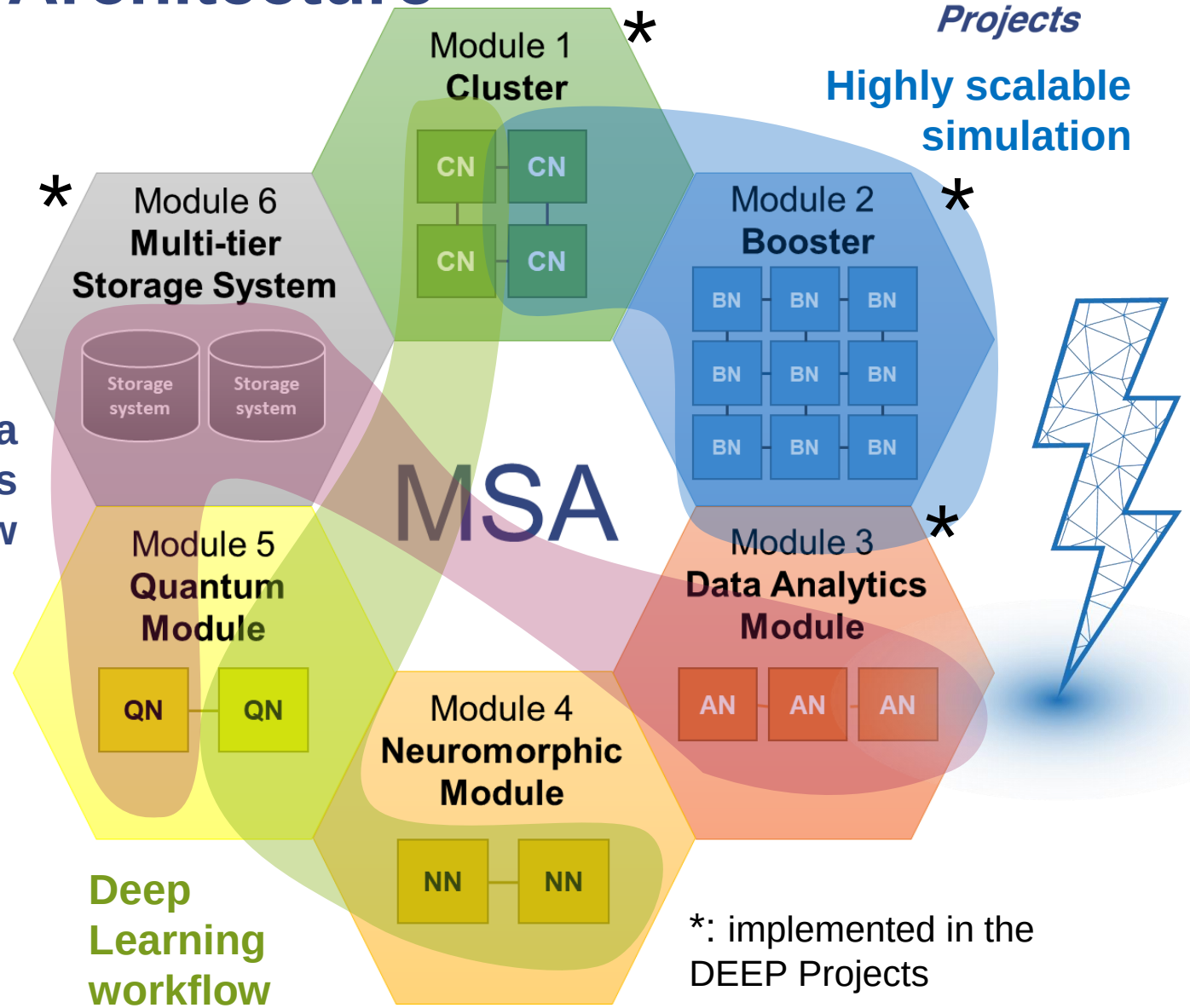
Composability of heterogeneous resources

Cost-effective scaling

Effective resource-sharing

Match workload diversity

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



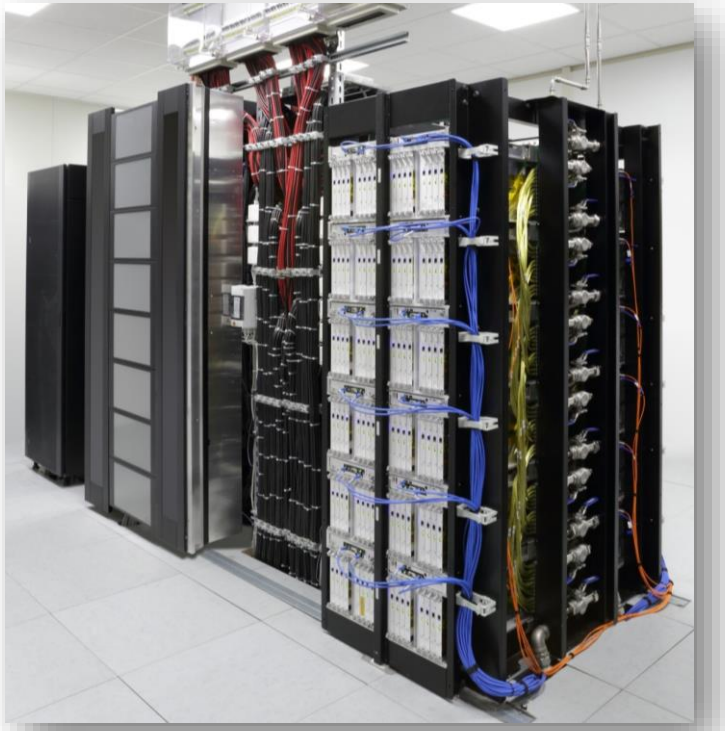
*: implemented in the DEEP Projects

Highly scalable simulation

- 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)

DEEP Series of Prototype Systems

© FZJ



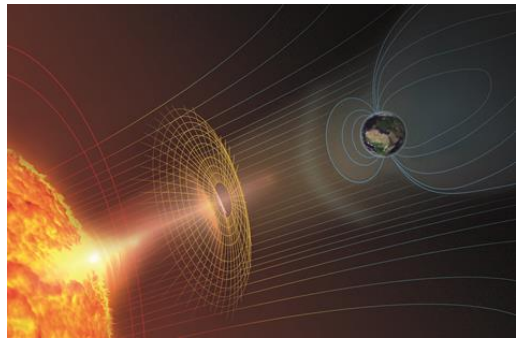
DEEP Prototype
128 Xeon + 284 KNC nodes
InfiniBand + 1.5Gbit Extoll
550 TFlop/s

DEEP-ER Prototype
16 Xeon + 8 KNL nodes
100Gbit Extoll
40 TFlop/s

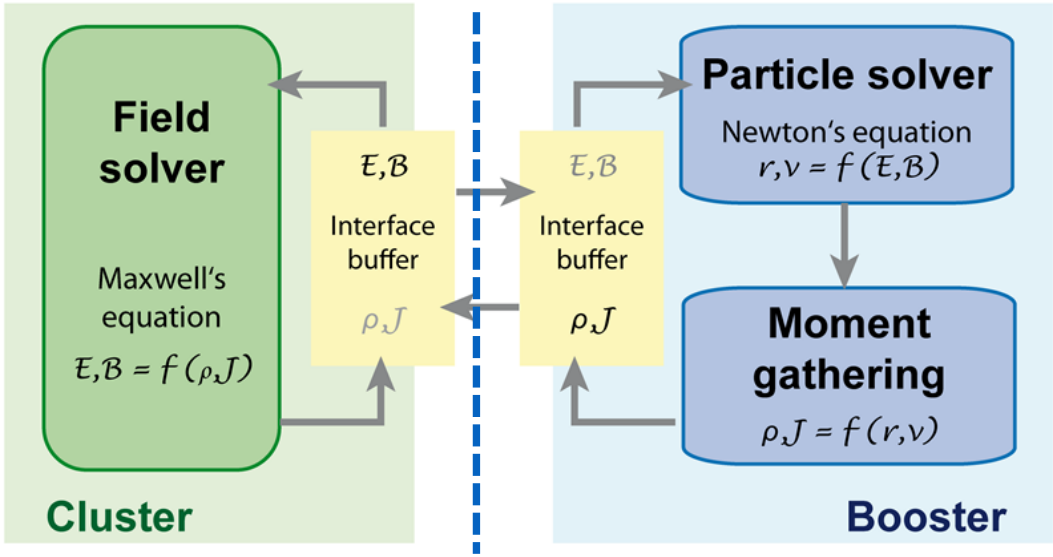
DEEP-EST Prototype
55 Cluster + 75 Booster + 16 Data Analytics
100 Gbit Extoll + InfiniBand + Eth
800 TFlop/s

Heterogenous Systems – Application View

- **Space Weather simulation**
  - Simulates plasma produced in solar eruptions and its interaction with the Earth magnetosphere
  - Particle-in-Cell (PIC) code
  - Authors: KU Leuven



- **Two solvers:**
  - **Field solver**: Computes electromagnetic (EM) field evolution
    - *Limited code scalability*
    - *Frequent, global communication*
  - **Particle solver**: Calculates motion of charged particles in EM-fields
    - *Highly parallel*
    - *Billions of particles*
    - *Long-range communication*



A. Kreuzer, J. Amaya, N. Eicker, E. Suarez, *"Application performance on a Cluster-Booster system",* 2018 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), HCW (20th International Heterogeneity in Computing Workshop), Vancouver (2018), p: 69 - 78. [doi: 10.1109/IPDPSW.2018.00019]

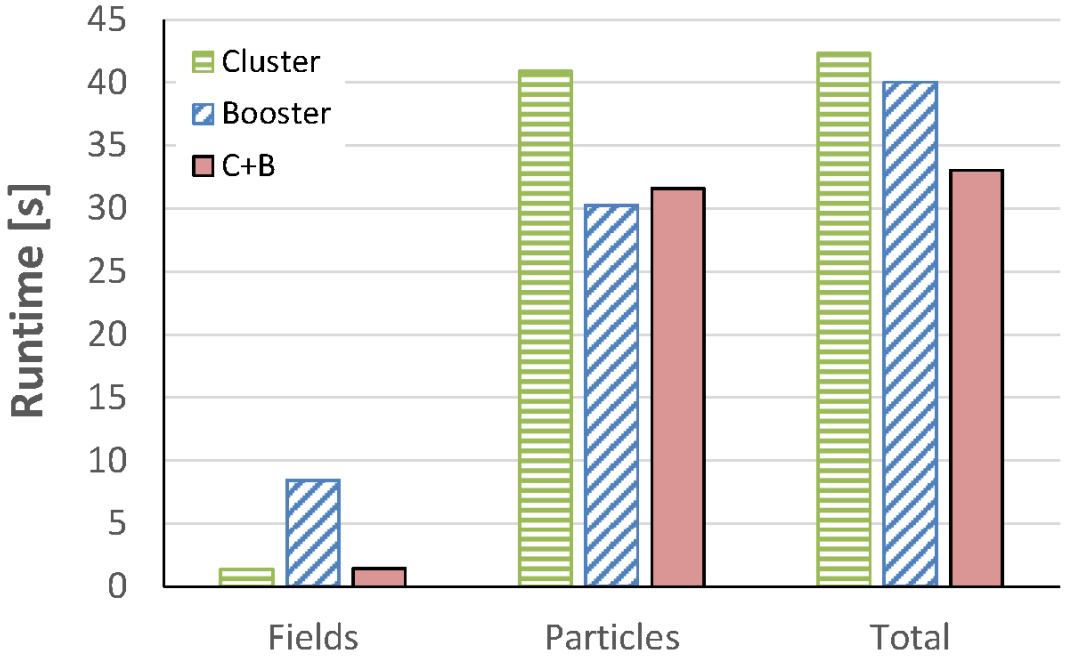
xPic – Small Scale Performance Results

- **Field solver**: 6× faster on Cluster
- **Particle solver**: 1.35 × faster on Booster
- **Overall performance gain:**

1× node
28% × gain compared to Cluster alone
21% × gain compared to Booster alone

8× nodes
38% × gain compared to Cluster only
34% × gain compared to Booster only

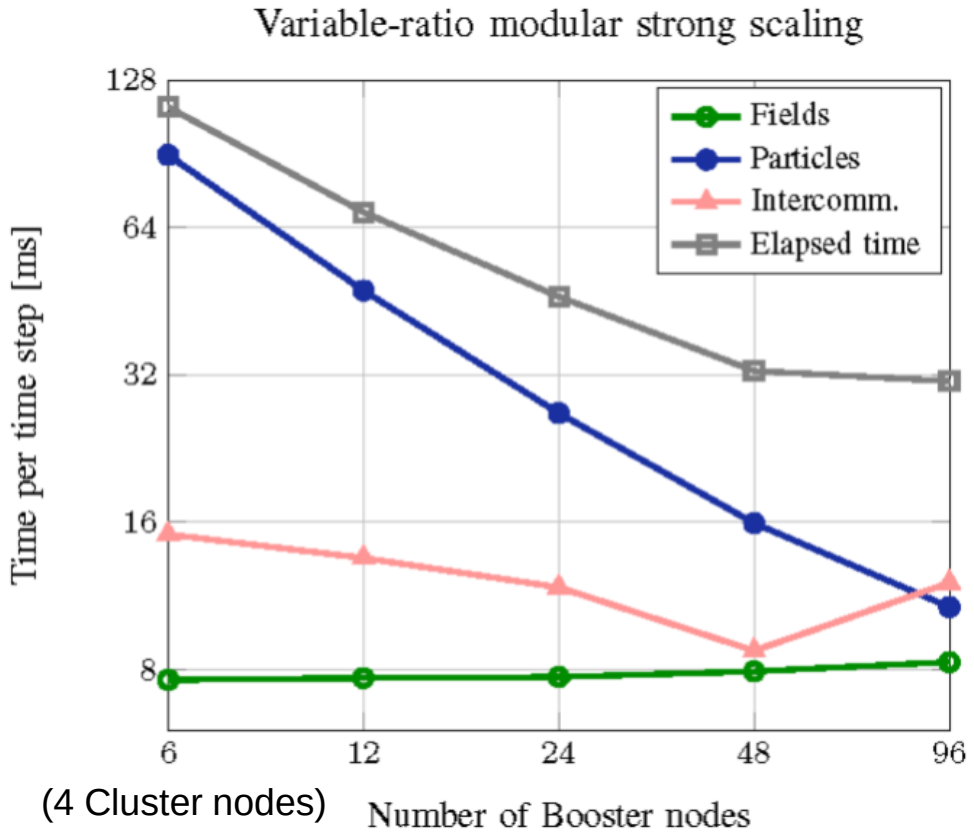
  - 3%-4% overhead per solver for C+B communication (point to point)



#cells per node	4096
#particles per cell	2048

A. Kreuzer et al. "*Application Performance on a Cluster-Booster System*", 2018 IEEE IPDPS Workshops (IPDPSW), Vancouver, Canada, p 69 - 78 (2018) [10.1109/IPDPSW.2018.00019]

xPic – Strong Scaling Behaviour

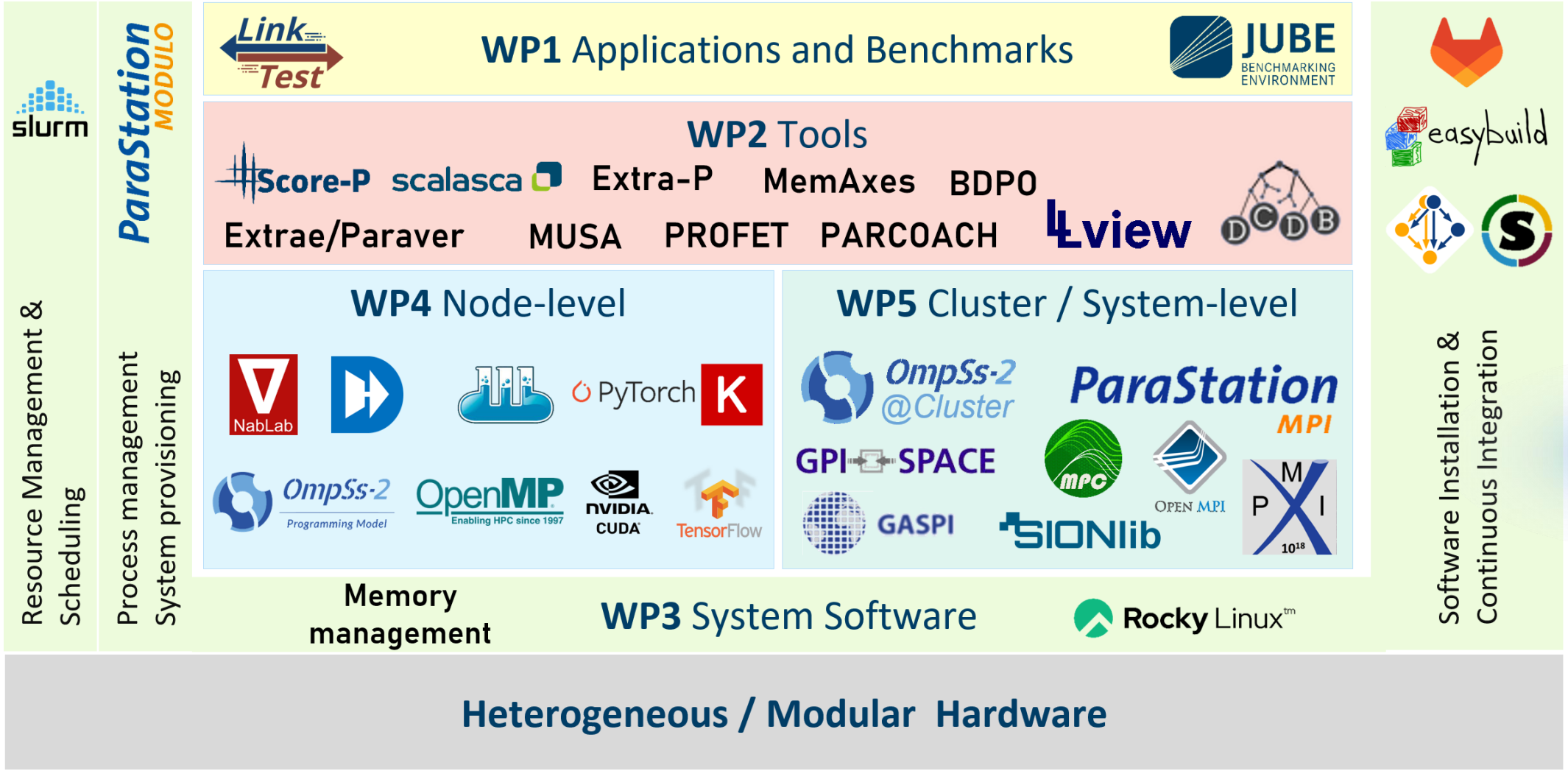


(4 Cluster nodes)

#cells per node	36864
#particles per cell	1024
#blocks per MPI process	12, 32 or 64

- JSC Jureca system – Intel® Xeon® plus Intel® Xeon Phi™ (KNL)
- Code portions can be scaled-up independently
  - Particles scale almost linearly on Booster
  - Fields kept constant on the Cluster (4CNs)
- A configuration is reached where same time is spent on Cluster and Booster
  - Additional 2× time-saving can be reached by co-scheduling “matching”xPic jobs

Integrated Exascale-Ready SW Stack



At the heart of the JUPITER system

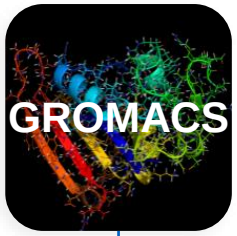
Public release at https://gitlab.jsc.fz-juelich.de/deep-sea/wp3/software/easybuild-repository-deep-sea

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



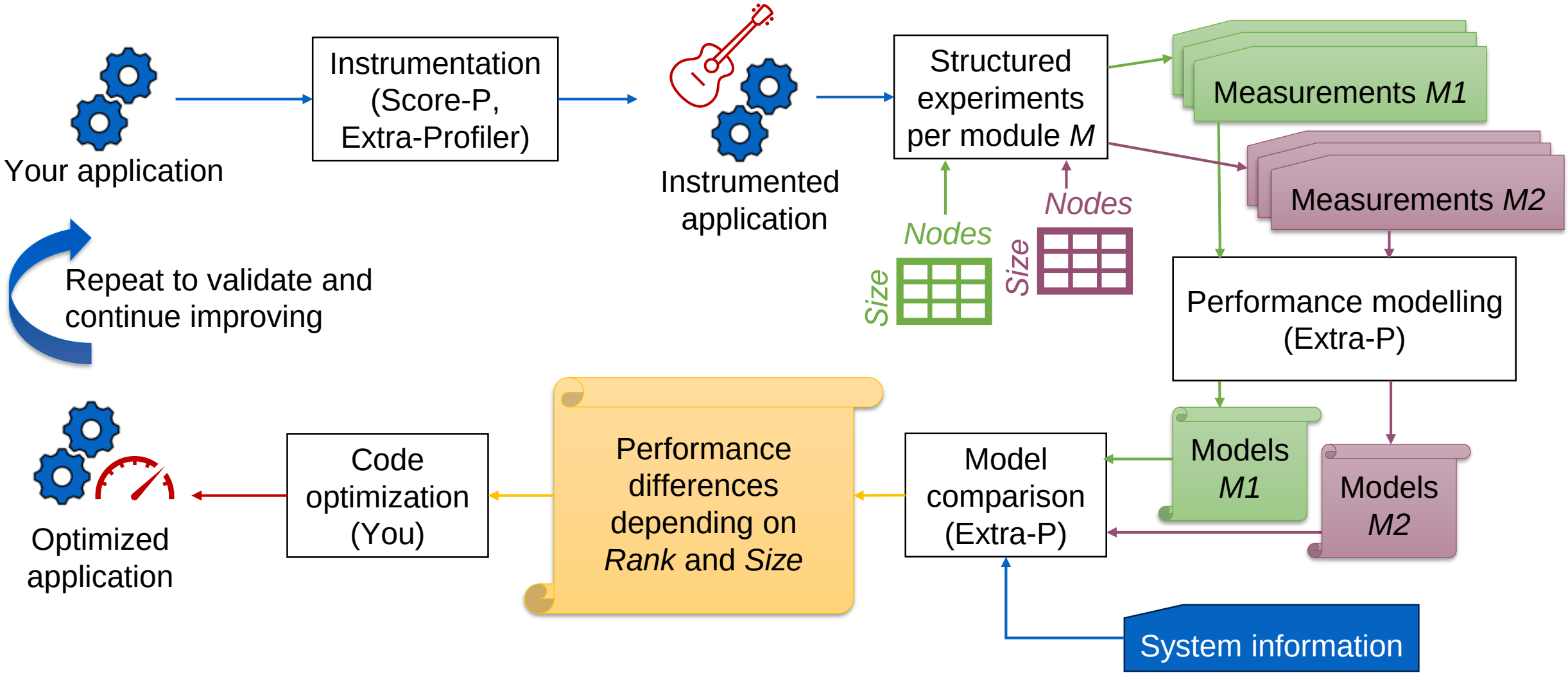
Molecular Dynamics
C++, OpenMP, MPI,
CUDA

Computational
Fluid Dynamics
Fortran, MPI

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

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

Application Mapping Optimisation Cycle



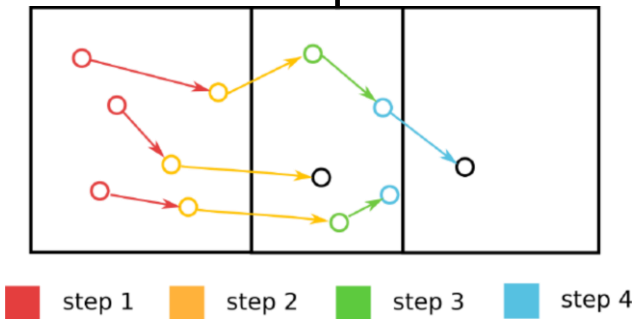
Based on material by A. Geiß, TU Darmstadt

Use Case: PATMOS

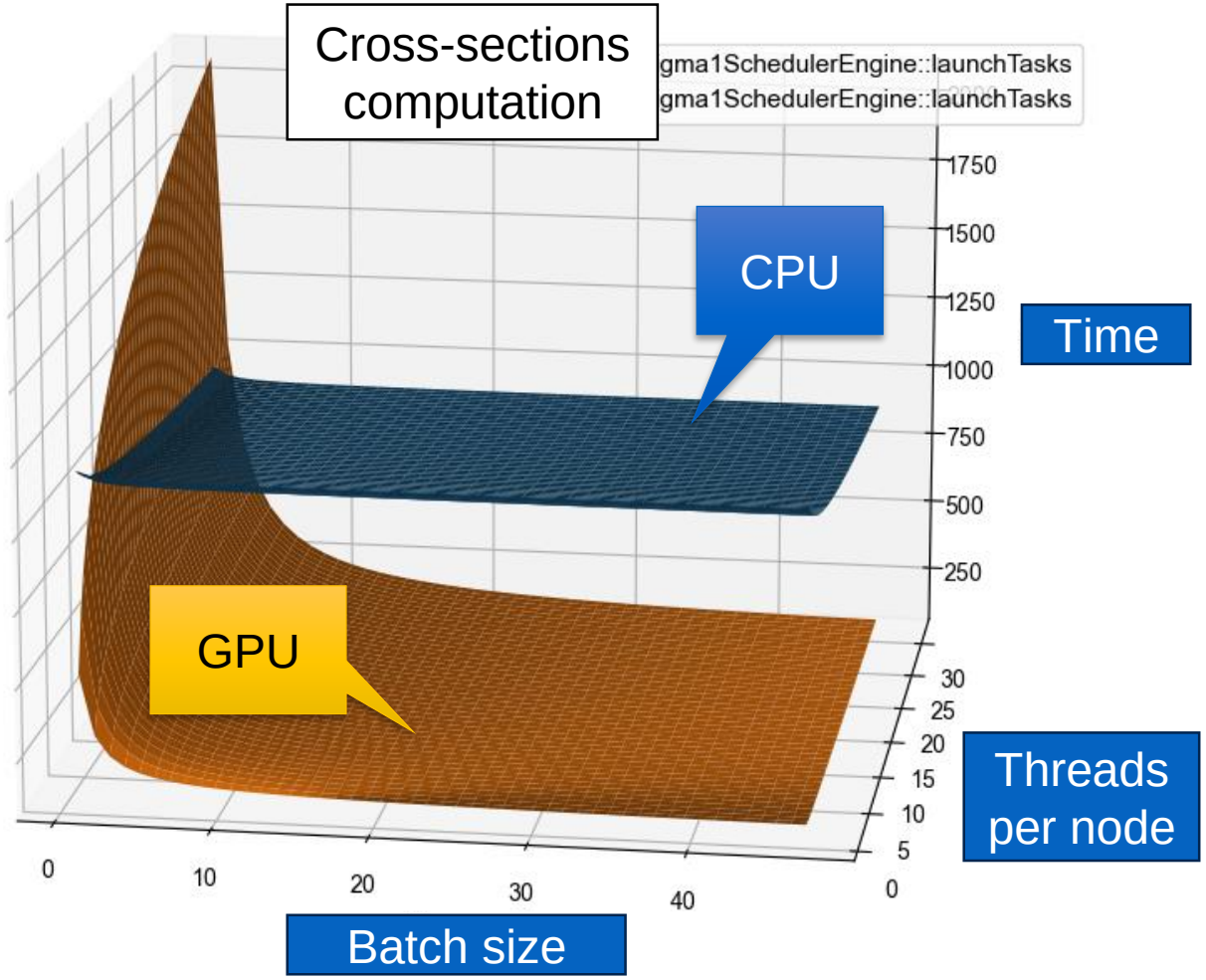
Solves the neutron transport equations to simulate evolution of physical quantities for complex systems

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

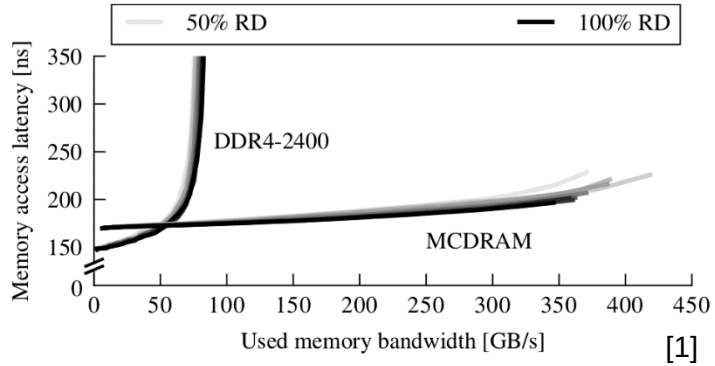


Based on material by T. Hovi, CEA

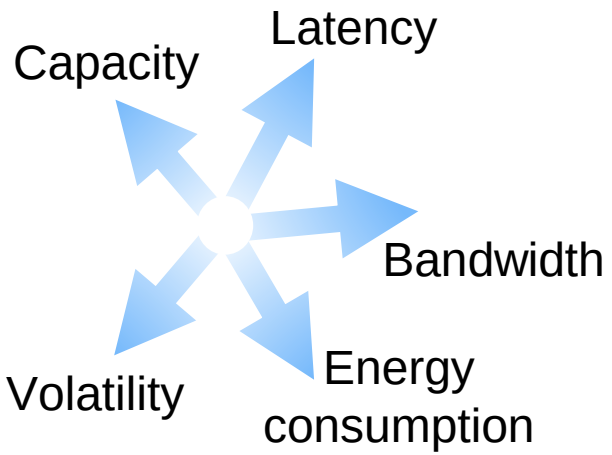
Heterogeneous/Hierarchical Memory

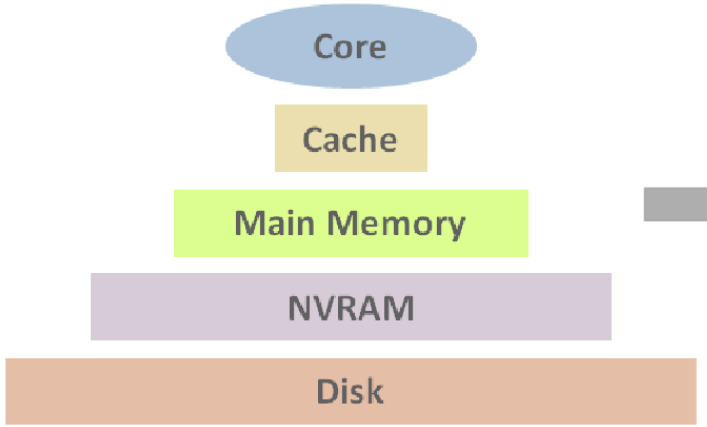
Examples…

- 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

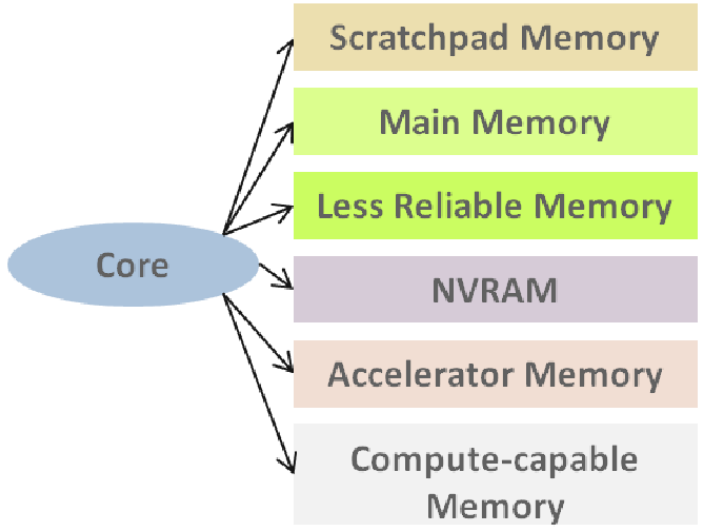


[1]





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

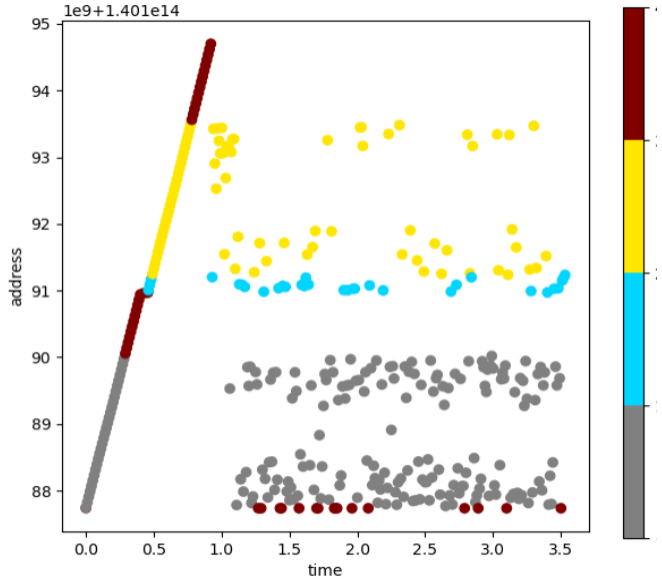
Transparent to the application

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)

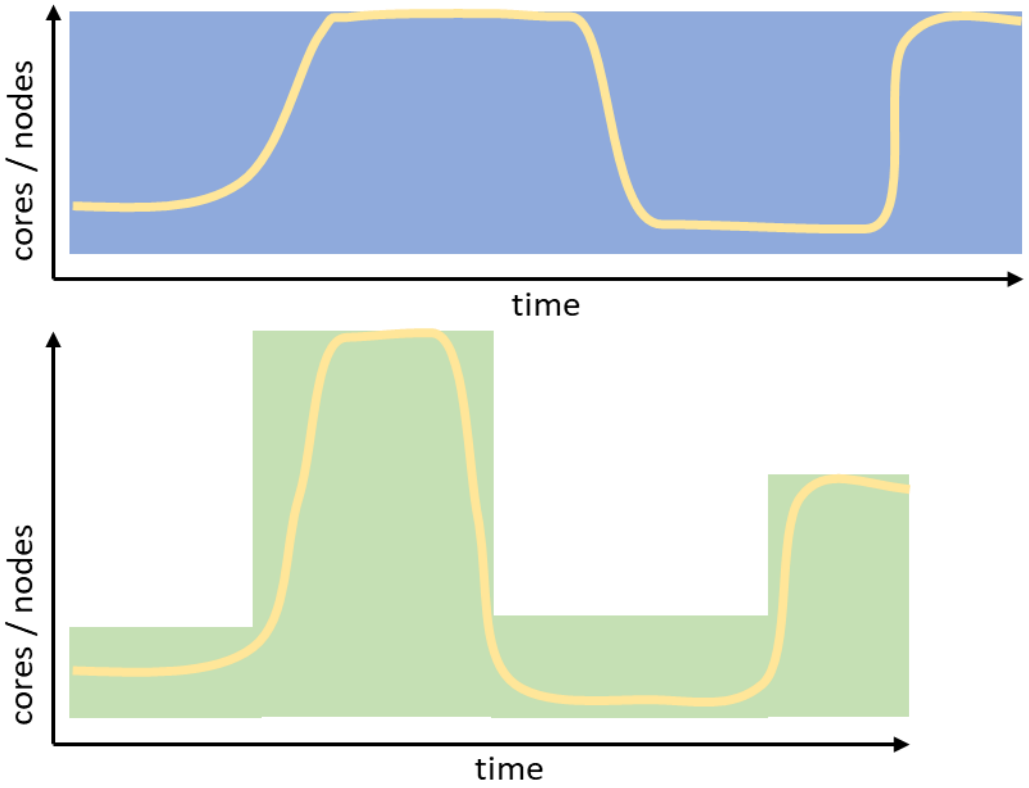
Ideal resource allocation for the workload in green

Malleable applications

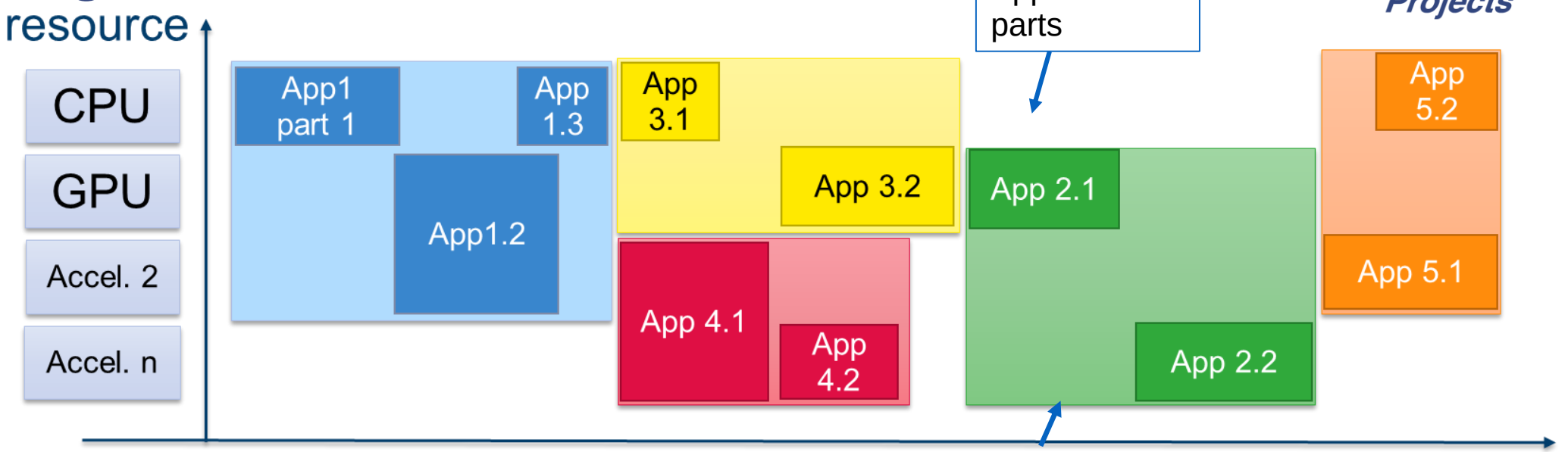
- Release resources not required
- Acquire more resources if advantageous

Change in # of nodes do require data redistribution in the workload

DEEP-SEA provides MPI & Slurm prototypes for enabling application-driven (active) malleability



Scheduling

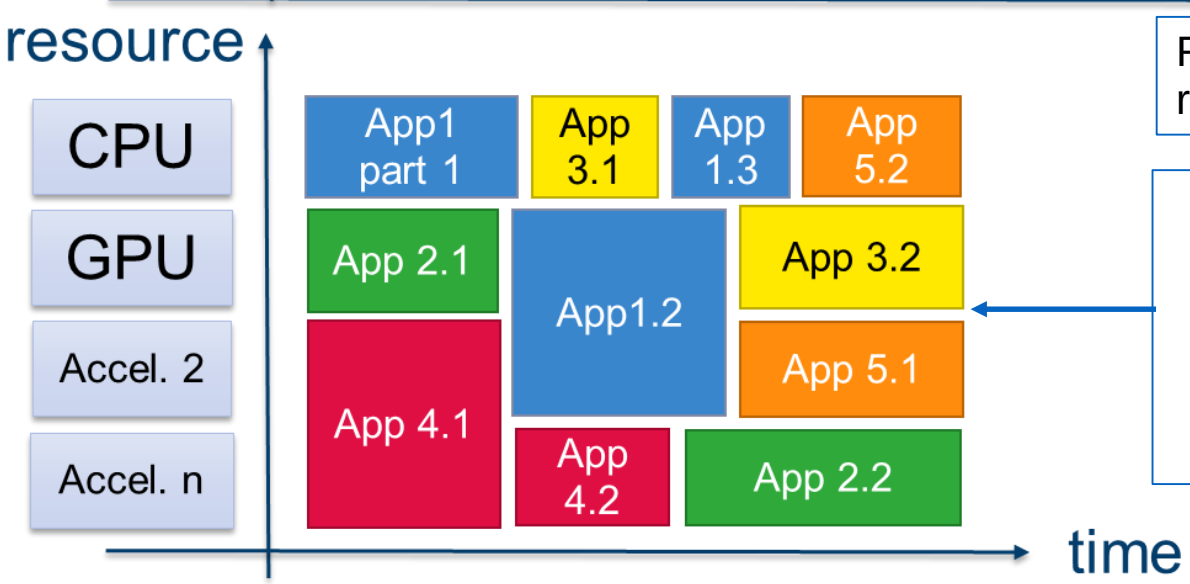


Application parts

Current behaviour

Resource reservation window

WIP



Ideal behaviour

Application part

=

Resource reservation window

Funding Acknowledgement

EuroHPC Joint Undertaking

SPONSORED BY THE

fwo

bpifrance

Federal Ministry of Education and Research

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

DEEP Projects

www.deep-projects.eu

@DEEPprojects