With MSA to ExaScale

Research on Parallel Architectures at JSC

14 SEPTEMBER 2023 I NORBERT EICKER I JSC & BERGISCHE UNIVERSITÄT WUPPERTAL

Mitglied der Helmholtz-Gemeinschaft

JÜLICH Forschungszentrum

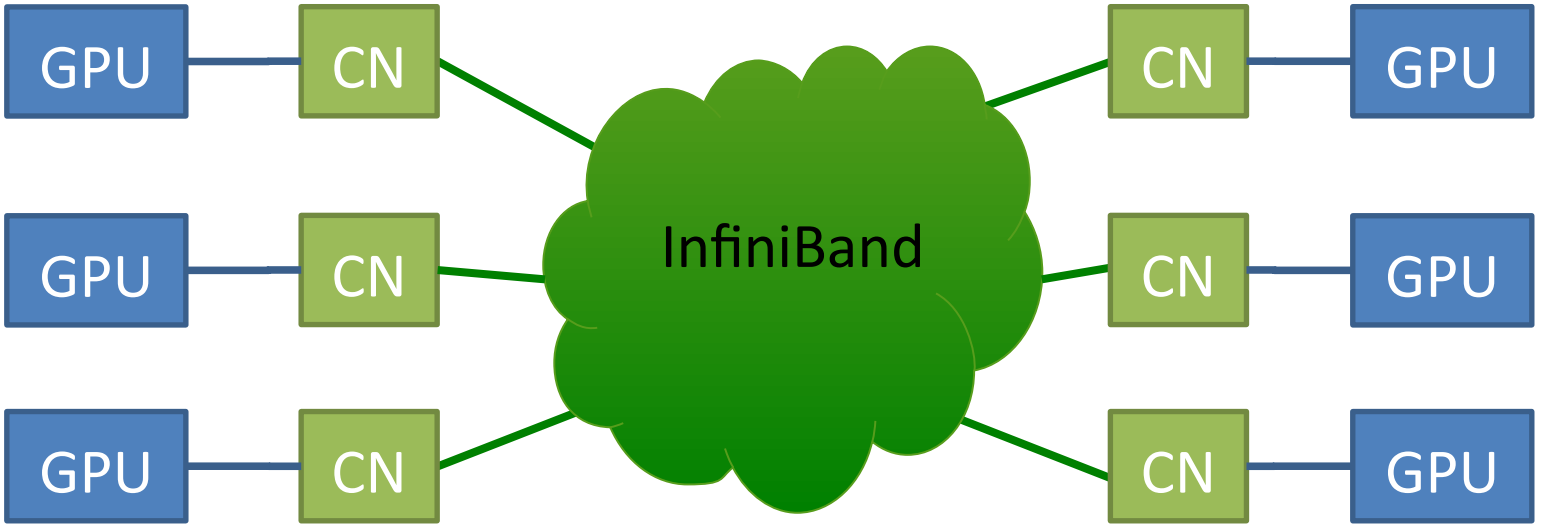
Shaping Change

Modular Supercomputing Architecture

is about

Organizing Heterogeneity

Heterogeneous Clusters



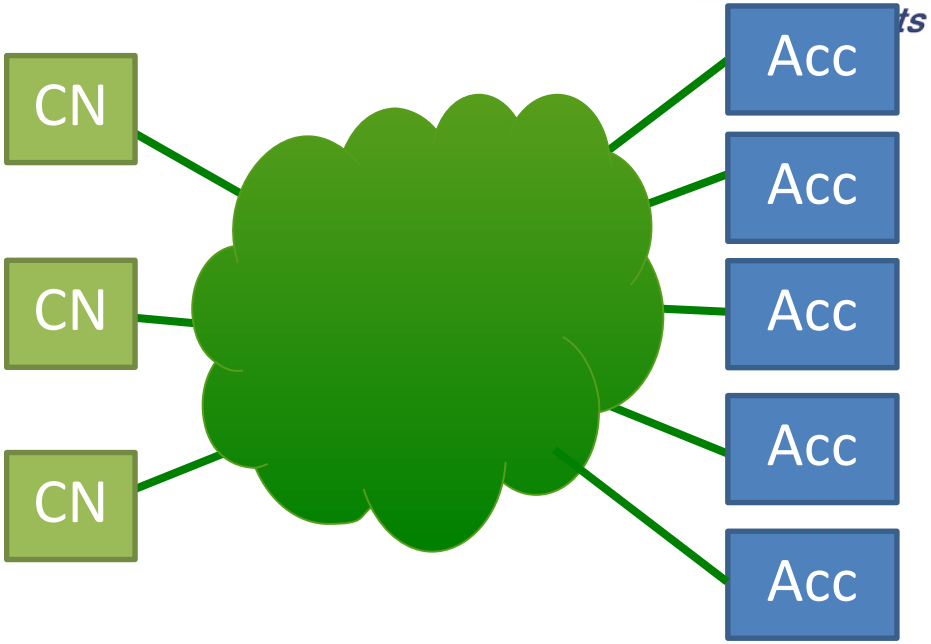
Flat IB-Topology

Simple Resourcemanagement

Static Assignment of CPU & GPU

Accelerator not autonomous

Alternative Integration



Go for more capable accelerators
- e.g. MIC

Attach all nodes to low-latency fabric

All nodes might act autonomously

Dynamical assignment of cluster-nodes and accelerators

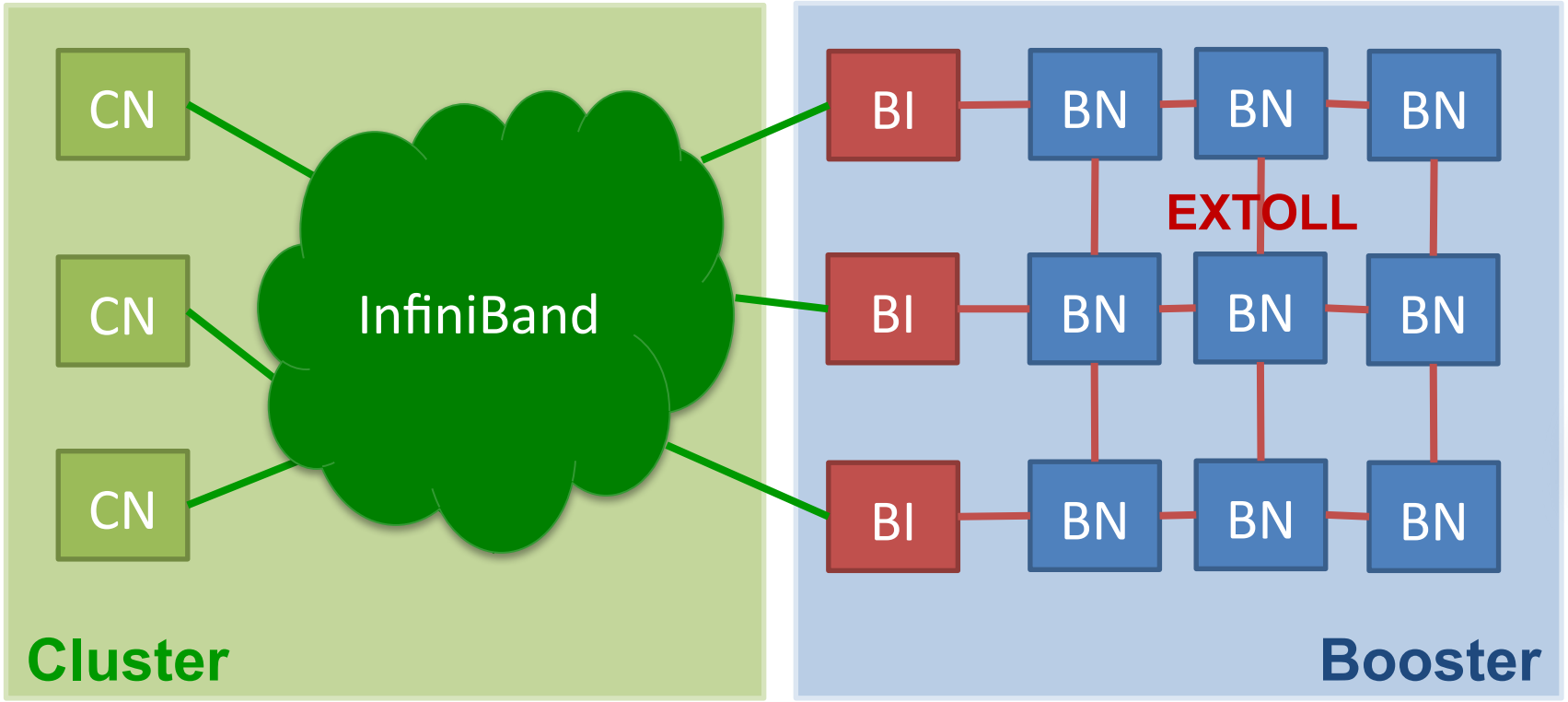
Accelerator bandwidth limited by PCIe

Ability to off-load more complex kernels (including communication)

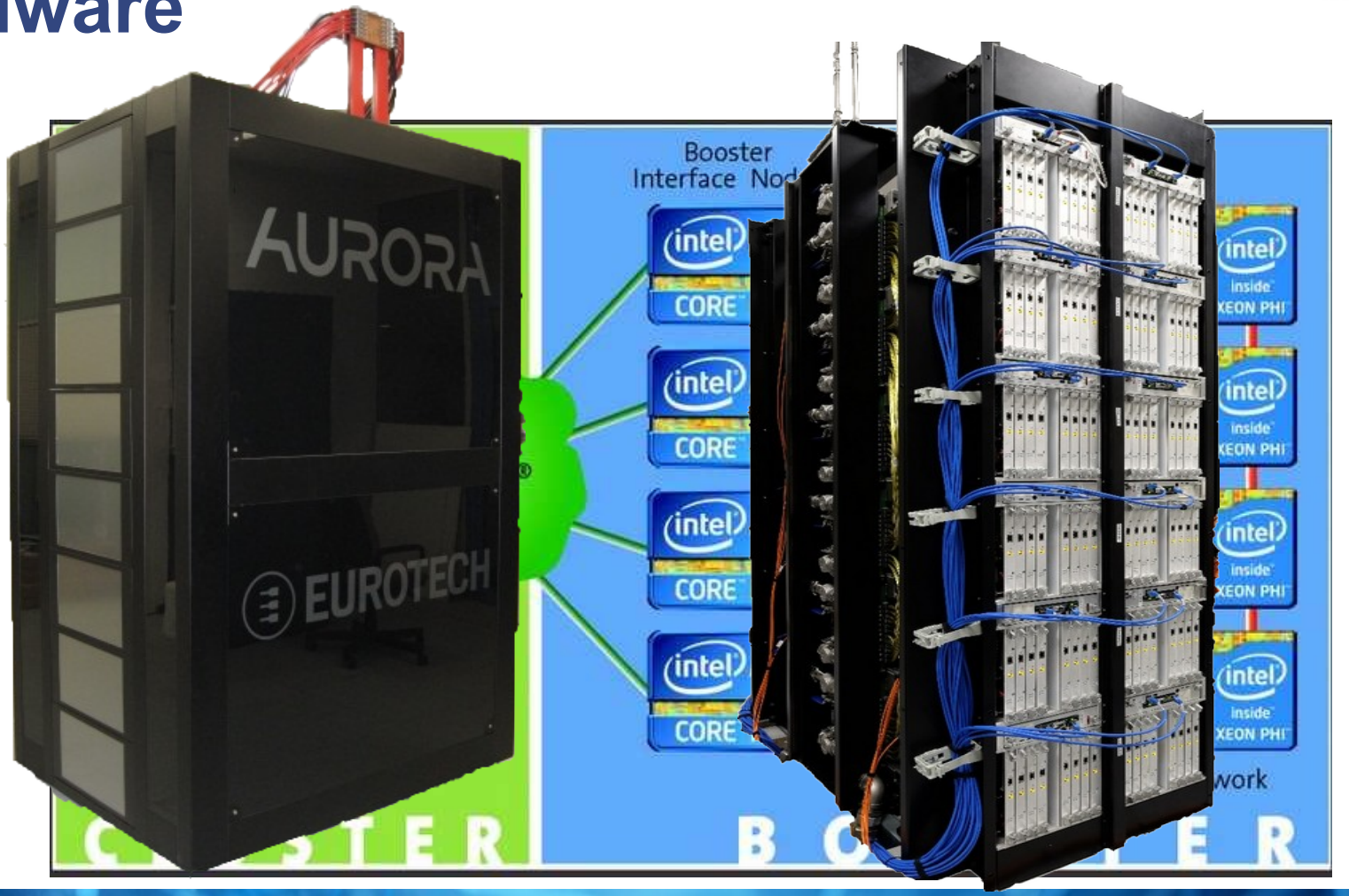
Communication between CPU and Accelerator less frequently
⇒ larger messages ⇒ less sensitive to latency

DEEP Cluster-Booster Architecture

The first ever CBA realized in the DEEP projects (Dynamical Exascale Entry Platform)



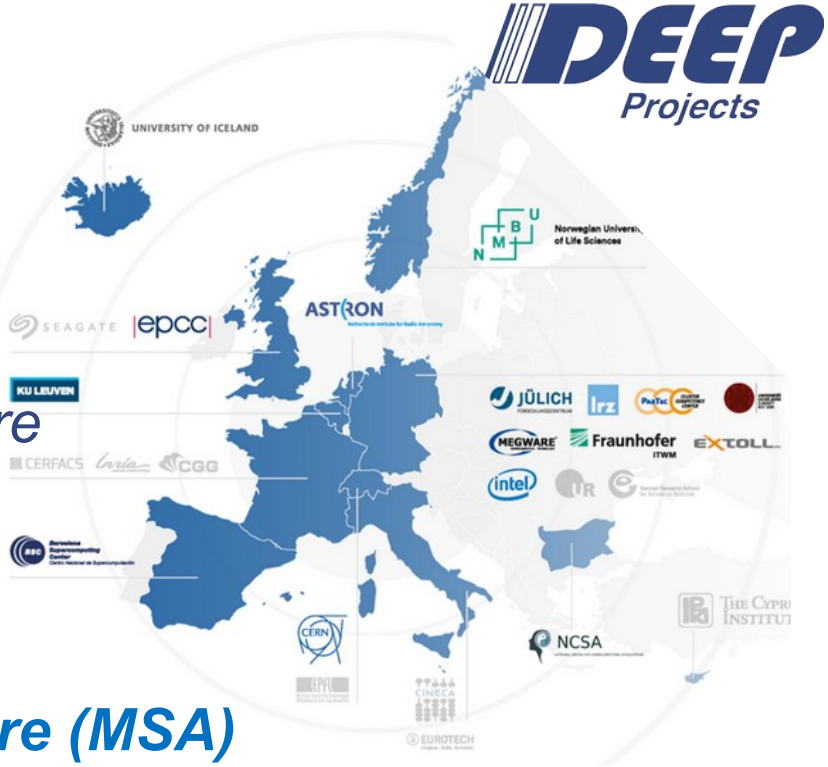
Hardware


AURORA
EUROTECH
Booster
Interface Nod
intel
CORE
intel
inside
XEON PHI
work

Past projects

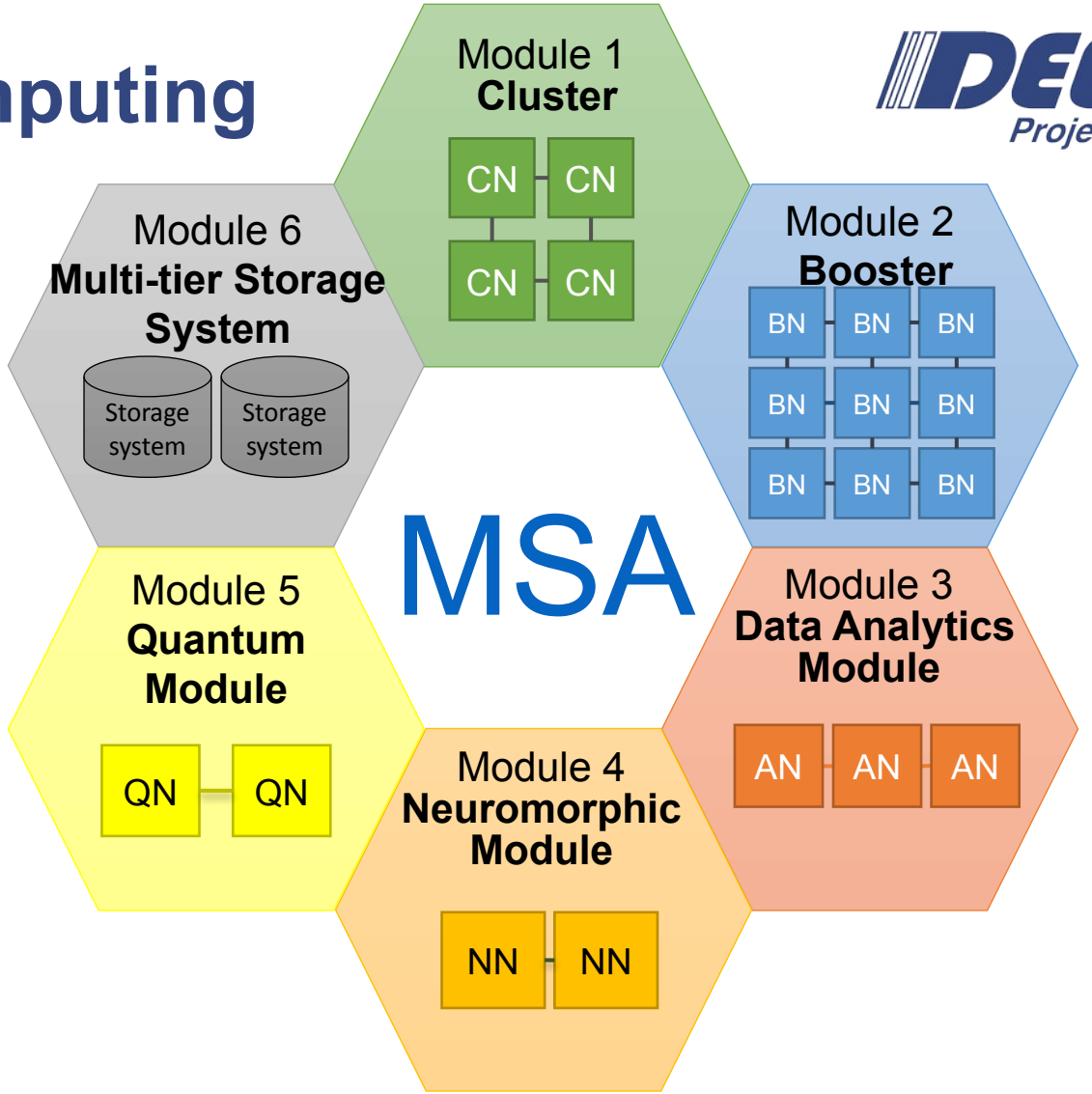
2011-2021 DEEP Projects

- DEEP (2011 – 2015)
  - *Introduce **Cluster-Booster** architecture*
- DEEP-ER (2013 – 2017)
  - *Add I/O and resiliency functionalities*
- DEEP-EST (2017 – 2021)
  - ***Modular Supercomputer Architecture (MSA)***

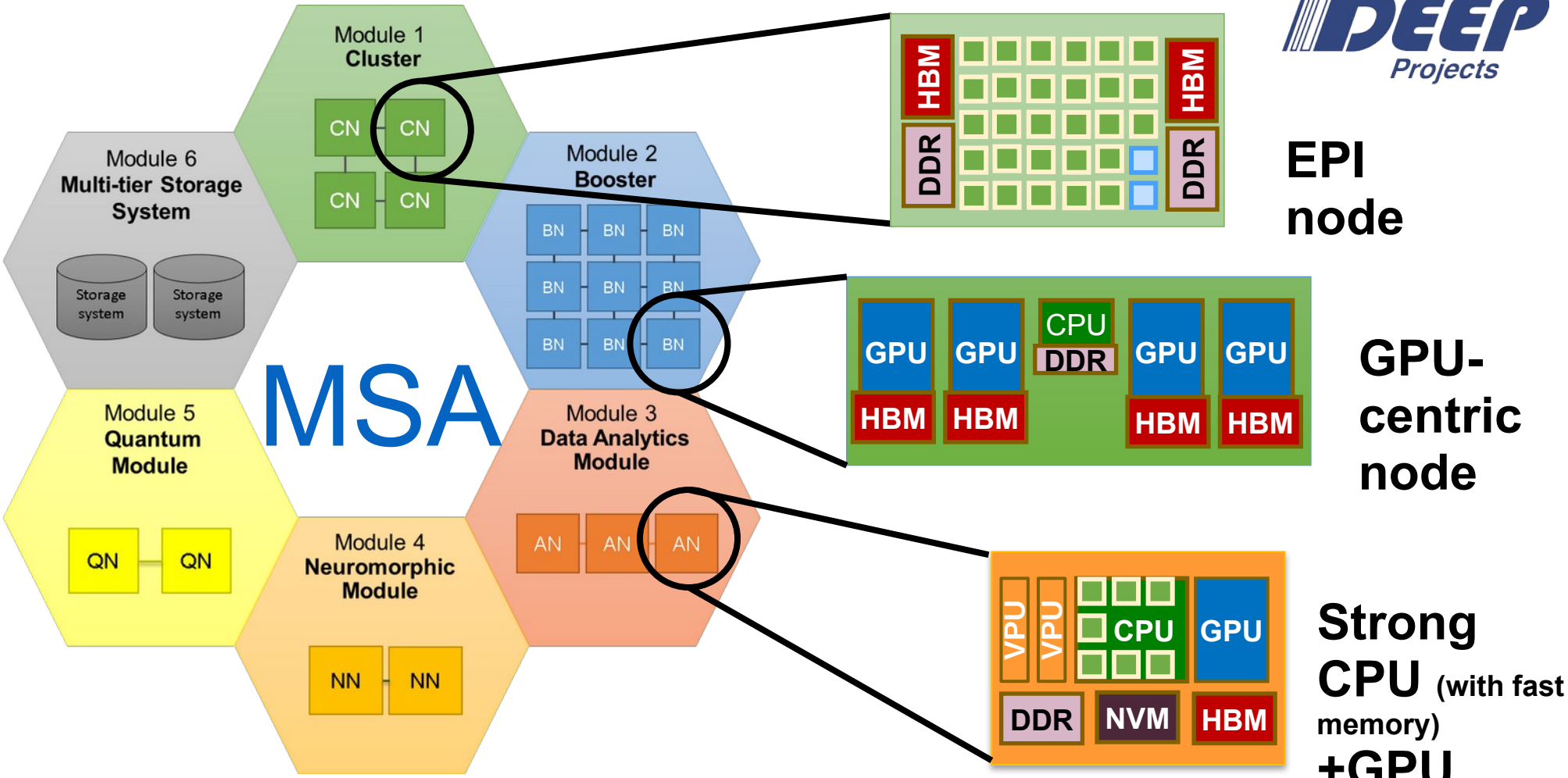


Modular Supercomputing

Composability of heterogeneous resources



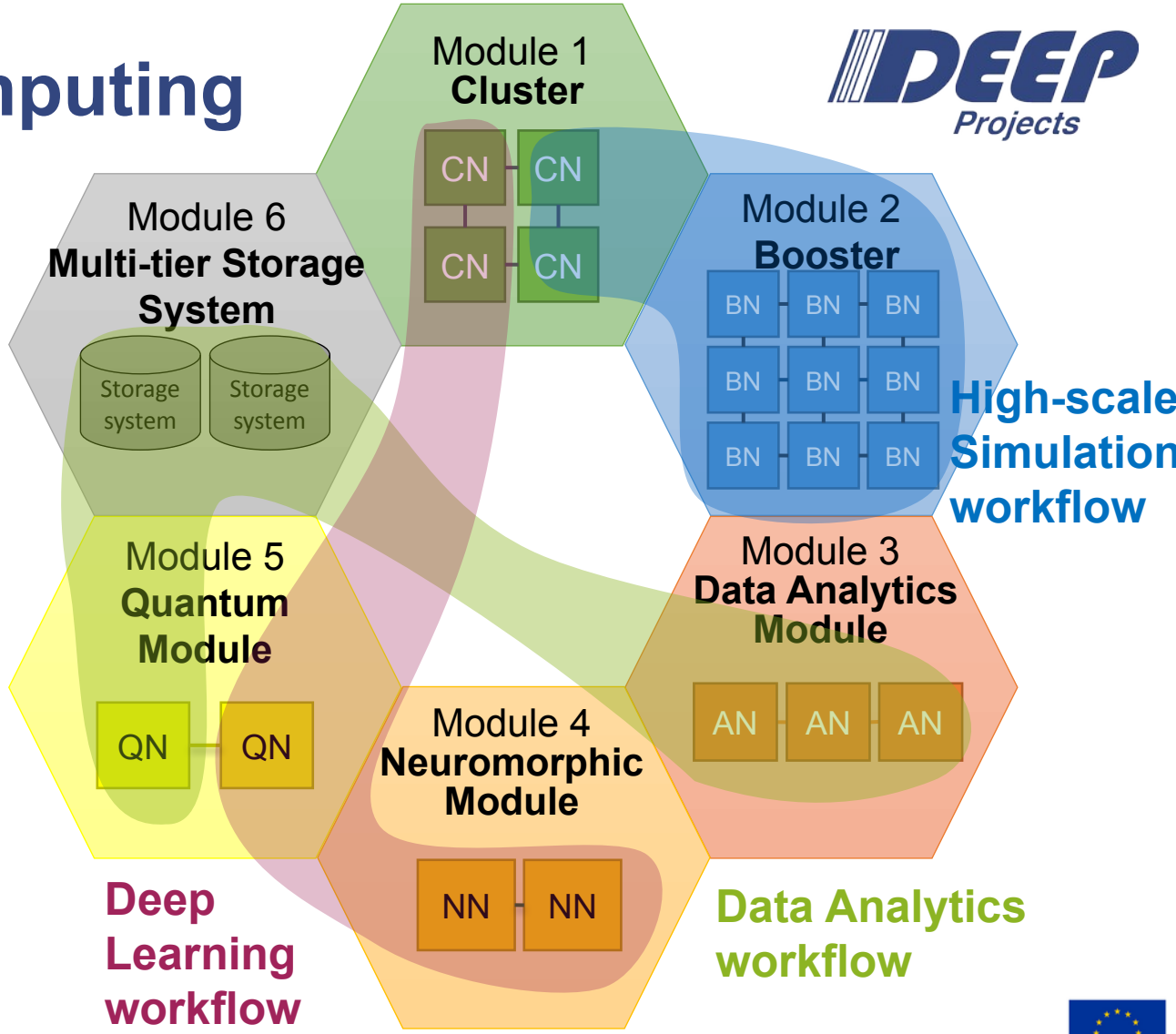
- E. Suarez, Th. Lippert, N.E., "*Modular Supercomputing Architecture: from idea to production*", Chapter 9 in Contemporary High Performance Computing: from Petascale toward Exascale, Volume 3, pp 223-251, CRC Press. (2019)
- E. Suarez, Th. Lippert, N.E., "Supercomputer Evolution at JSC", Proceedings of the 2018 NIC Symposium, Vol.49, p.1-12, (2018) [online: http://juser.fz-juelich.de/record/844072].


Module 1
Cluster
CN
CN
CN
CN
Module 6
Multi-tier Storage System
Storage system
Storage system
Module 2
Booster
BN
BN
BN
BN
BN
BN
BN
BN
BN
MSA
Module 5
Quantum Module
QN
QN
Module 4
Neuromorphic Module
NN
NN
Module 3
Data Analytics Module
AN
AN
AN
HBM
DDR
HBM
DDR
GPU
GPU
CPU
DDR
GPU
GPU
HBM
HBM
HBM
HBM
VPU
VPU
CPU
GPU
DDR
NVM
HBM
EPI node
GPU-centric node
Strong CPU (with fast memory) +GPU

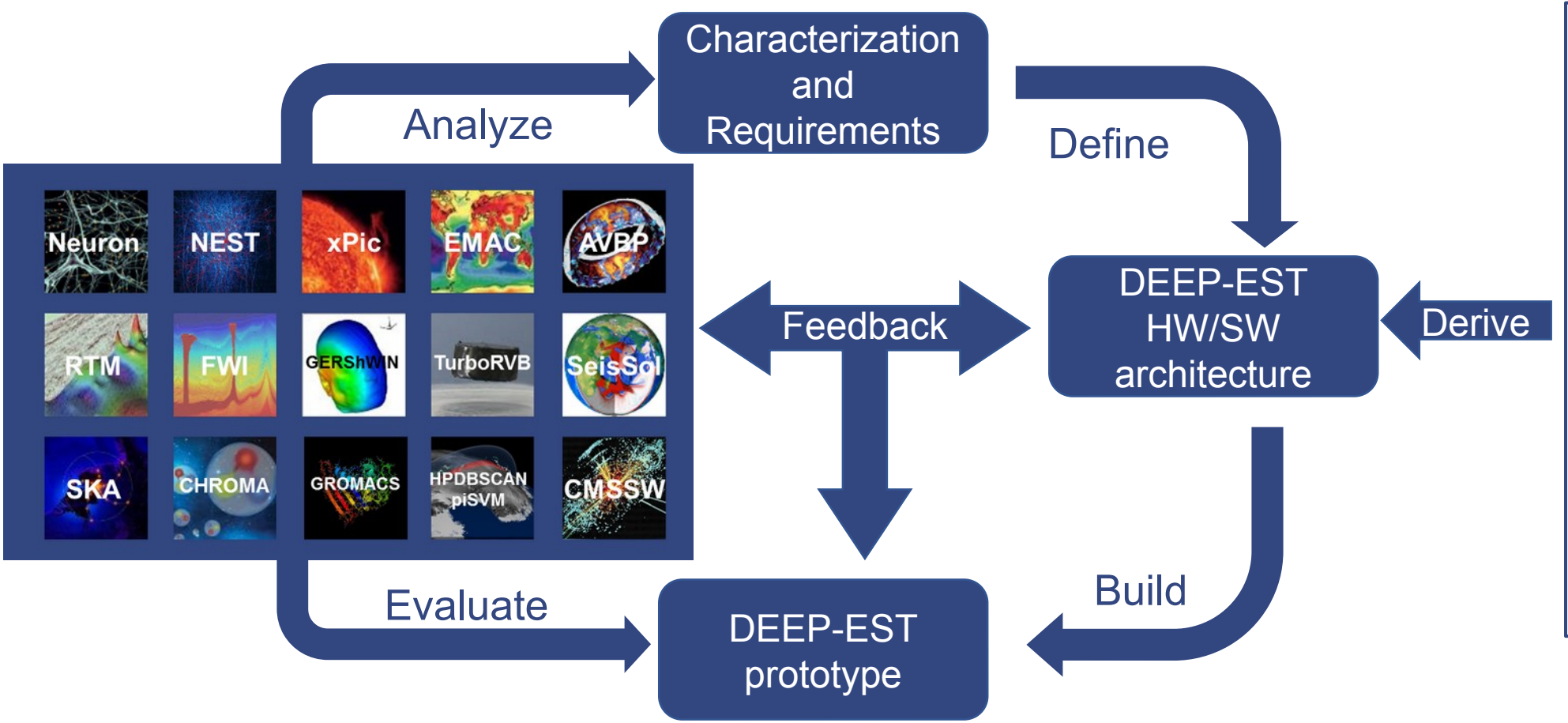
Modular Supercomputing

- Fit application diversity
  - Large-scale simulations
  - Data analytics
  - Machine- and Deep Learning
  - Artificial Intelligence

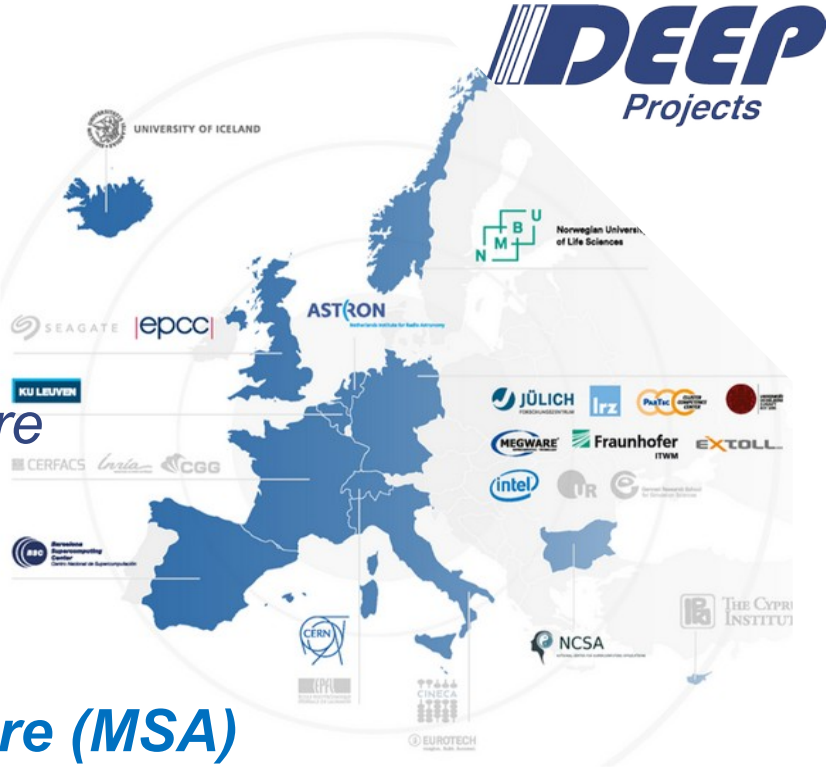
Composability of heterogeneous resources



Co-design Approach



Ursprünge



2011-2021 DEEP Projekte

- DEEP (2011 – 2015)
  - *Introduce **Cluster-Booster** architecture*
- DEEP-ER (2013 – 2017)
  - *Add I/O and resiliency functionalities*
- DEEP-EST (2017 – 2021)
  - ***Modular Supercomputer Architecture (MSA)***

2021-2024 SEA Projekte

- **DEEP-SEA**, IO-SEA, RED-SEA

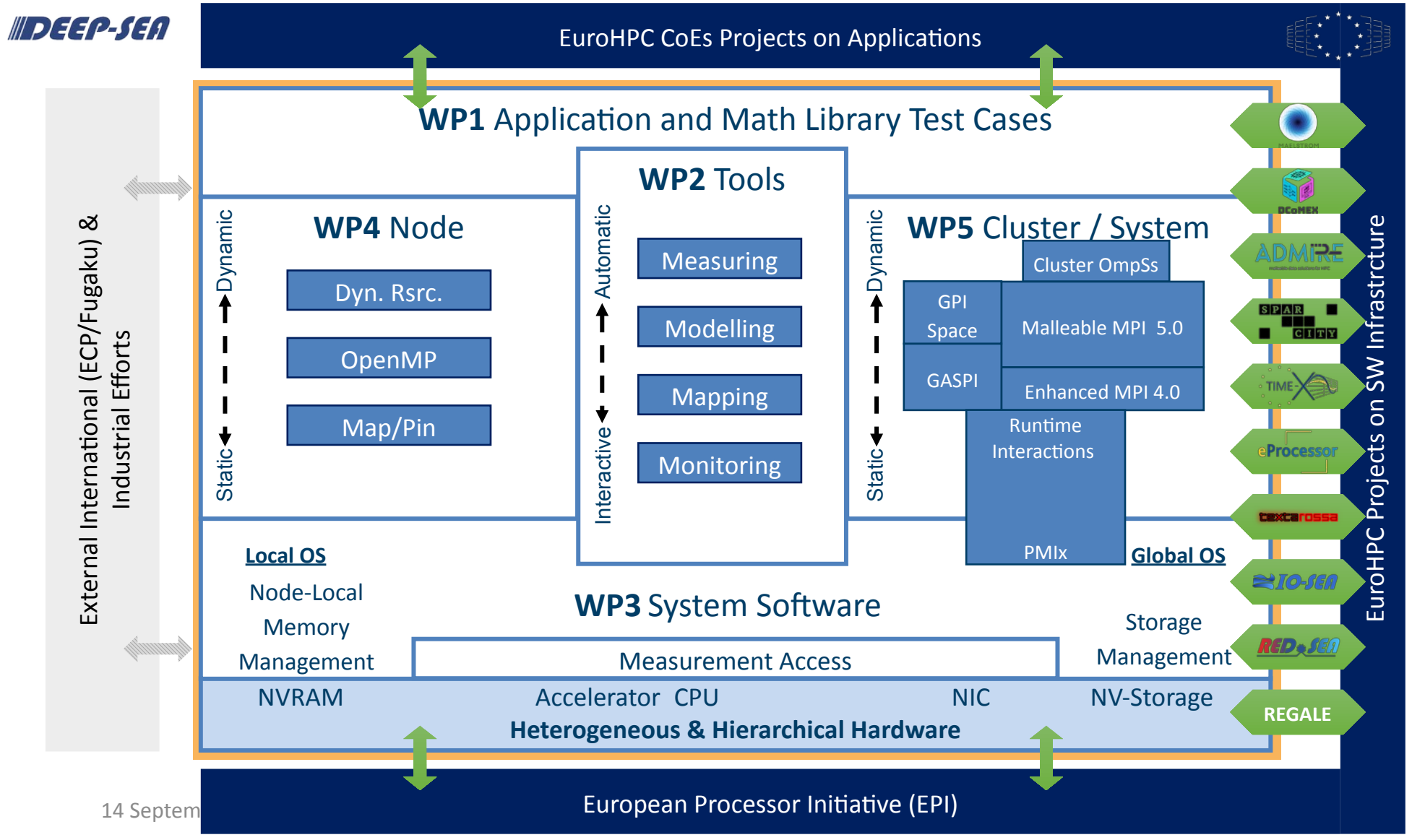
SEA Projects

Motivation

Heterogeneity in HPC is there an will not go away
- in processor, memory and network
- Different codes run better on different components
- Programming models might be even vendor-specific

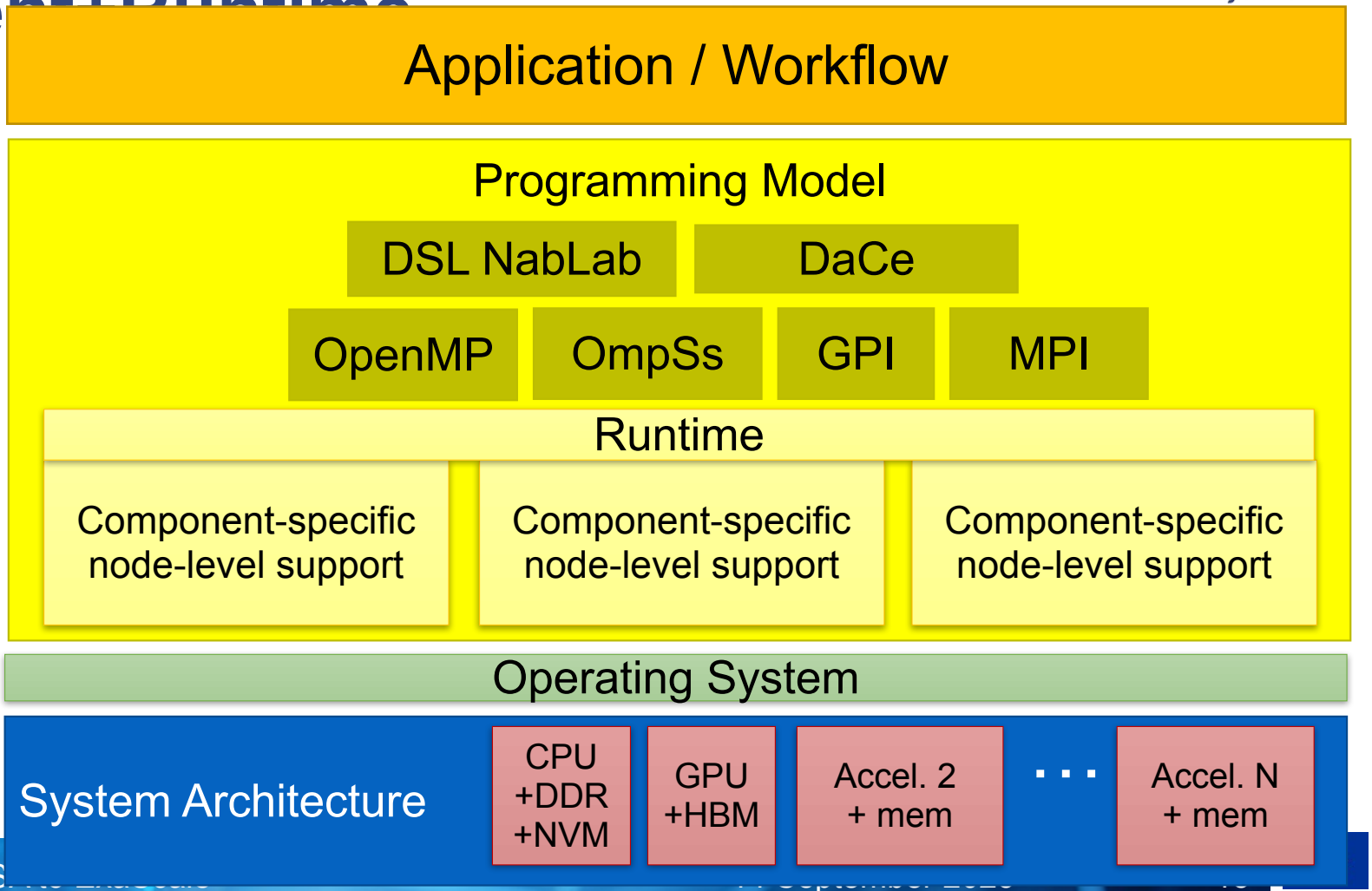
Burden on application developers, who must:
- Understand very well their codes **and** the HW
- Decide where to run each part of their codes
- Port codes to different technologies
- Optimize codes for different platforms

Goal: ease and/or automatize some of these tasks
- Provide solutions with high TRL (near production)


DEEP-SEA
EuroHPC CoEs Projects on Applications
WP1 Application and Math Library Test Cases
External International (ECP/Fugaku) & Industrial Efforts
WP4 Node
Dynamic
Static
Dyn. Rsrc.
OpenMP
Map/Pin
WP2 Tools
Automatic
Interactive
Measuring
Modelling
Mapping
Monitoring
WP5 Cluster / System
Dynamic
Static
Cluster OmpSs
GPI Space
GASPI
Malleable MPI 5.0
Enhanced MPI 4.0
Runtime Interactions
PMIx
Local OS
Node-Local Memory Management
Global OS
WP3 System Software
Storage Management
Measurement Access
NVRAM
Accelerator CPU
NIC
NV-Storage
Heterogeneous & Hierarchical Hardware
MAELSTROM
DCoMEX
ADMIRE
SPARCITY
TIME-X
ePROCESSOR
textarossa
IO-SEA
RED-SEA
REGALE
EuroHPC Projects on SW Infrastrcture
European Processor Initiative (EPI)

OS+Programming environment+Runtime

- **Support for accelerators & memory**
- **Malleability**
- **Interoperability**
- **Composability**
- **Performance portability**
- **Resiliency**



Enabling Applications

Skalierbarkeit, innovative Methoden sowie Architekturen für zukünftige Exascale-Systeme

Weiterentwicklungen des ICON Erdsystem-Modells

ExaOcean

Leistungsverbesserung des ICON-O Ozeanmodells auf heterogenen Exascale-Supercomputern mit Methoden des Maschinellen Lernens

Daniel Ruprecht, Peter Korn, Christopher Kadow, Lars Hoffmann

Philip Freese, Fabricio Lapollo, Maximilian Witte, Yen-Sen Lu

IFCES2

Optimierung von Simulationsalgorithmen für Exascale-Supercomputer zur Berechnung des Erdsystemmodells ICON

Fabian Senf, Panagiotis Adamidis, Estela Suarez, Nobert Eicker,Daniel Klocke, Carsten Clauss, Matthias Lieber, Wolfgang E. Nagel

Roxana Cremer, Xingran Wang, Manoel Römmer, Fatemeh Chegini, Simon Pickartz, Sonja Happ, Johann Biedermann

SPONSORED BY THE

Federal Ministry of Education and Research

Gefördert vom Bundesministerium für Bildung und Forschung (BMBF) unter Förderkennzeichen 16ME0679K und 16ME0689K.

Partner von ExaOcean & IFCES2

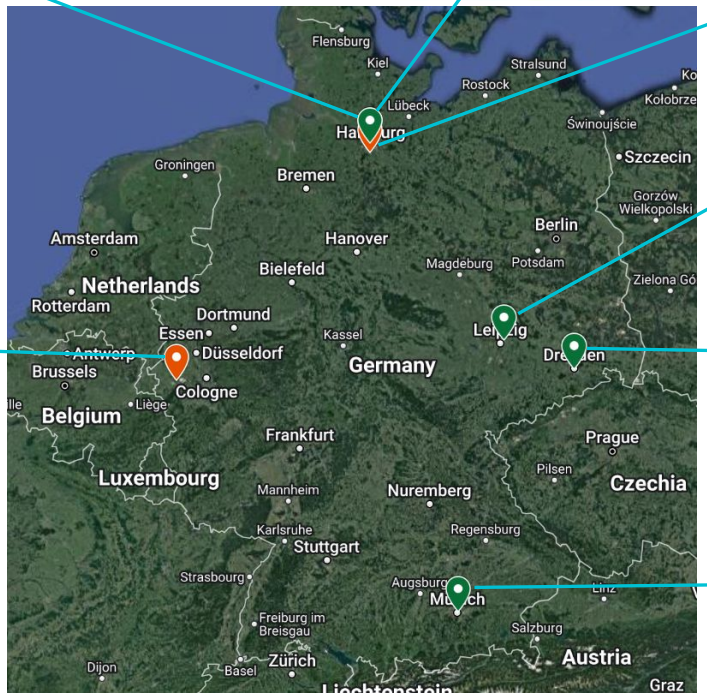
TUHH

TROPOS
Leibniz Institute for Tropospheric Research

MAX-PLANCK-INSTITUT FÜR METEOROLOGIE

DKRZ
DEUTSCHES KLIMARECHENZENTRUM

TUHH
Technische Universität Hamburg



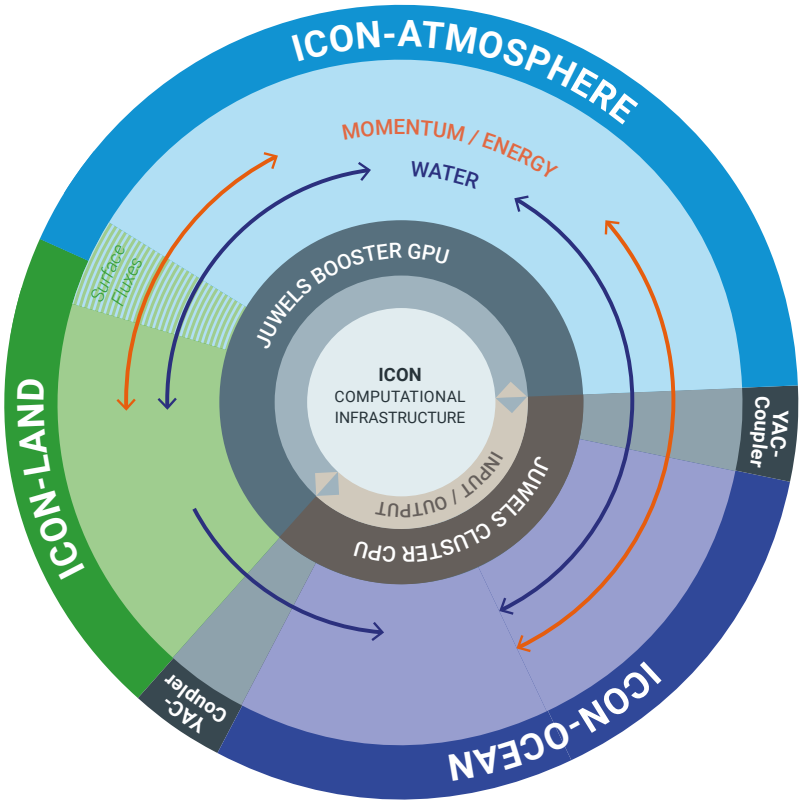
TROPOS
Leibniz Institute for Tropospheric Research

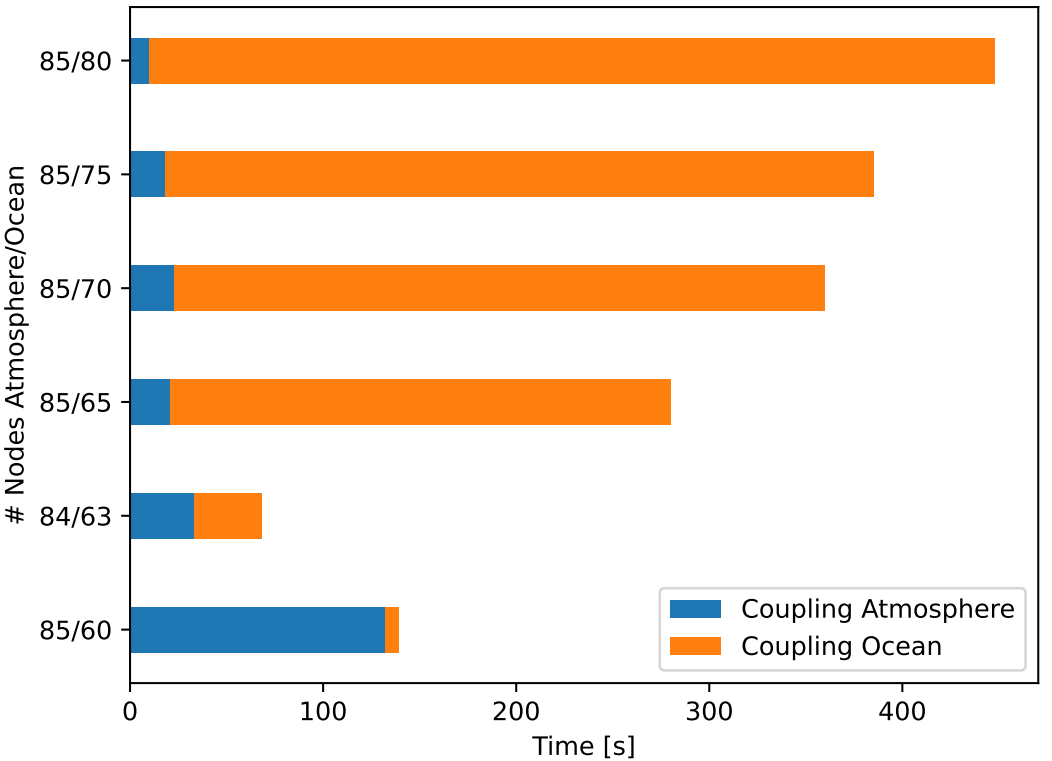
JÜLICH
Forschungszentrum

TECHNISCHE UNIVERSITÄT DRESDEN

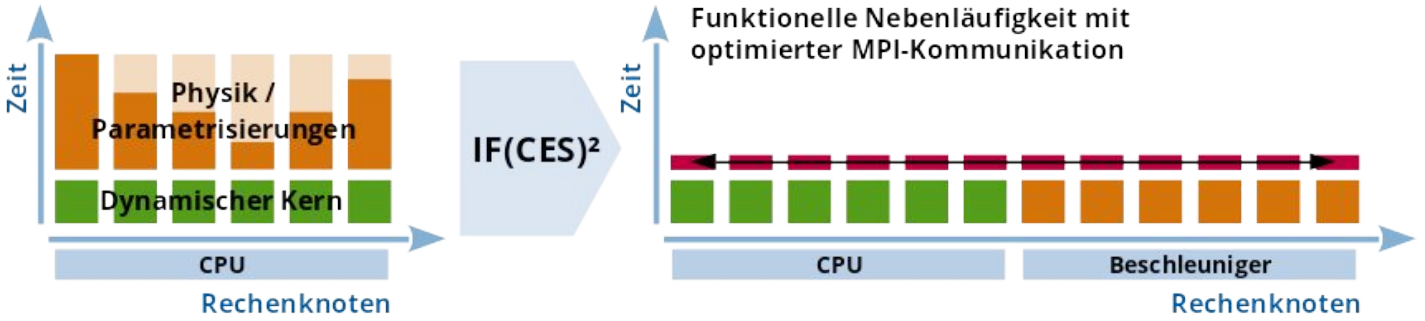
ParTec
MODULAR SUPERCOMPUTING

ICON MSA study

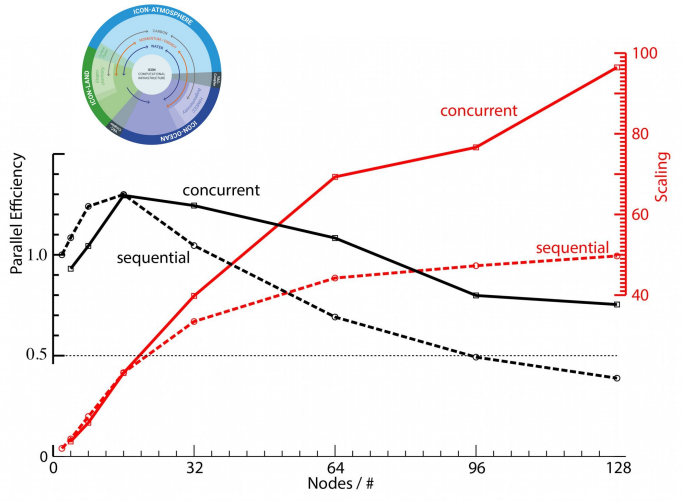




3. Functional Concurrency



- Functional Concurrency ermöglicht Modellkomponenten besser zu verteilen und der Heterogenität der zukünftigen HPC-Systeme Rechnung zu tragen
  ⇒ **verfeinerte Auflösung**
- smarte Kommunikation essenziell

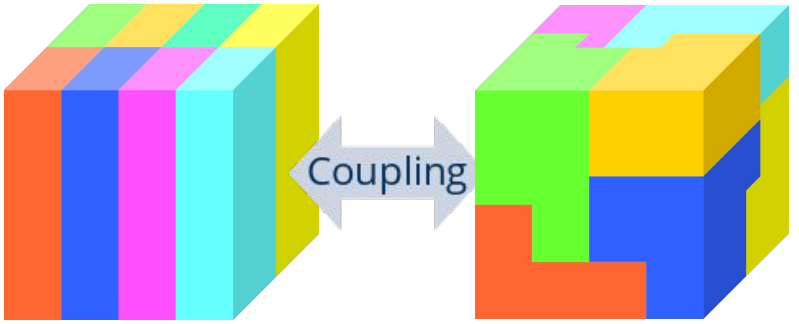


Linardakis et al. 2022. https://doi.org/10.5194/gmd-2022-214

4. Last-Balancierung für flexible Domänenaufteilungen

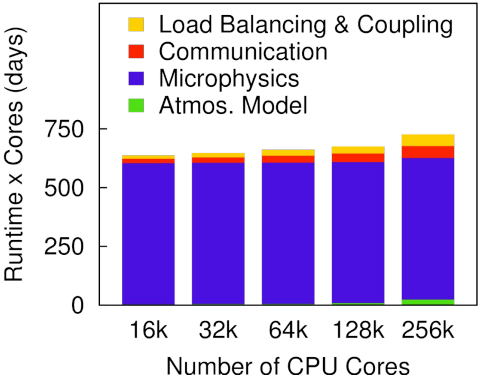
TUHH

- viele Atmosphärenphänomene (z.B. Wolken oder Atmosphärenchemie) benötigen nur an bestimmten Orten einen hohen Detailgrad
  ⇒ **verbesserte Adaptivität**
- Last-Balancierung kann völlig neue Wege zur Beschreibung komplexer und detailreicher Atmosphären-Prozesse ermöglichen
- effiziente Kommunikation essenziell



statische 2D Gebietszerlegung

dynamische Gebietszerlegung



Strong scaling benchmark on BlueGene/Q with 1024 x 1024 grid

MSA in Production

Today & Tomorrow

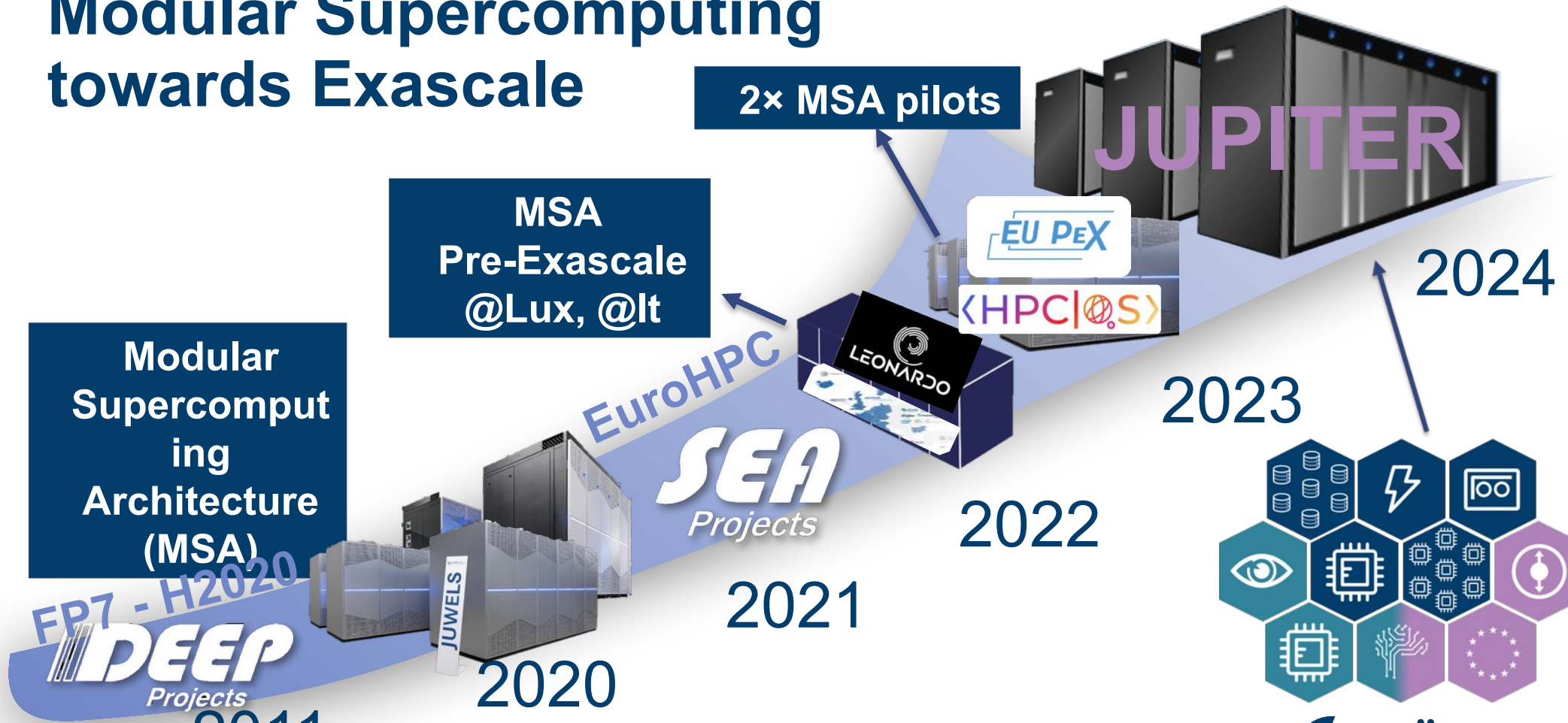
Modular supercomputer JUWELS

JUWELS Cluster	JUWELS Booster
Intel Xeon (Skylake) processor	AMD EPYC Rome 7402 processor
InfiniBand EDR network	3,700 NVIDIA A100 GPUs
2,500 compute nodes	InfiniBand HDR DragonFly+
10 PFLOP/s peak (CPU-based)	**70 PFLOP/s peak** (GPU-based)

JUWELS is a Blueprint for JUPITER

(modulo specific hardware vendors and generation)

Modular Supercomputing towards Exascale



JUPITER – Timeline

17.12.2021: Call for Expression of Interest (EoI) for Hosting Entity

14.02.2022: Deadline EoI Submission

16.05.2022: Hearings

15.06.2022: Hosting site decision & announcement

14.12.2022: Signature hosting agreement

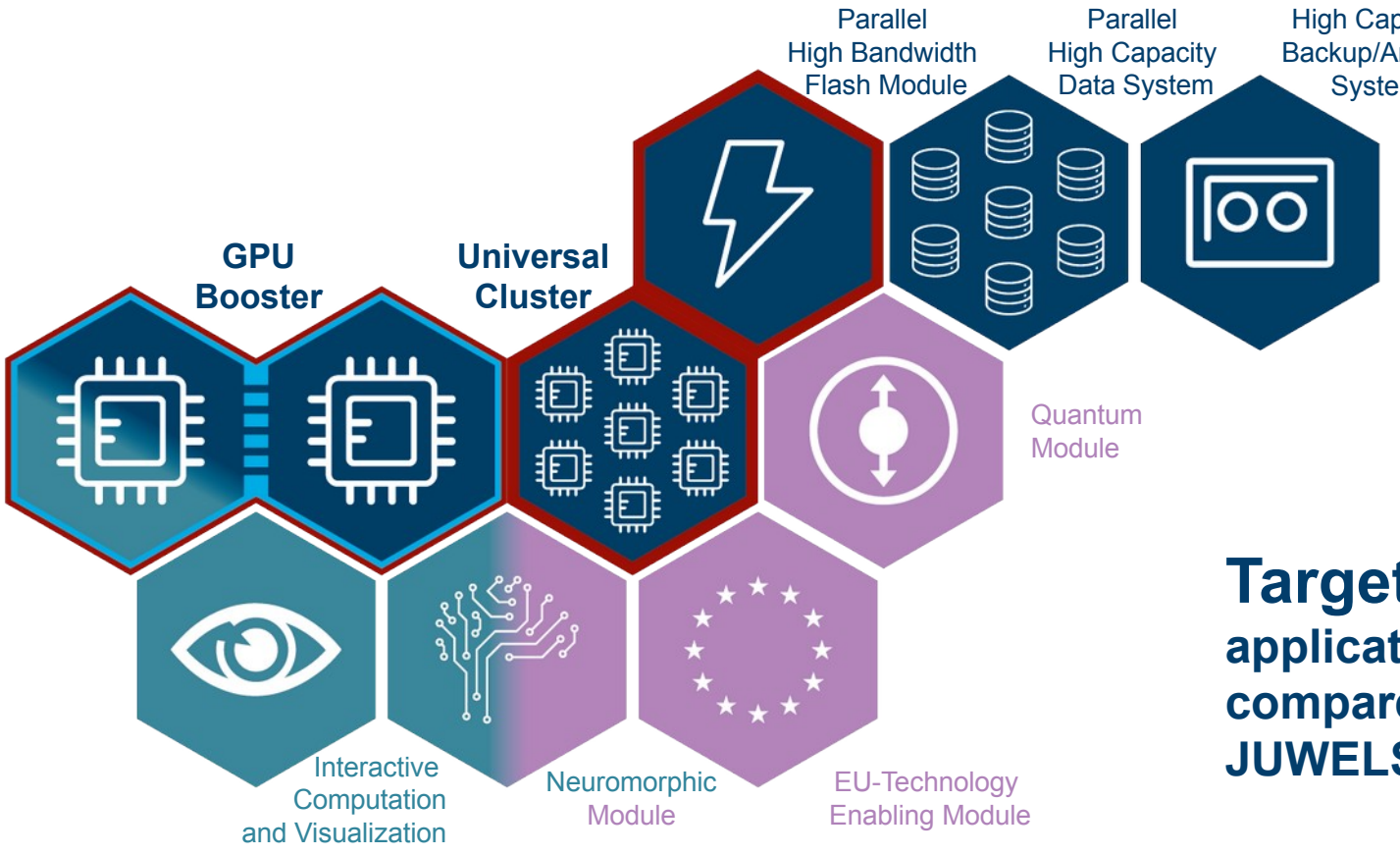
Q1-Q3 2023: Procurement (Competitive Dialogue)

Q1/2024: Start installation of JUPITER

End of 2024: Put in operation JUPITER

The acquisition and operation of the EuroHPC supercomputer is funded jointly by the EuroHPC Joint Undertaking, through the European Union's Digital Europe programme, as well as by Germany through the BMBF and the MKW.

JUPITER – Architecture and Options



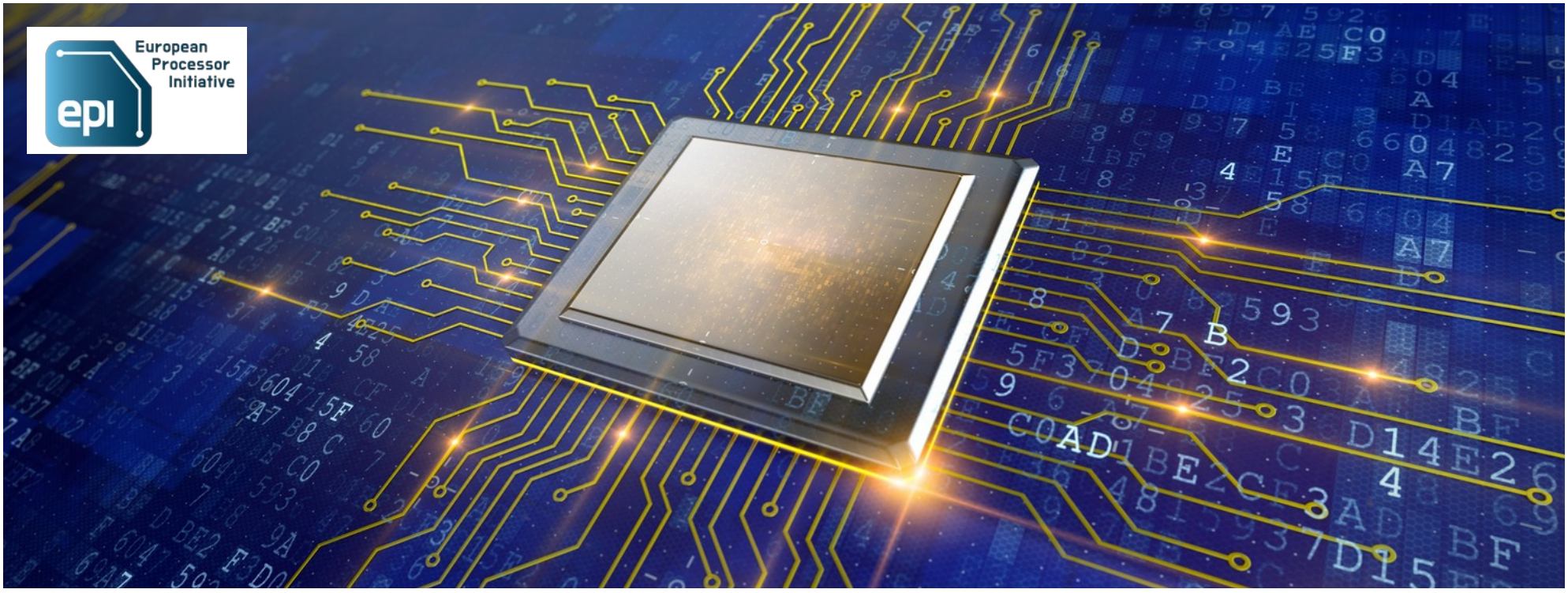
Basic Configuration

Optional Modules

Future Technology Modules

Target >20×
application performance compared to JUWELS Booster

Research on HW building blocks



The European Processor Initiative

EPI Consortium

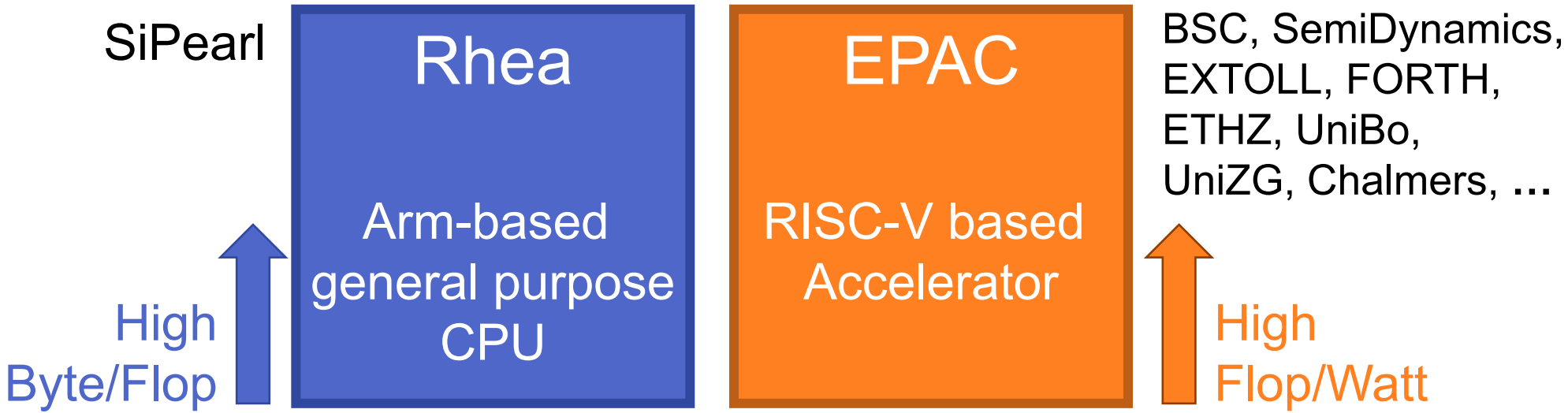
Work presented here has been performed by the EPI Consortium:

- www.european-processor-initiative.eu
- @EuProcessor
- European Processor Initiative
- European Processor Initiative

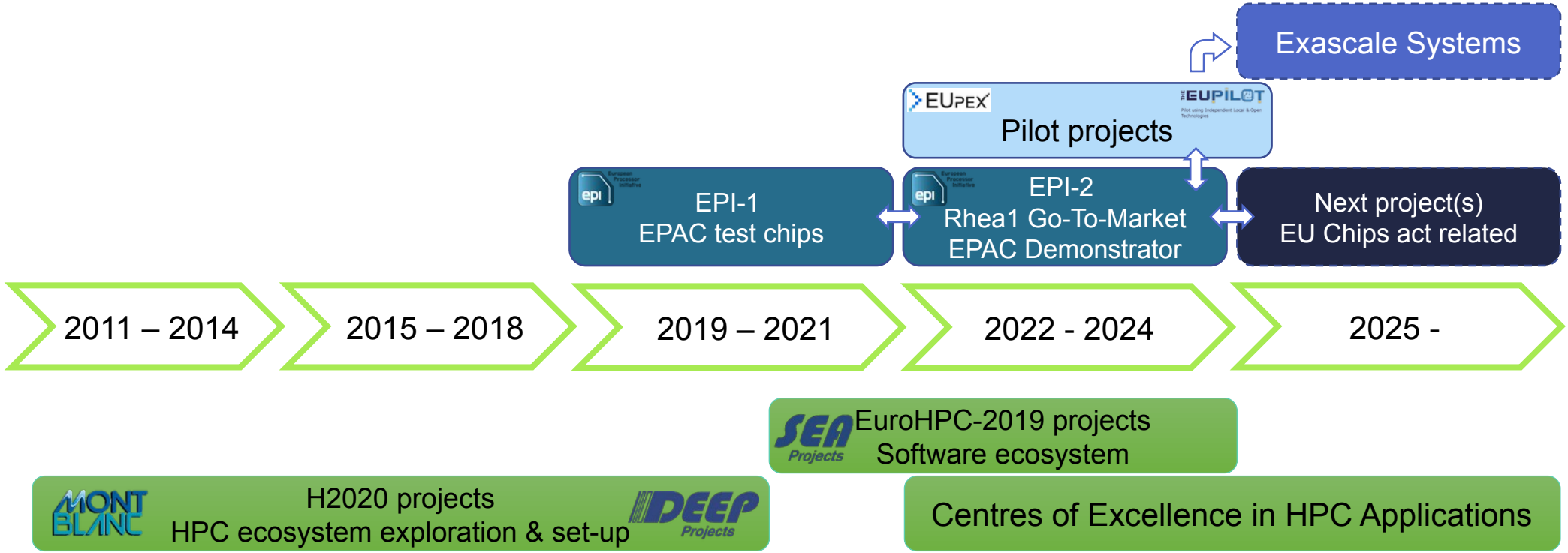
Main Objective

Develop European **microprocessor and accelerator** techn.

- Strengthen competitiveness of EU industry and science



Overall Technology Roadmap



Summary

Broad spectrum of activities towards Parallel Architectures

- Develop, Implement and Assess new concepts
- Deploy into Production Systems

  DEEP ⇒ JURECA ⇒ JUWELS ⇒ JUPITER

Keep the Software Stack in the focus

- All types of Middleware and Tools need to be enahnced

Support the Application Development

- Participation in dedicated projects, Simulation Labs

Prepare for integration of future (uncommon) Hardware

- Quantum (⟨HPC|QS⟩), Neuromorphic, ???