DEEP Projects

Towards a Modular Supercomputing Architecture for Exascale

Estela Suarez, Jülich Supercomputing Centre (JSC)

30.06.2020 – ESiWACE Virtual Workshop

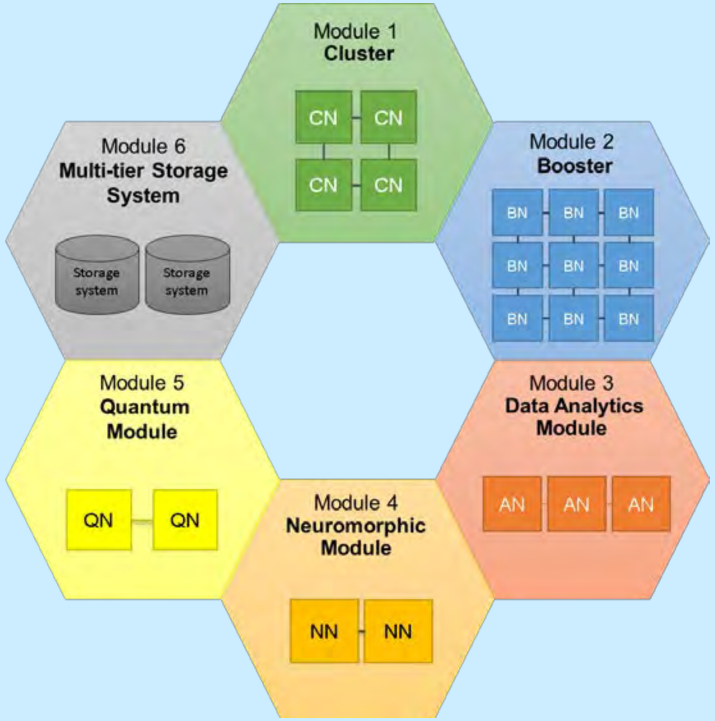
JÜLICH Forschungszentrum

The Research leading to these results has received funding from the European Commission's FP7 and H2020 Programmes, under Grant Agreements n° 287530, 610476, and 754304.

Outline

Towards a Modular Supercomputing Architecture for Exascale

- Architecture
  - Homogeneous vs. heterogeneous clusters
  - **Modular Supercomputing Architecture (MSA)**
- Software
  - Software stack
  - Programming environment
  - Scheduling and resource management
- Application example
- Conclusions and Next steps

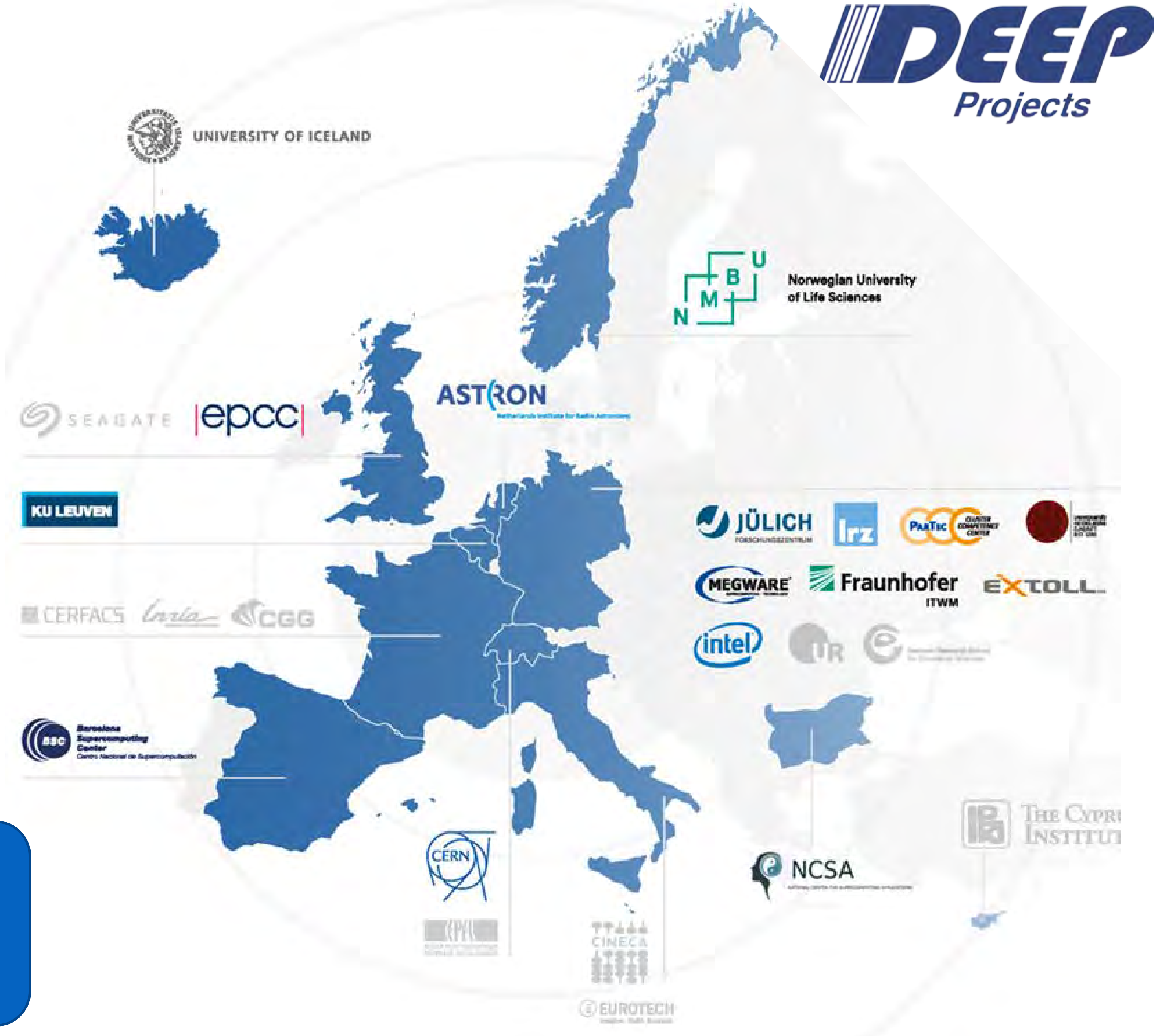


The DEEP Projects

- **DEEP**
  - Introduced the Cluster-Booster architecture
- **DEEP-ER**
  - Added I/O and resiliency functionalities
- **DEEP-EST**:
  - Extends the concept to a Modular Supercomputer Architecture

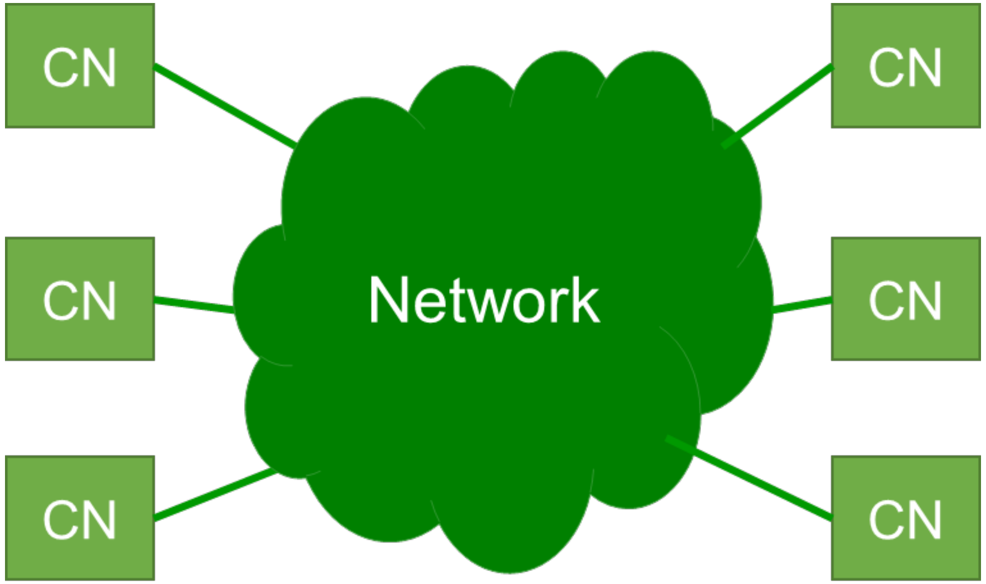
Co-design
Hardware
Software
Applications

27 partners
Time: 2011 - 2021
EU funding: 30 M€



Homogenous cluster

General purpose CPUs attached to a high-speed network



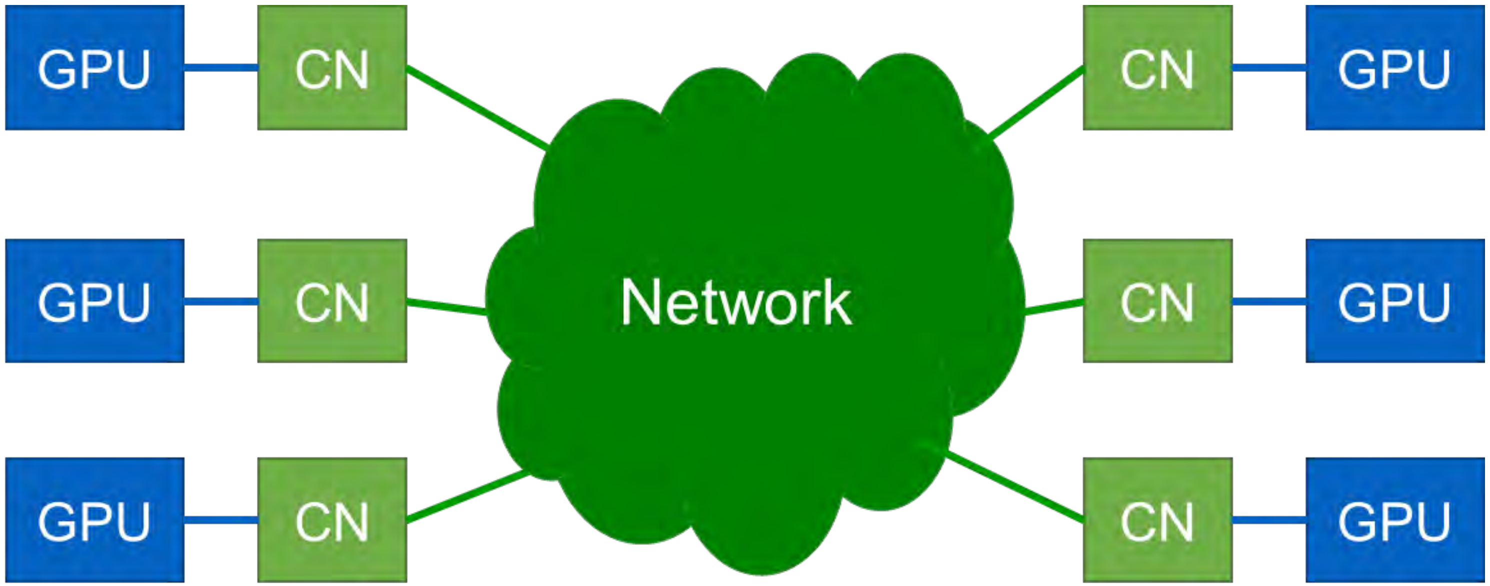
+: Easy to use
+: Very flexible

–: Power hungry

CN: Cluster Node (general purpose processor)

Traditional heterogeneous cluster

Attach accelerators (e.g. GPUs) to each CPU



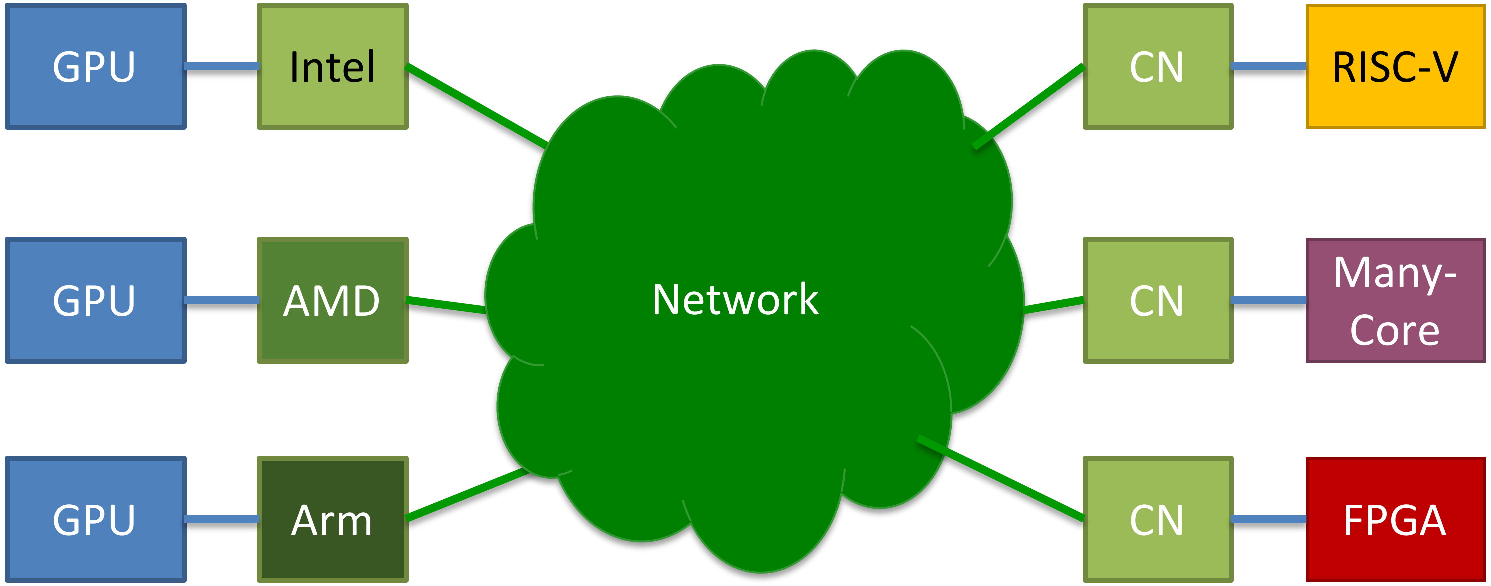
+: Energy efficient
+: Easy management

–: Static assignment accelerators-CPUs
–: Expensive scale-up

CN: Cluster Node (general purpose processor)
GPU: Graphics Processing Unit (or any other accelerator)

Highly heterogeneous cluster

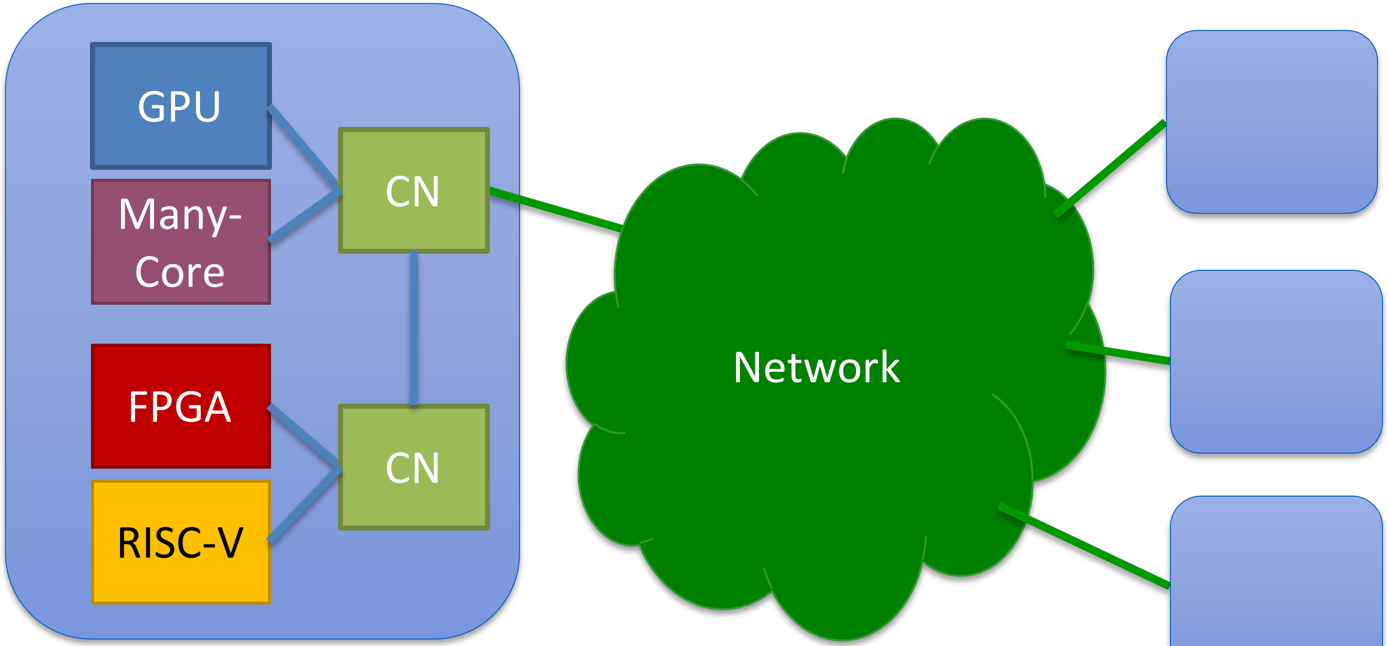
Many different general purpose and acceleration devices



How to organize and orchestrate this heterogeneity?

Highly heterogeneous cluster

Heterogeneous node

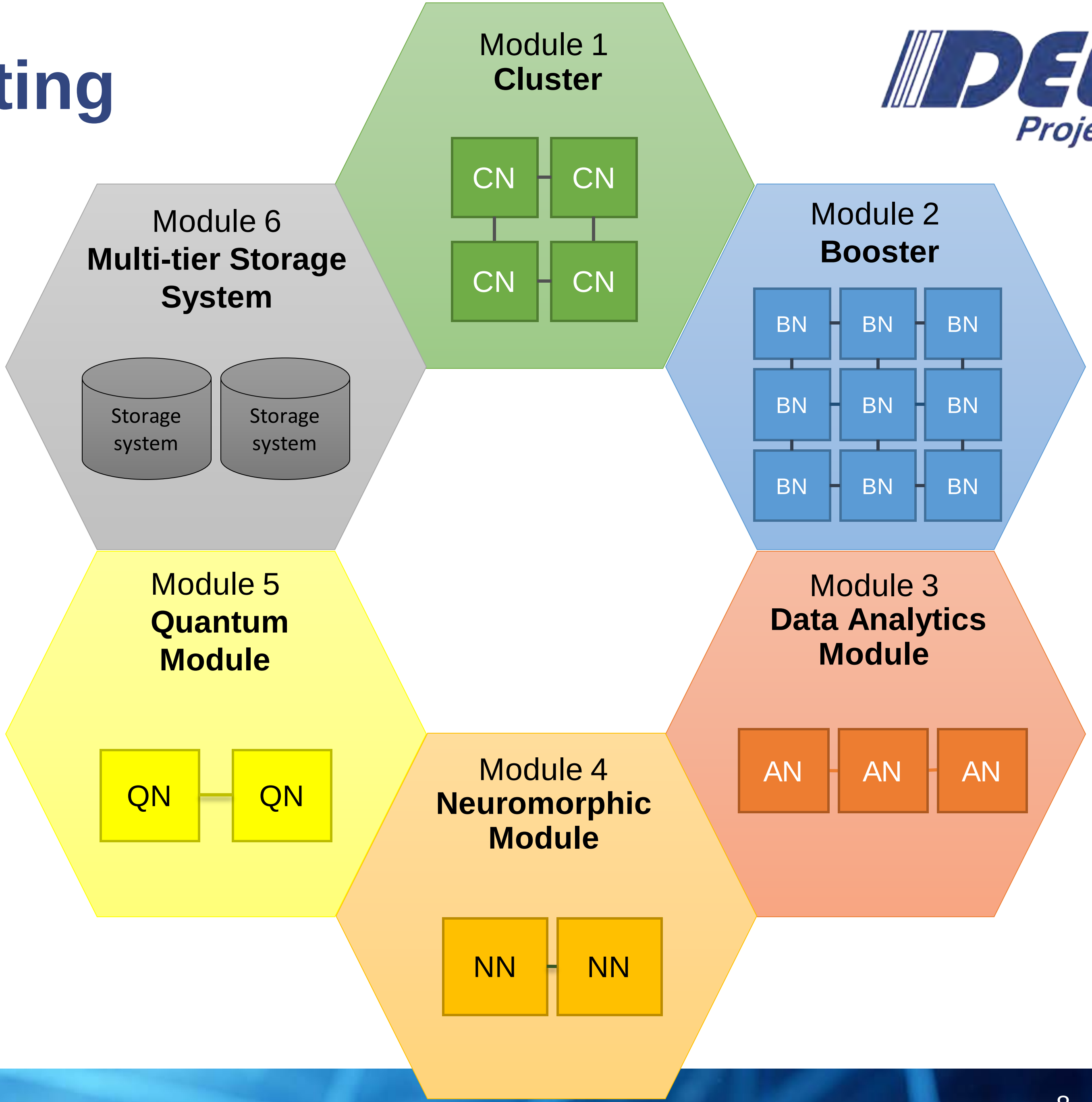


CN = Intel AMD Arm

Modular Supercomputing

Composability of heterogeneous resources

- Cost-efficient scaling
- Effective resource-sharing

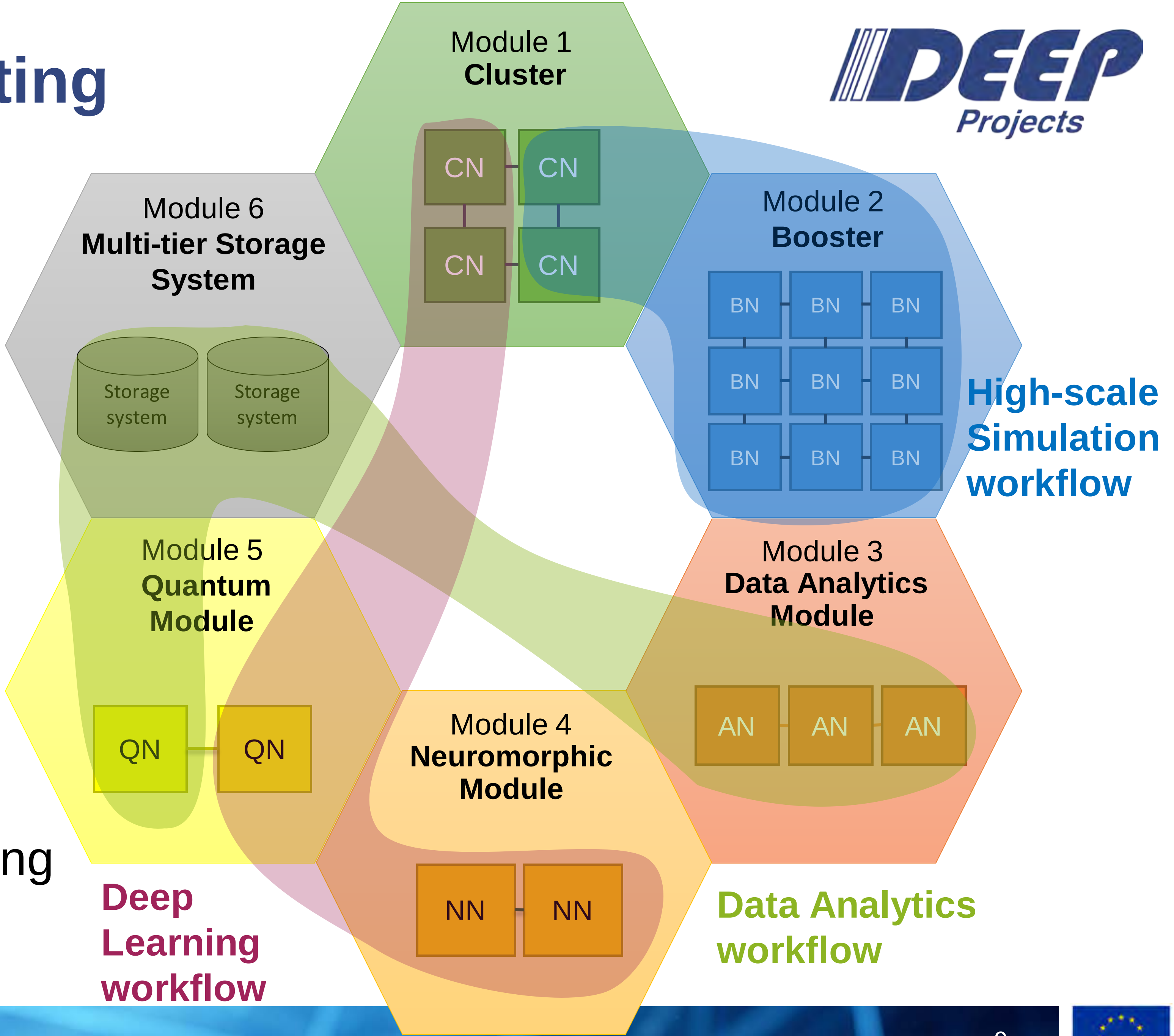


- **E. Suarez***, N. Eicker, Th. Lippert, "*Modular Supercomputing Architecture: from idea to production*", Chapter 9 in Contemporary High Performance Computing: from Petascale toward Exascale, Volume 3, pp 223-251, CRC Press. (2019)
- **E. Suarez***, N. Eicker, and Th. Lippert, "Supercomputer Evolution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

Modular Supercomputing

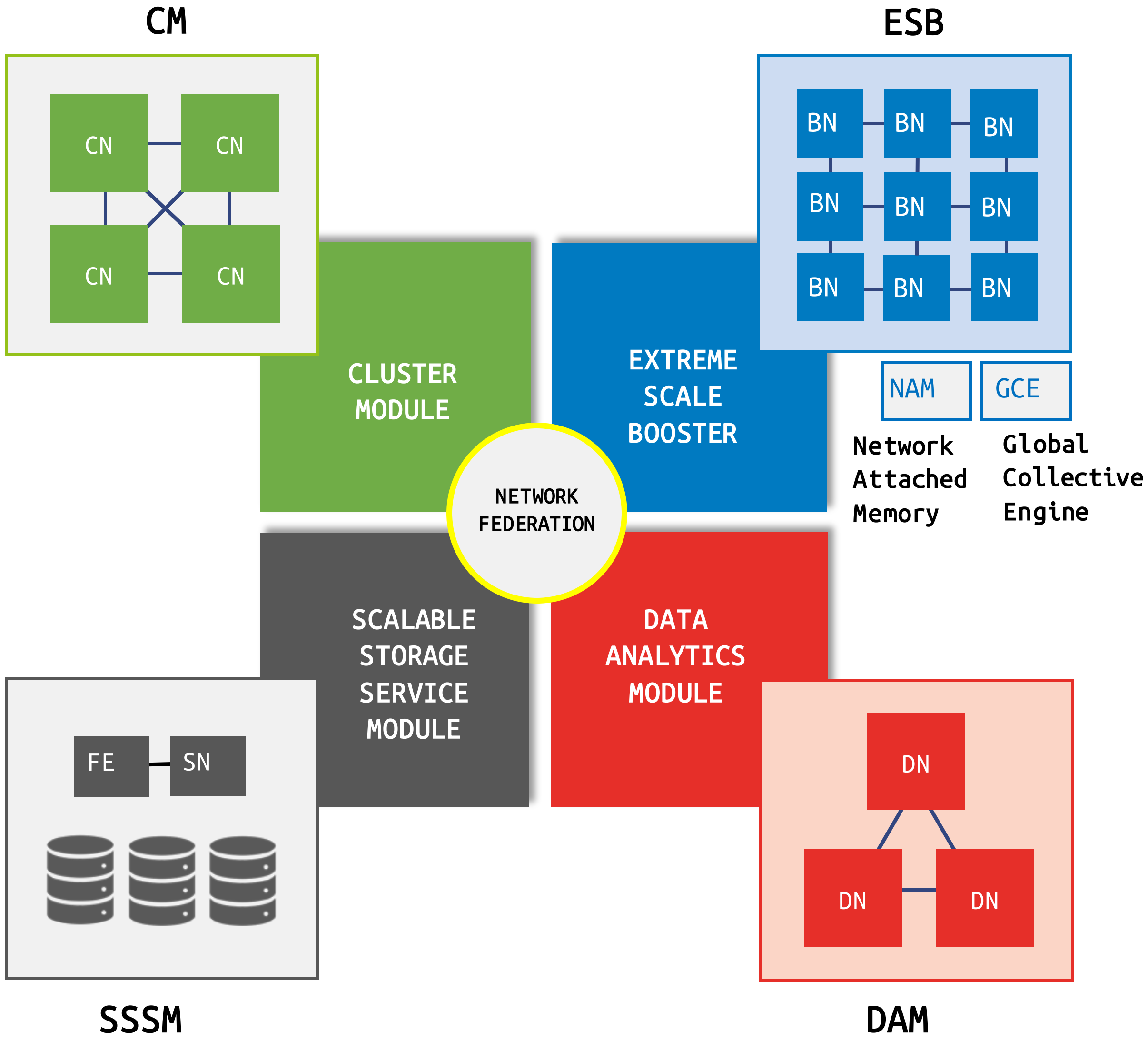
Composability of heterogeneous resources

- Cost-efficient scaling
- Effective resource-sharing
- Fit application diversity
  - Large-scale simulations
  - Data analytics
  - Machine- and Deep Learning
  - Artificial Intelligence



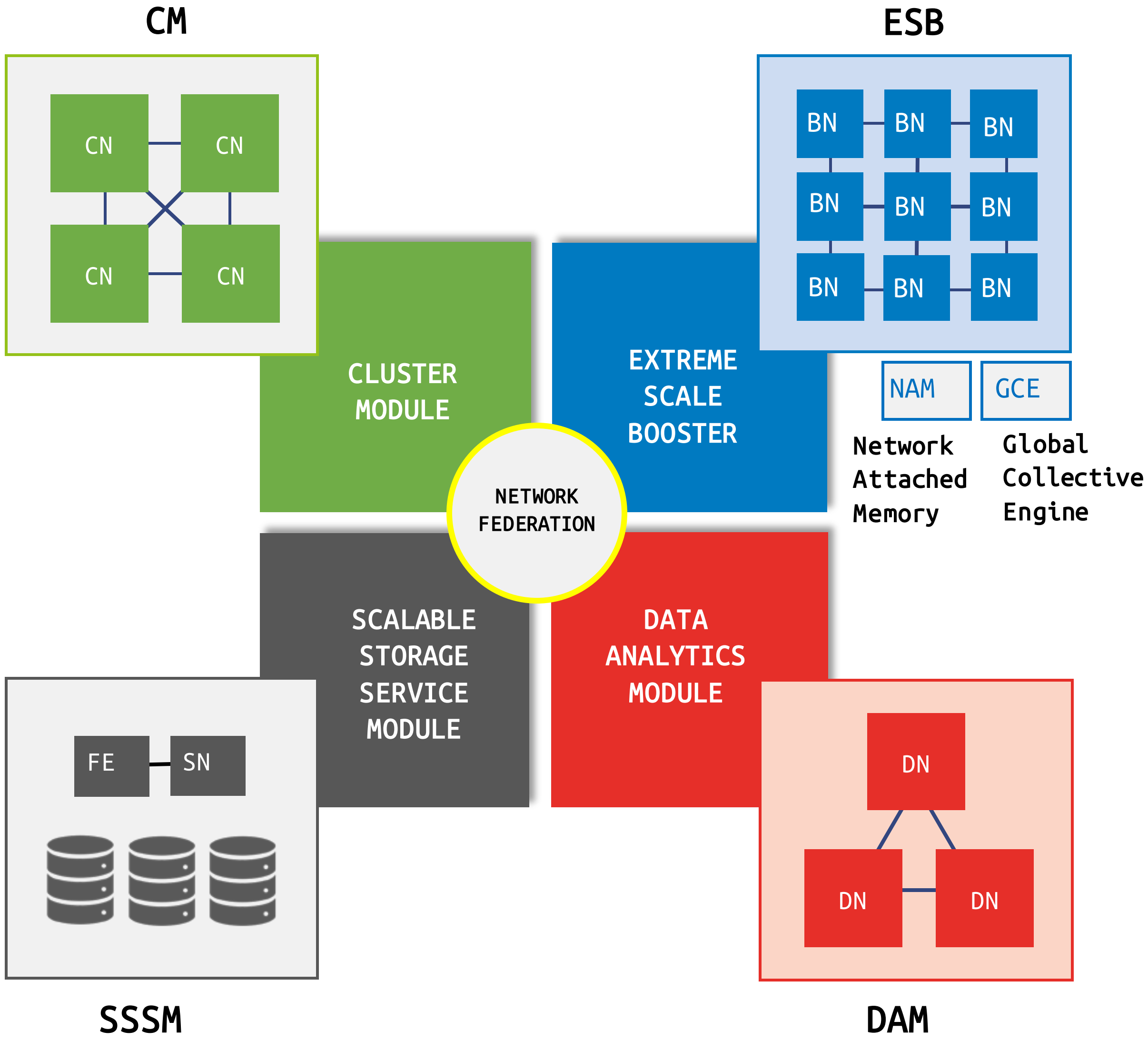
DEEP-EST modular prototype

- **Cluster Module (CM)**
  - 50× Xeon Skylake
  - InfiniBand EDR
- **Extreme Scale Booster (ESB)**
  - 75× Intel Xeon (weak SKU)
  - 75× NVIDIA V100 GPU
  - EXTOLL (100Gbit/s)
- **Data Analytics Module (DAM)**
  - 16× Xeon Cascade Lake (+NVM)
  - 16× NVIDIA GPU
  - 16× FPGA (Intel)
  - 40Gb Ethernet + 100Gb EXTOLL



DEEP-EST modular prototype

Source: DEEP Projects



Early-Access Program
https://www.deep-projects.eu/access.htm

MSA in production

a) JURECA Cluster

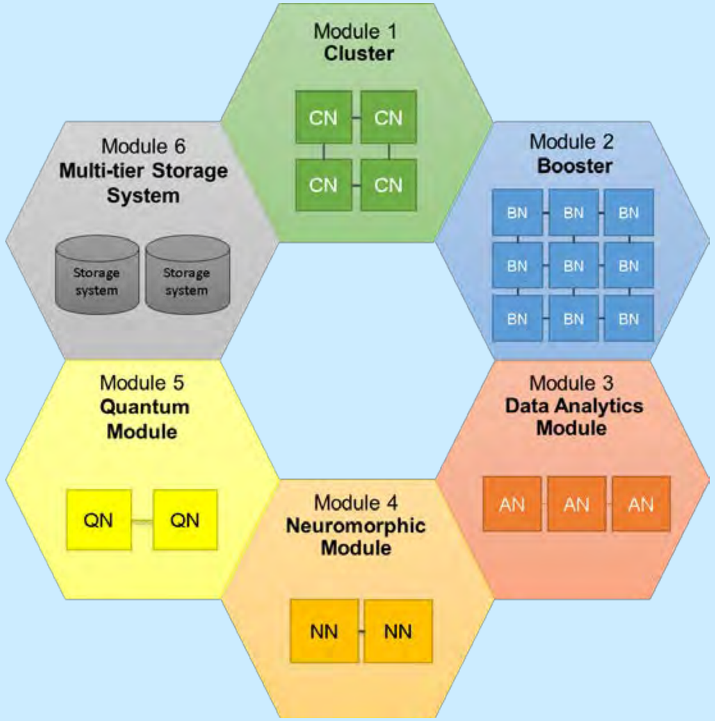
b) JURECA Booster

	Cluster	Booster
Processor	Intel Xeon (Haswell)	Xeon Phi (KNL)
Interconnect	InfiniBand EDR	OmniPath
Node count	1,872	1,640
Peak Perf. (PFlops)	1,8 (CPU) + 0.4 (GPU)	5

Outline

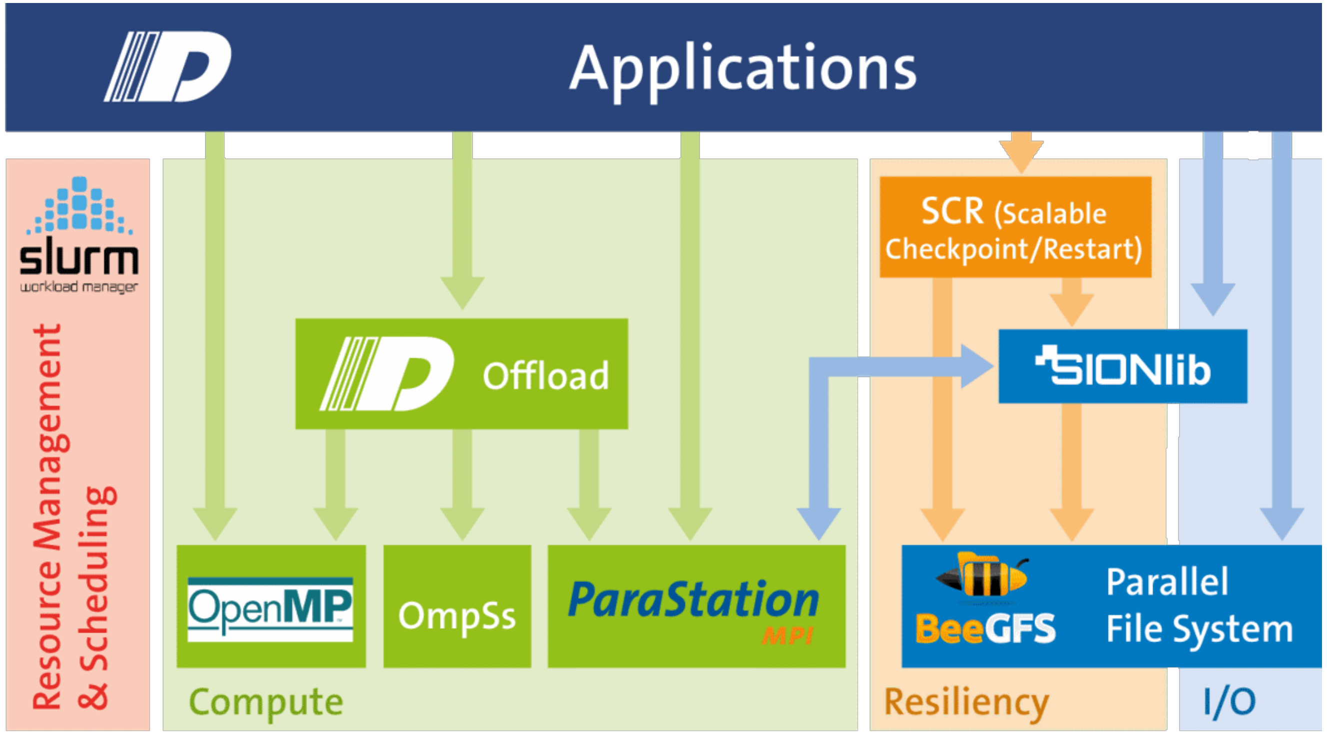
Towards a Modular Supercomputing Architecture for Exascale

- Architecture
  - Homogeneous vs. heterogeneous clusters
  - Modular Supercomputing Architecture (MSA)
- Software
  - Software stack
  - Programming environment
  - Scheduling and resource management
- Application example
- Conclusions and Next steps



Software environment

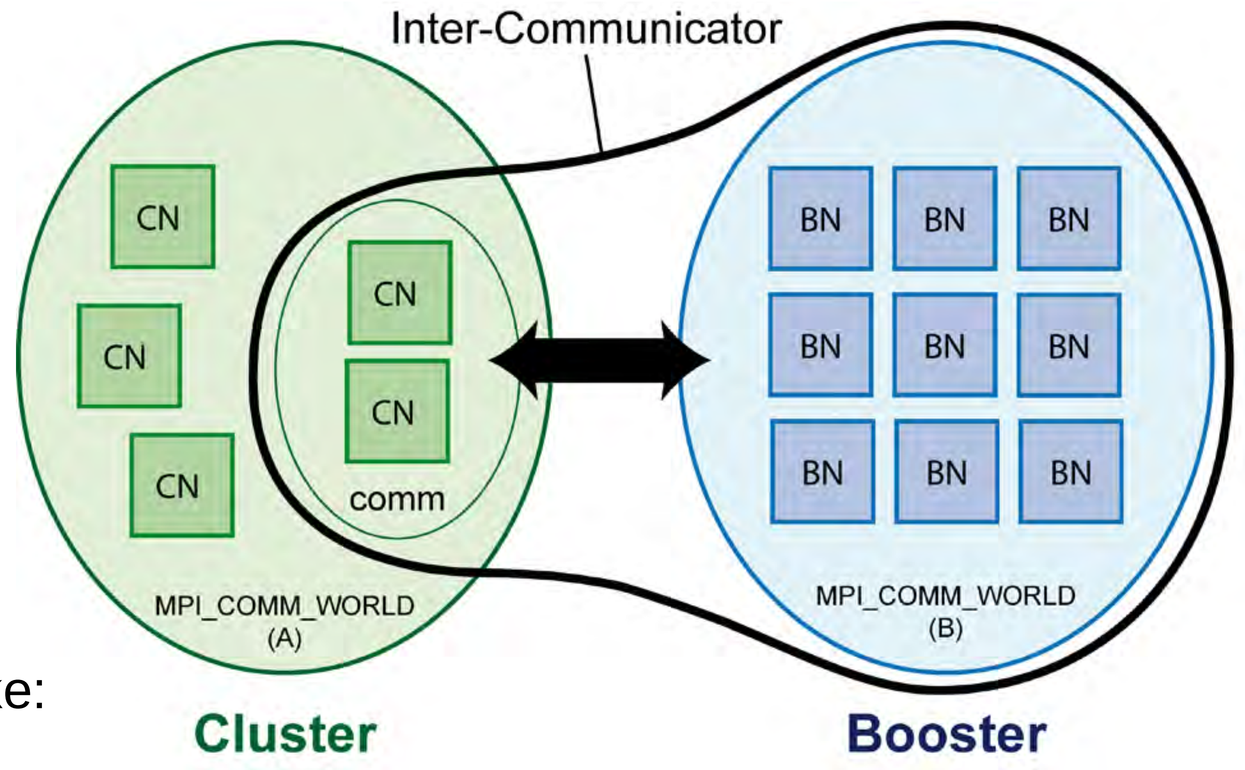
- **Low-level SW**: Cluster-Booster protocol
- **Scheduler**: SLURM
- **Filesystem**: BeeGFS
- **Compilers**: Intel, gcc, PGI
- **Debuggers**: Intel Inspector, TotalView
- **Programming**: ParaStation MPI (mpich), OpenMP, OmpSs
- **Performance analysis tools**: Scalasca, Extrae/Paraver, Intel Advisor, VTune…
- **Benchmarking tools**: JUBE
- **Libraries**: SIONlib, SCR, HDF5…



- **Eicker et al.**, *Bridging the DEEP Gap - Implementation of an Efficient Forwarding Protocol*, Intel EU Exascale Labs Report 2013 34-41, (2014)
- **Clauss et al.**, *Dynamic Process Management with Allocation-internal Co-Scheduling towards Interactive Supercomputing*, COSH@HiPEAC, (2016)

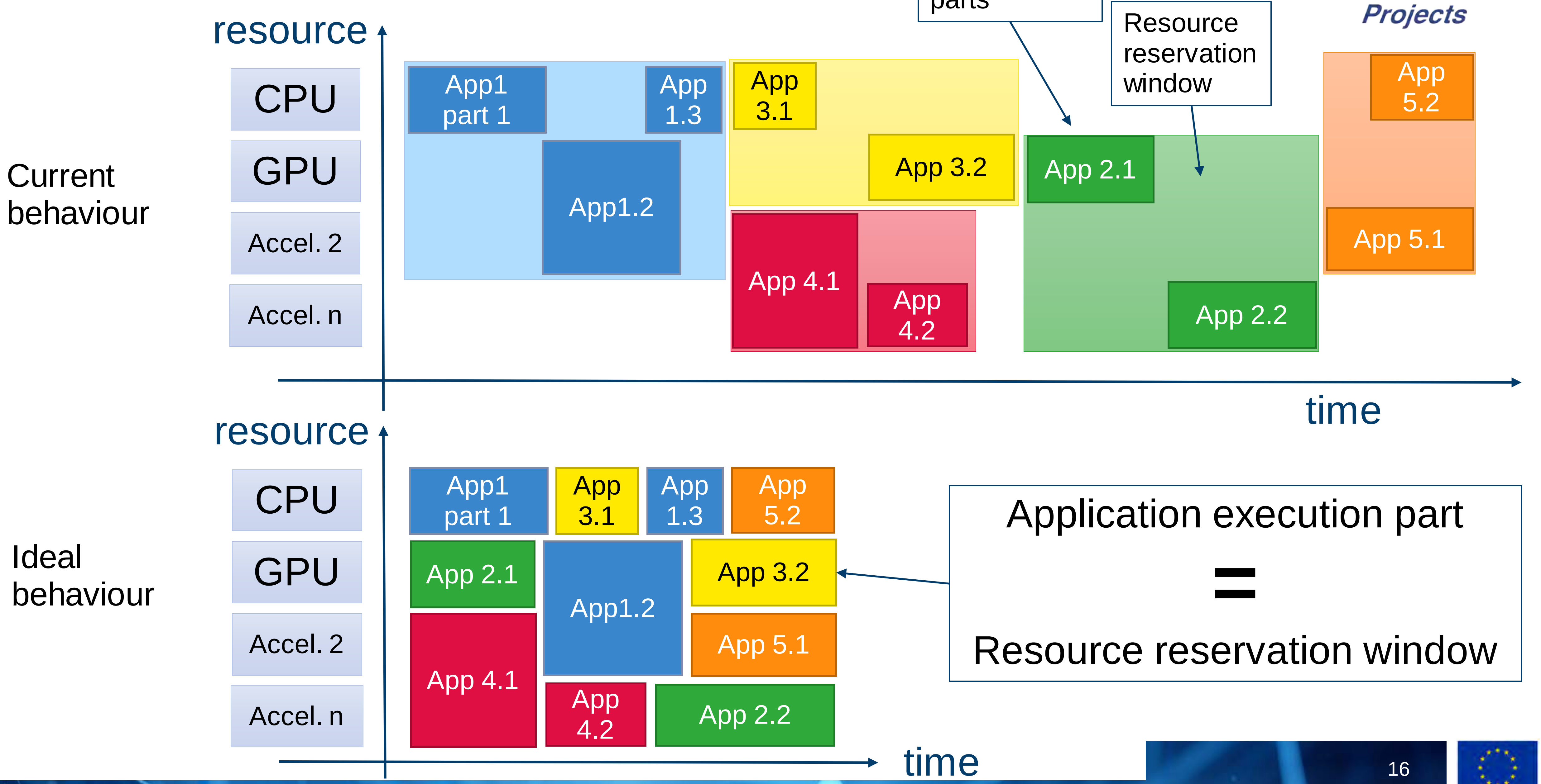
Programming Environment

- One application can run:
  - Using only Cluster nodes
  - Using only Booster nodes
  - Distributed over Cluster and Booster
    - *In this case two executables are created*
    - *Collective offload process*
- ParaStation Global MPI
  - Enables distributing code
  - Uses MPI standard collective instructions like:
    - **`MPI_Comm_spawn()`**
    - **`MPI_Connect()`**
    - **`MPI_Comm_Split()`**
  - Inter-communicator
    - *Connects the 2 MPI_Comm_worlds*



- **Clauss et al.**, *Dynamic Process Management with Allocation-internal Co-Scheduling towards Interactive Supercomputing*, COSH@HiPEAC, (2016)

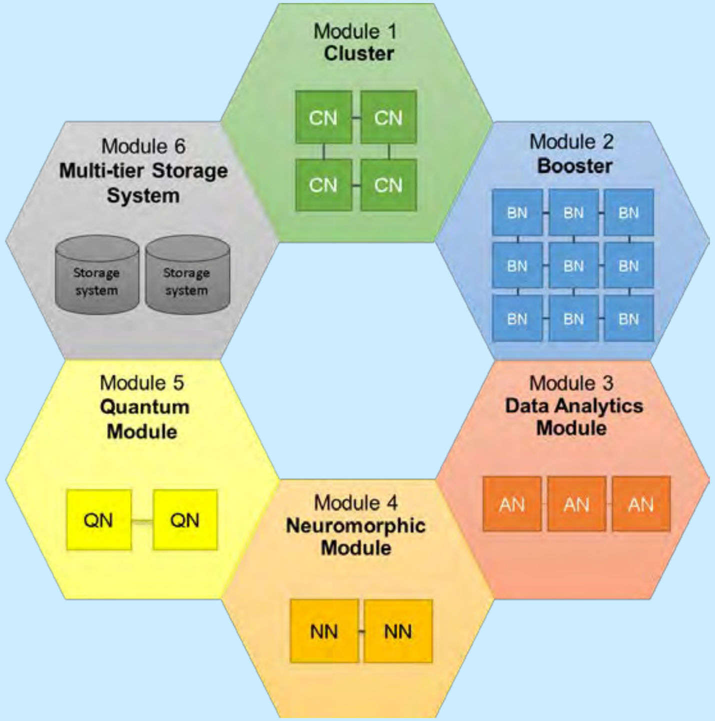
Resource management



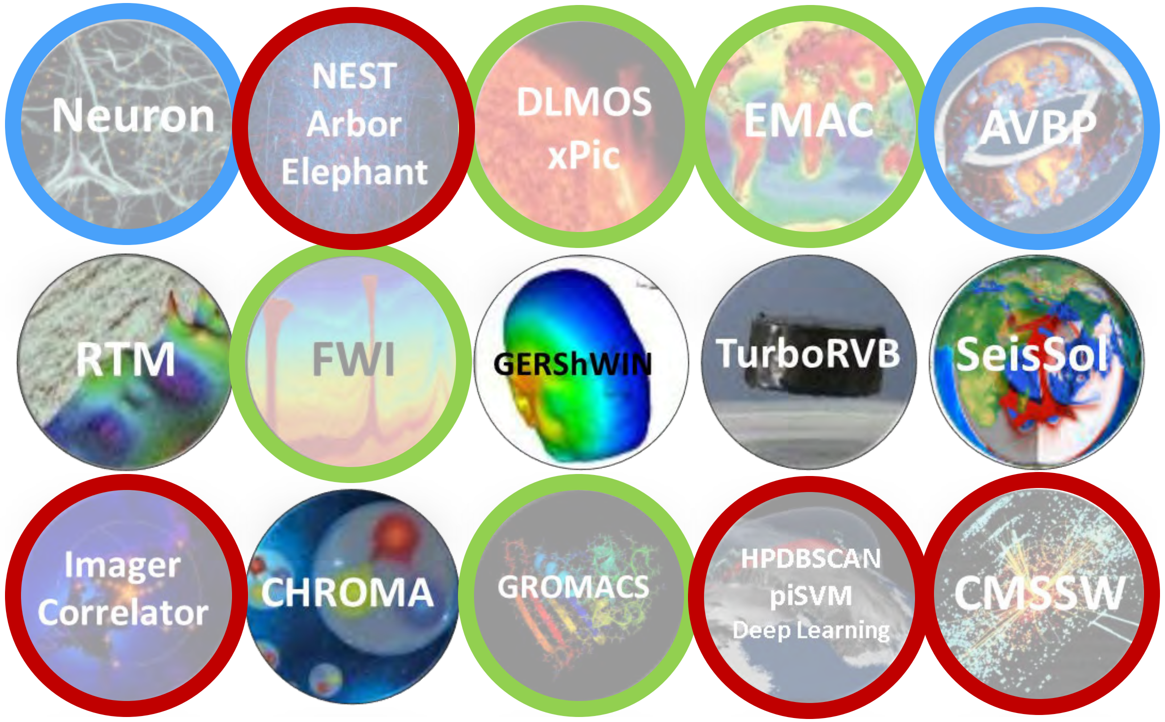
Outline

Towards a Modular Supercomputing Architecture for Exascale

- Architecture
  - Homogeneous vs. heterogeneous clusters
  - Modular Supercomputing Architecture (MSA)
- Software
  - Software stack
  - Programming environment
  - Scheduling and resource management
- Application example
- Conclusions and Next steps



Architecture Use-Modes



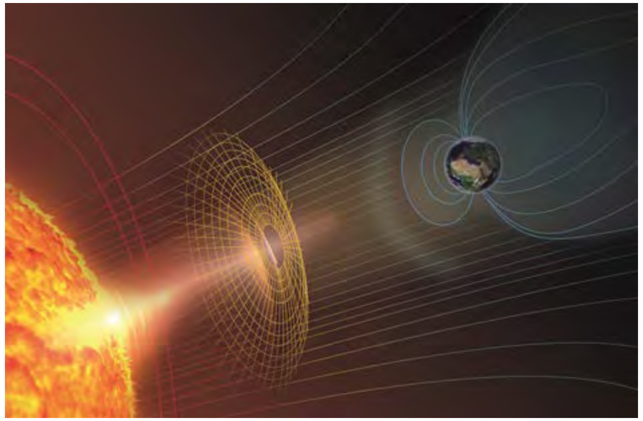
Cluster-Booster use mode

Code partition
Workflow
I/O forward

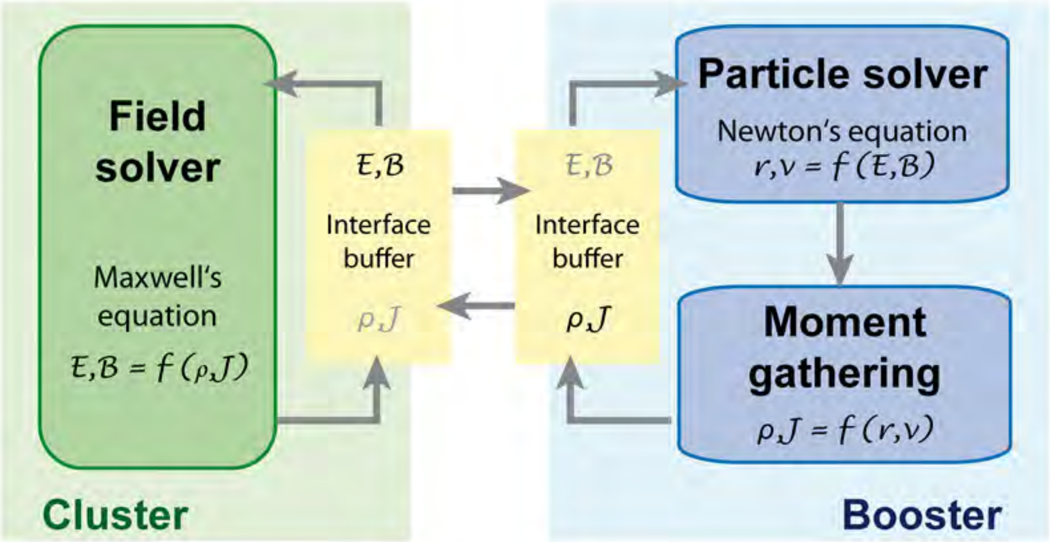
- **Kreuzer, et al.,** *Application Performance on a Cluster-Booster System.* IPDPSW – HCW (2018) [10.1109/IPDPSW.2018.00019]
- **Kreuzer et al.** *The DEEP-ER project: I/O and resiliency extensions for the Cluster-Booster architecture.* HPCC'18 proceedings (*2018)* [10.1109/HPCC/SmartCity/DSS.2018.00046]
- Wolf et al., *PIC algorithms on DEEP: The iPiC3D case study*. PARS-Mitteilungen 32, 38-48 (2015)
- Christou et al., *EMAC on DEEP*, Geoscientific model devel.(2016) [10.5194/gmd-9-3483-2016]
- Kumbhar et al., *Leveraging a Cluster-Booster Architecture for Brain-Scale Simulations*, Lecture Notes in Computer Science 9697 (2016) [10.1007/978-3-319-41321-1_19]
- Leger et al., *Adapting a Finite-Element Type Solver for Bioelectromagnetics to the DEEP-ER Platform*. ParCo 2015, Advances in Parallel Computing, 27 (2016) [10.3233/978-1-61499-621-7-349]

Application use case: xPic

- **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 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]

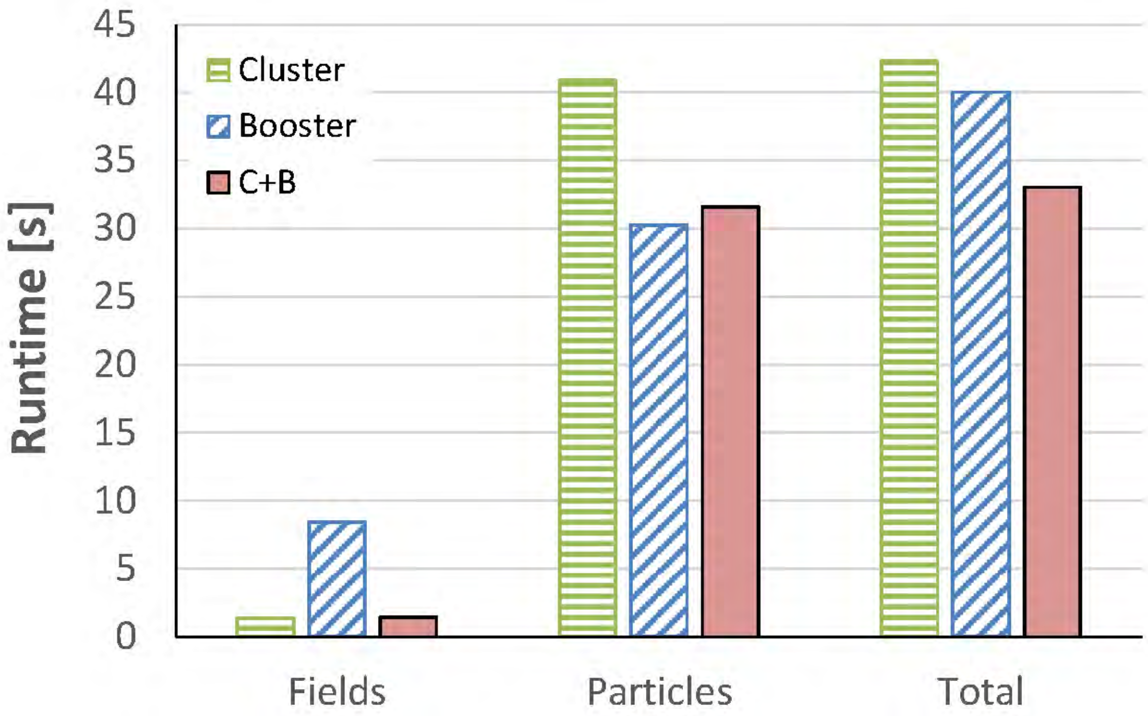
xPic – (1-node) Performance Results

- **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

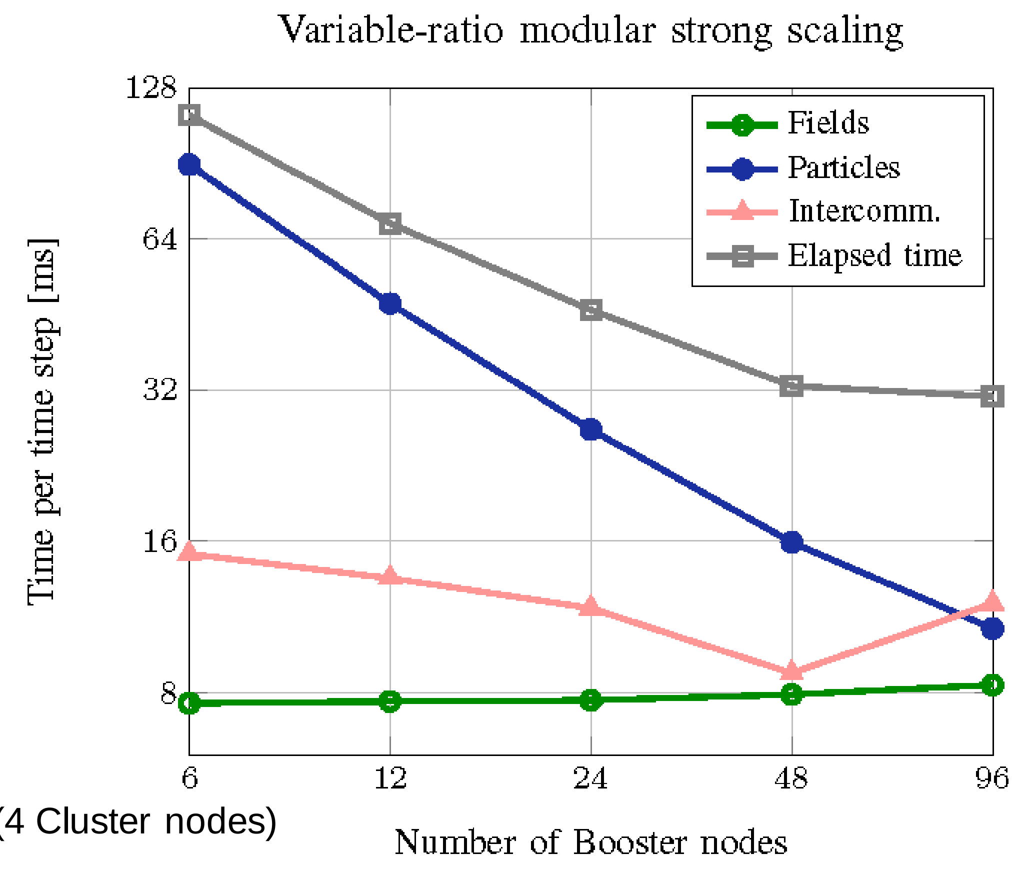
– 3%-4% overhead per solver for C+B communication (point to point)



#cells per node	4096
#particles per cell	2048
Compilation flags	-openmp, -mavx (Cluster) -xMIC-AVX512 (Booster)

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 on JURECA



(4 Cluster nodes)

#cells per node	36864
#particles per cell	1024
#blocks per MPI process	12, 32 or 64
Compilation flags	-mavx (Cluster) -openmp, xMIC-AVX512 (Booster)

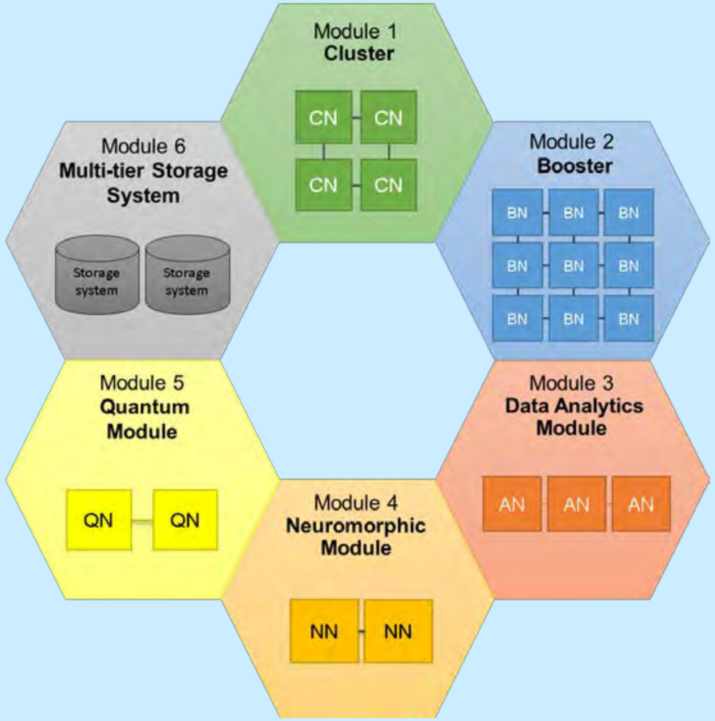
- Code portions can be scaled-up independently
  - Particles scale almost linearly on Booster
  - Fields kept constant on Cluster (4CNs)
- A configuration is reached where same time is spent on Cluster and Booster
  - Additional 2× time-saving can be enabled via overlapping

Work performed by J. de Amicis (JSC)

Outline

Towards a Modular Supercomputing Architecture for Exascale

- Architecture
  - Homogeneous vs. heterogeneous clusters
  - Modular Supercomputing Architecture (MSA)
- Software
  - Software stack
  - Programming environment
  - Scheduling and resource management
- Application example
- Conclusions and Next steps



Conclusions

- **The Modular Supercomputing Architecture (MSA)**
  – Orchestrates heterogeneity at system level
  – Allows scaling hardware in economical way
    - *Booster → Exascale*
  – Serves very diverse application profiles
    - *Maximum flexibility for users, without taking anything away*
    - *Still can use modules individually*
- **Distribute applications to run each code-part on suitable hardware**
  – Straight-forward implementation for workflows
  – Partition at MPI-level interesting for multi-physics / multi-scale codes
  – Monolithic codes can run inside the best suited module, without code-division

NEXT steps

- **Software development**
  - Develop tools to map applications to hardware
  - Improve support for malleability and dynamical resource allocation
  - Better scheduling of heterogeneous jobs/workflows
  - Facilitate exploitation of new memory technologies
  - Modularize more codes

- **Current / Upcoming implementations of MSA**
  - DEEP prototypes, JURECA, JUWELS (in 2020)
  - MeluXina (Luxembourg EuroHPC Petascale system)
  - Leonardo (Italy EuroHPC Pre-Exascale system)
  - Tianhe-3 *(heterogeneous flexible architecture)*
    - *https://www.r-ccs.riken.jp/R-CCS-Symposium/2019/slides/Wang.pdf*
  - Pilot system (as pre-Exascale demonstrator)
  - And if everything goes well, then…. **Exascale**!

With thanks to:
the full team in the DEEP projects

www.deep-projects.eu
@DEEPprojects

DEEP Projects

The DEEP projects have received funding from the European Union's Seventh Framework Programme (FP7) for research, technological development and demonstration and the Horizon2020 (H2020) funding framework under grant agreement no. FP7-ICT-287530 (DEEP), FP7-ICT-610476 (DEEP-ER) and H2020-FETHPC-754304 (DEEP-EST).