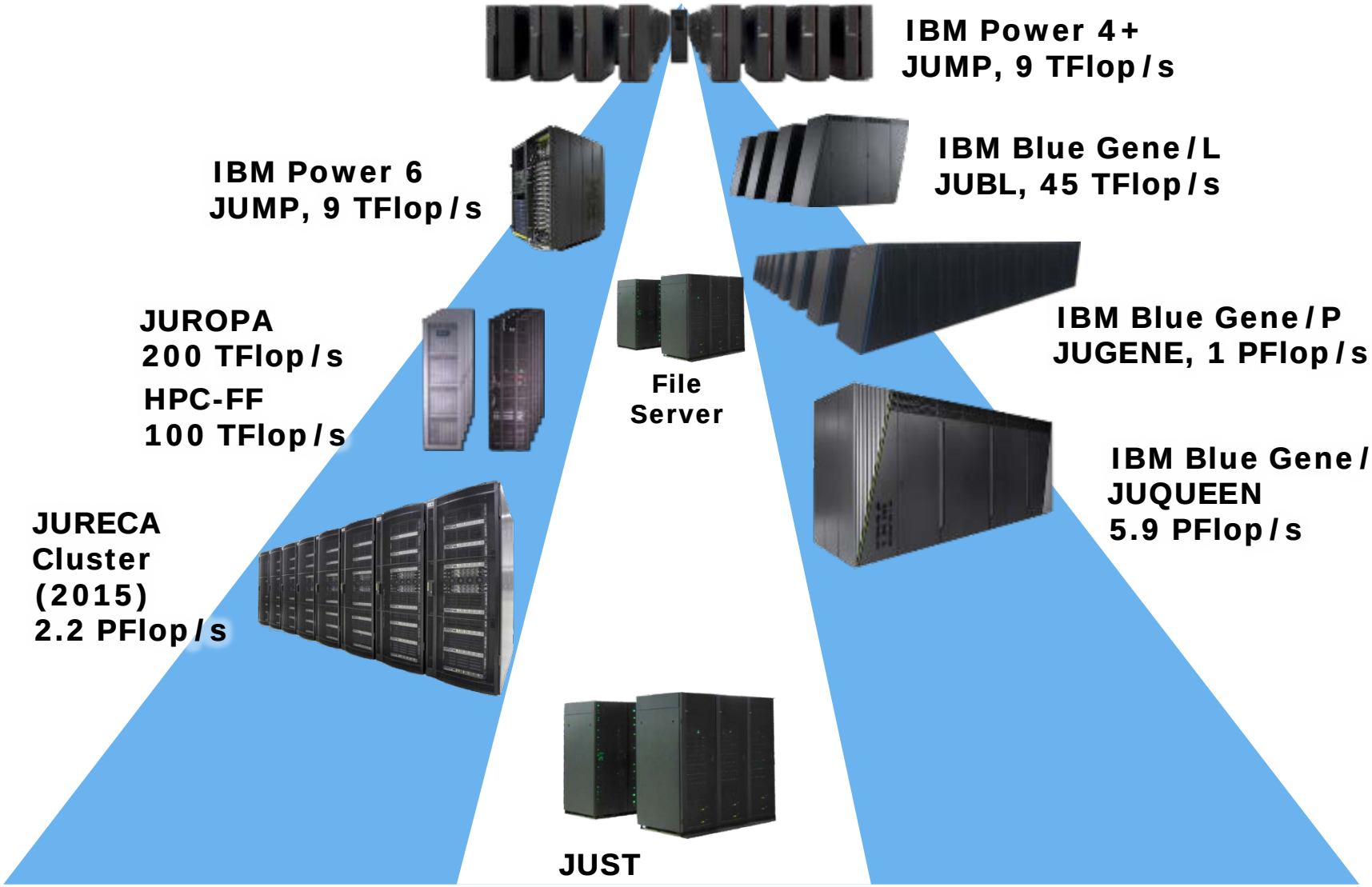
JÜLICH
FORSCHUNGSZENTRUM

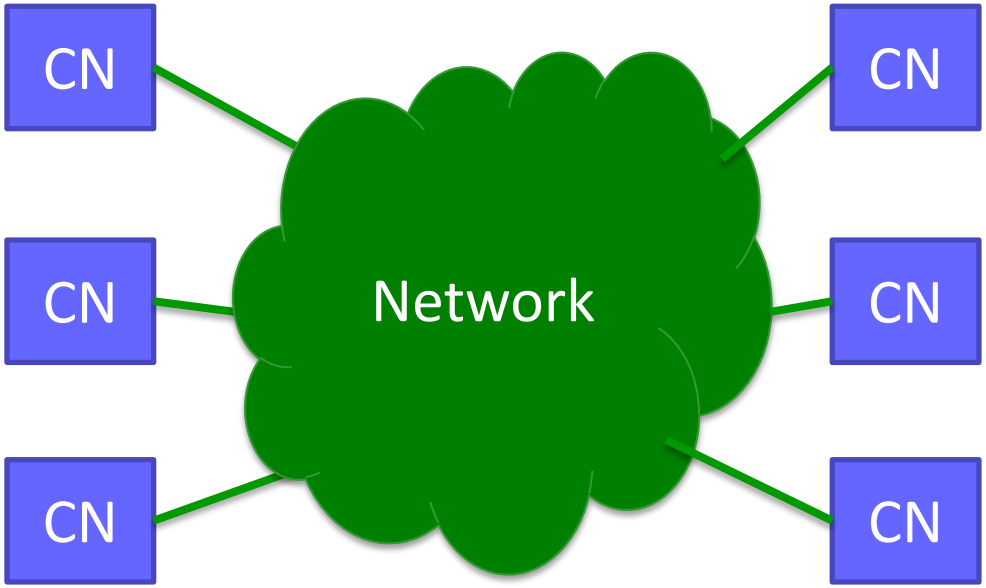
Towards Modular Supercomputing with Slurm

2017-09-25 | Dorian Krause et al.,
Jülich Supercomputing Centre, Forschungszentrum Jülich

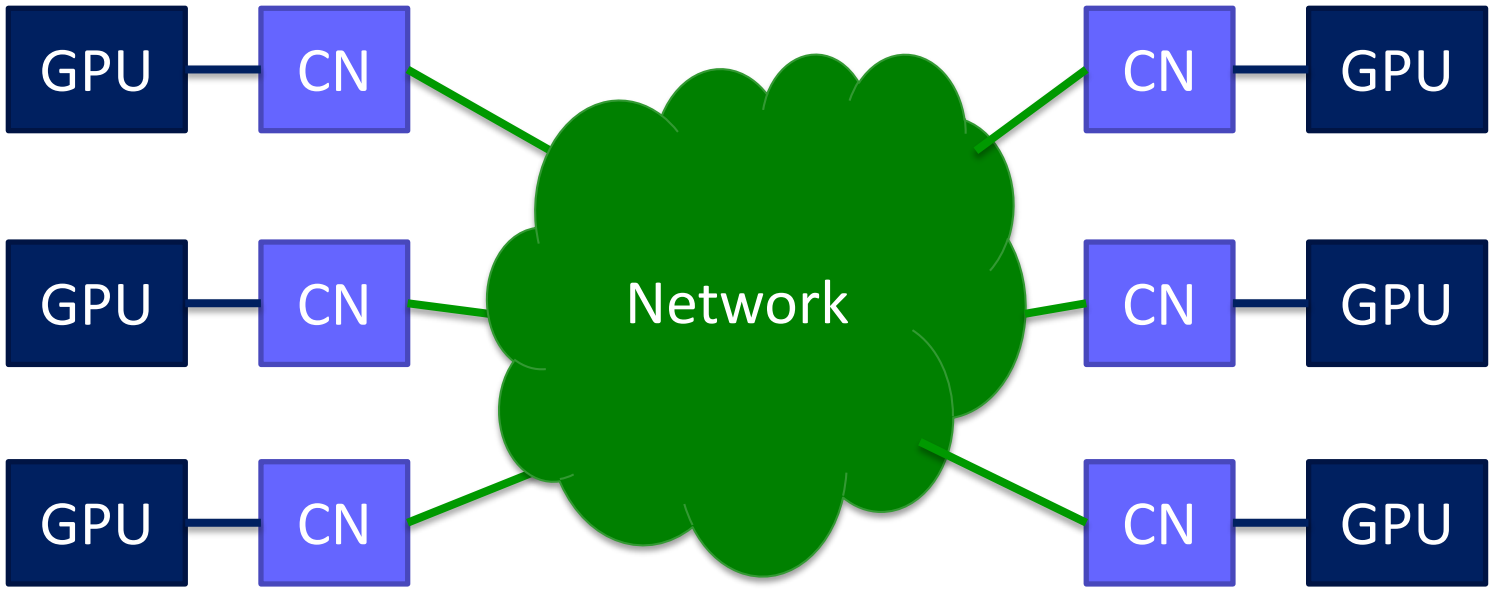
Dual-Architecture Supercomputing Facility



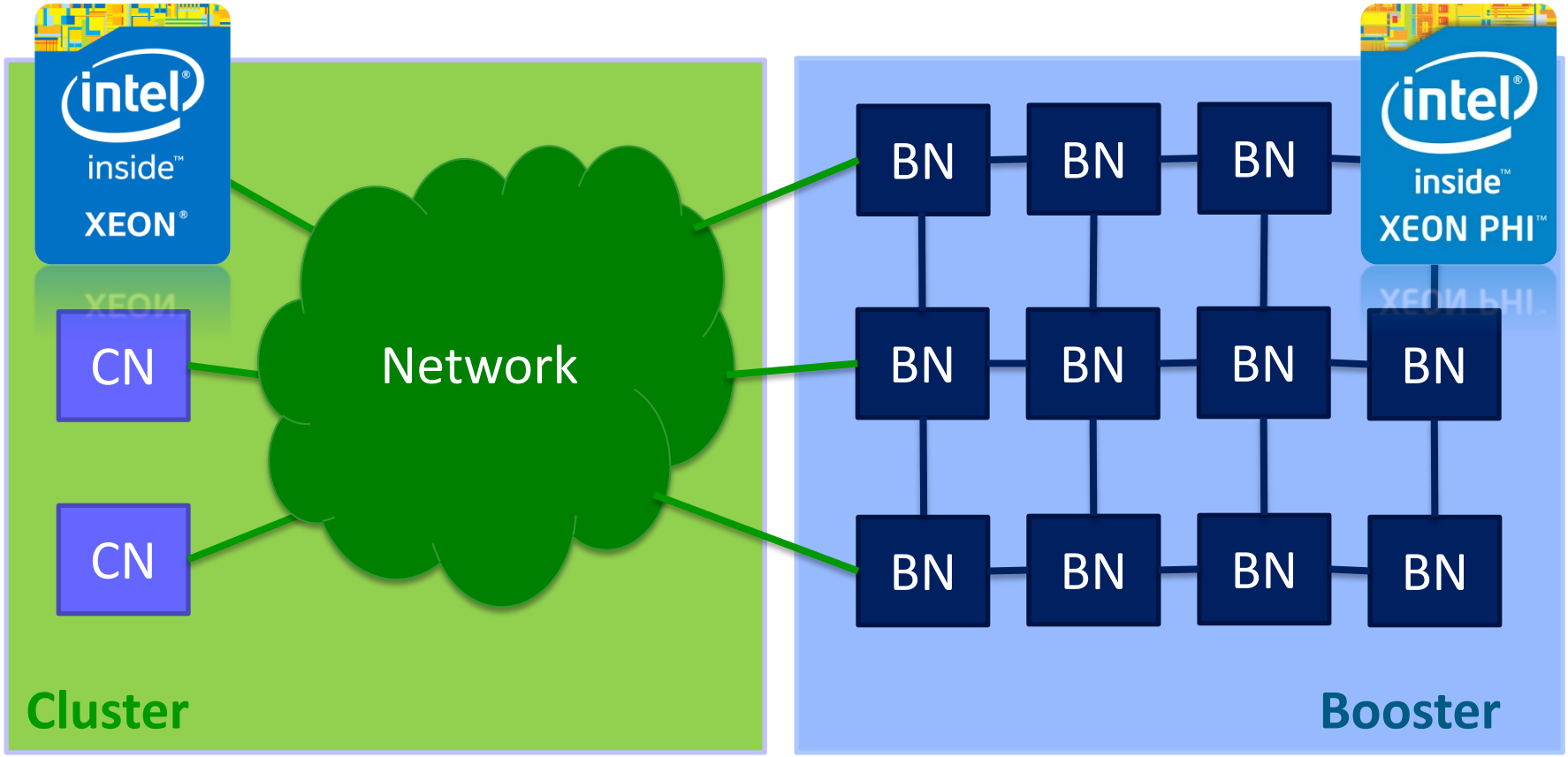
Homogeneous Cluster



Heterogeneous Cluster



Cluster-Booster Architecture



The DEEP projects: DEEP, DEEP-ER, DEEP-EST

JÜLICH FORSCHUNGSZENTRUM

www.deep-projects.eu

EU-Exascale projects

27 partners

Total budget: 44 M€

EU-funding: 30 M€

Nov 2011 – Jun 2020

All combine:

- Hardware
- Software
- Applications

in a strong co-design

The DEEP projects: DEEP, DEEP-ER, DEEP-EST

JÜLICH FORSCHUNGSZENTRUM

www.deep-projects.eu

EU-Exascale projects

27 partners

Total budge

EU-funding

Nov 2011 –

All combin

- Hardwa
- Softwa
- Applica

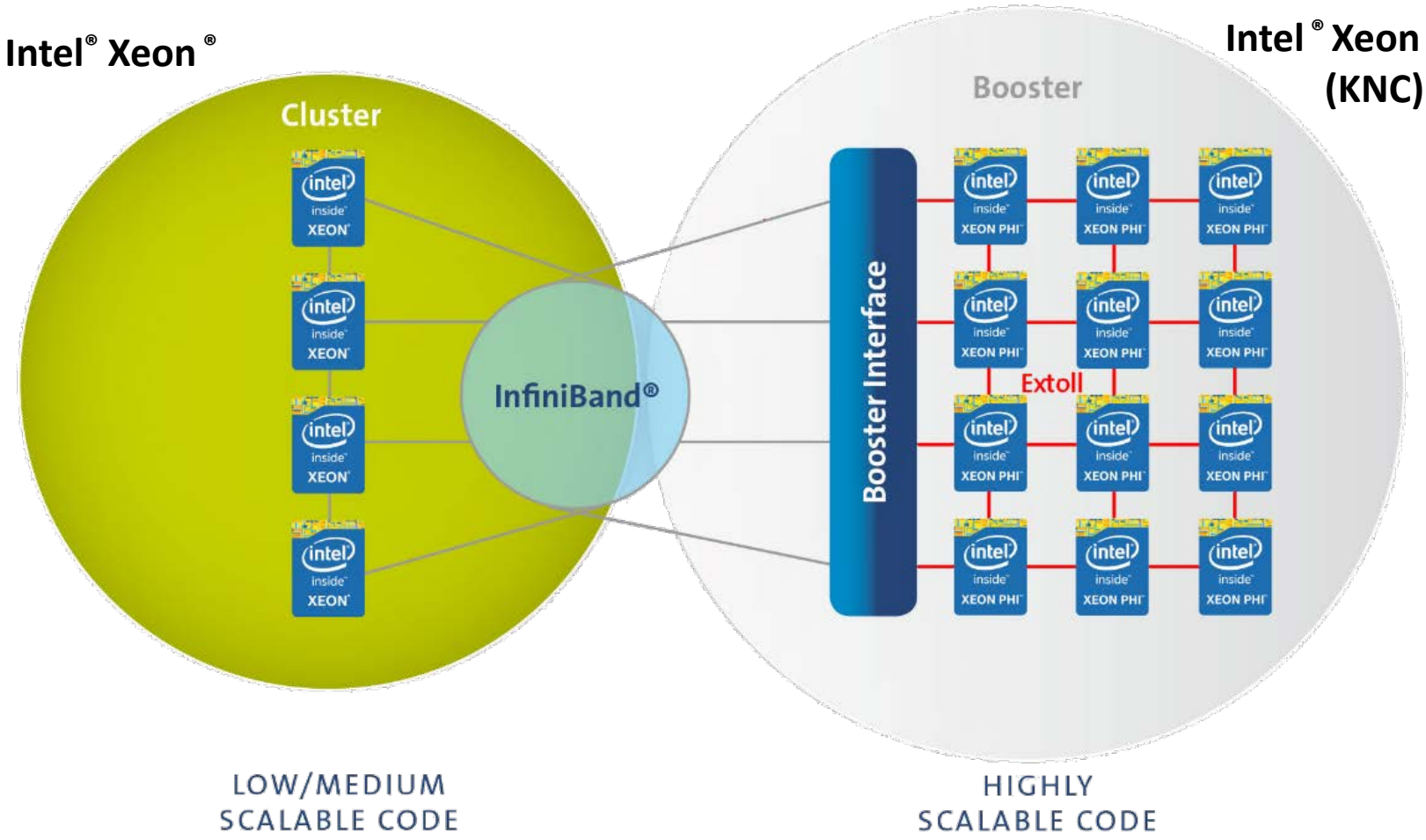
in a strong

DEEP:
Cluster-Booster Architecture + Software Environment

DEEP-ER:
I/O + Resiliency

DEEP-EST:
Modular Supercomputing

DEEP Architecture



DEEP Prototype

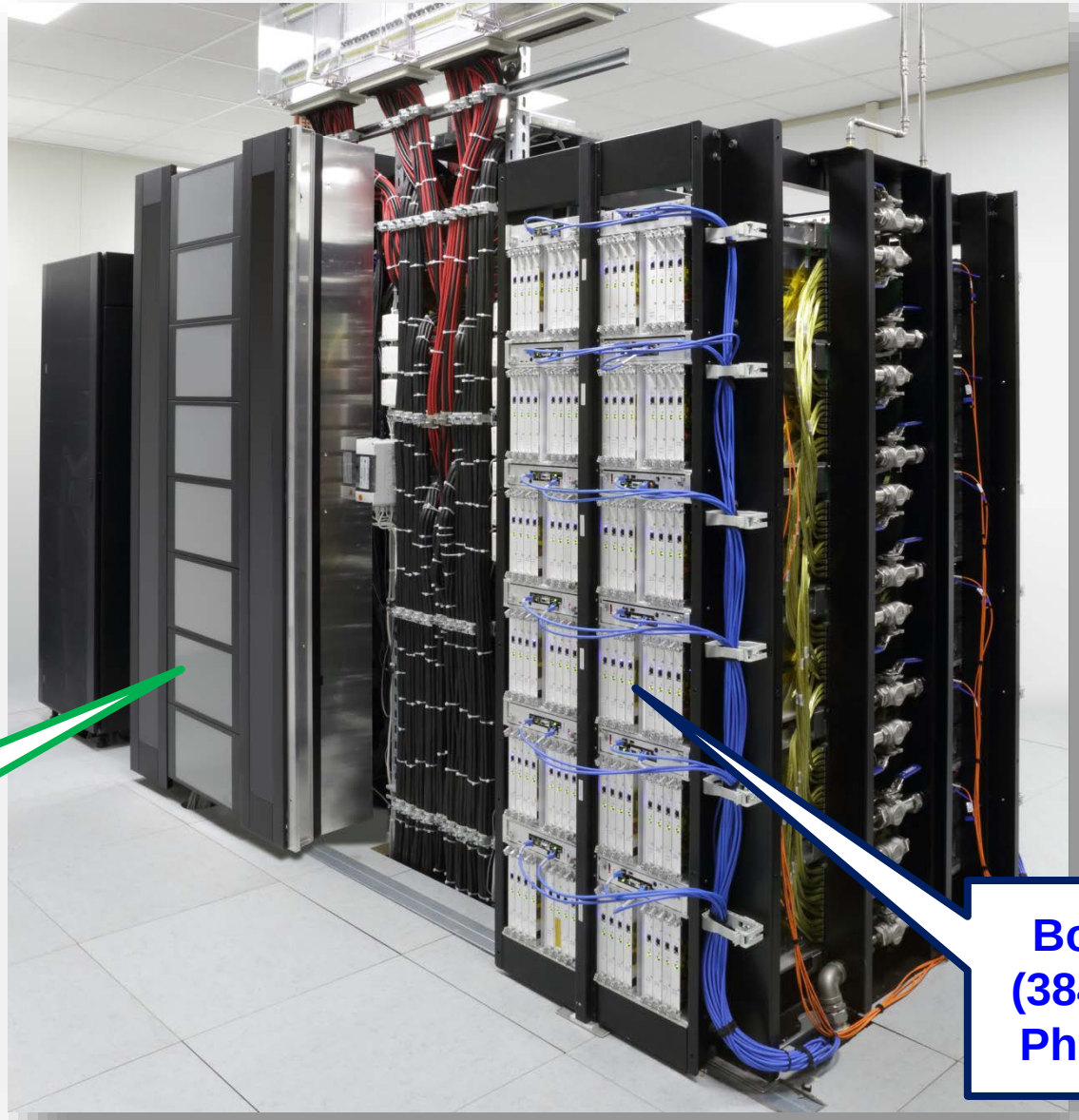
Installed at JSC

1,5 racks

500 TFlop/s peak

3.5 GFlop/s/W

Water cooled

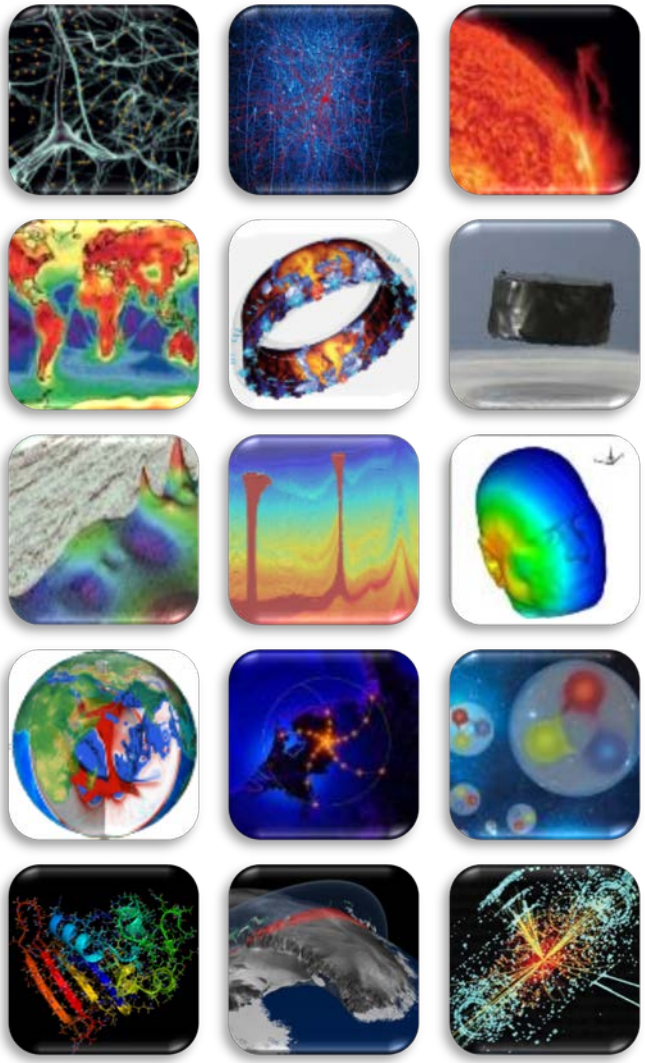


Application-Driven Approach

JÜLICH FORSCHUNGSZENTRUM

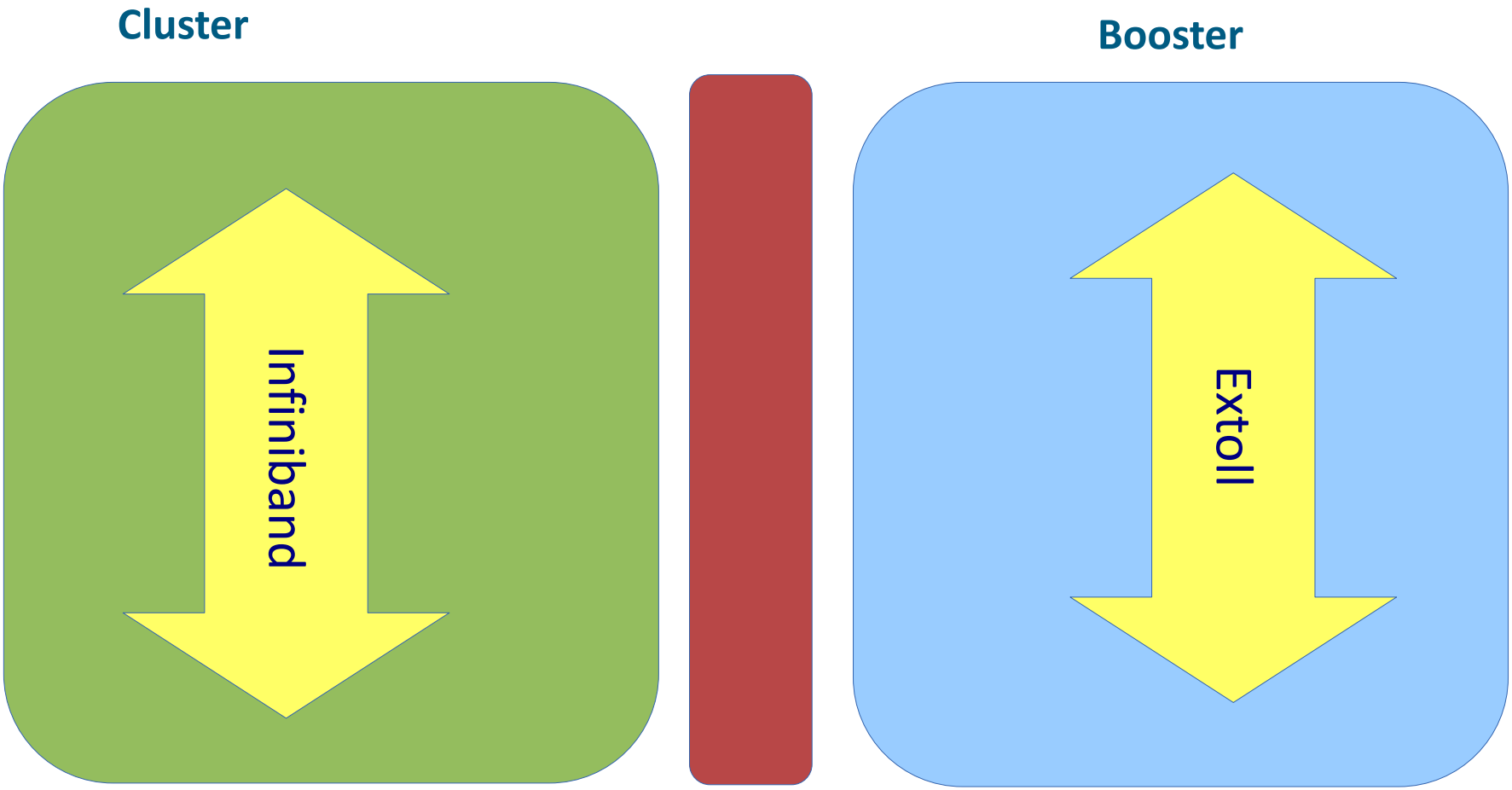
DEEP projects applications (15):

- Brain simulation (EPFL + NMBU)
- Space weather simulation (KU Leuven)
- Climate simulation (Cyprus Institute)
- Computational fluid engineering (CERFACS)
- High temperature superconductivity (CINECA)
- Seismic imaging (CGG + BSC)
- Human exposure to electromagnetic fields (INRIA)
- Geoscience (LRZ)
- Radio astronomy (Astron)
- Lattice QCD (University of Regensburg)
- Molecular dynamics (NCSA)
- Data analytics in Earth Science (UoI)
- High Energy Physics (CERN)



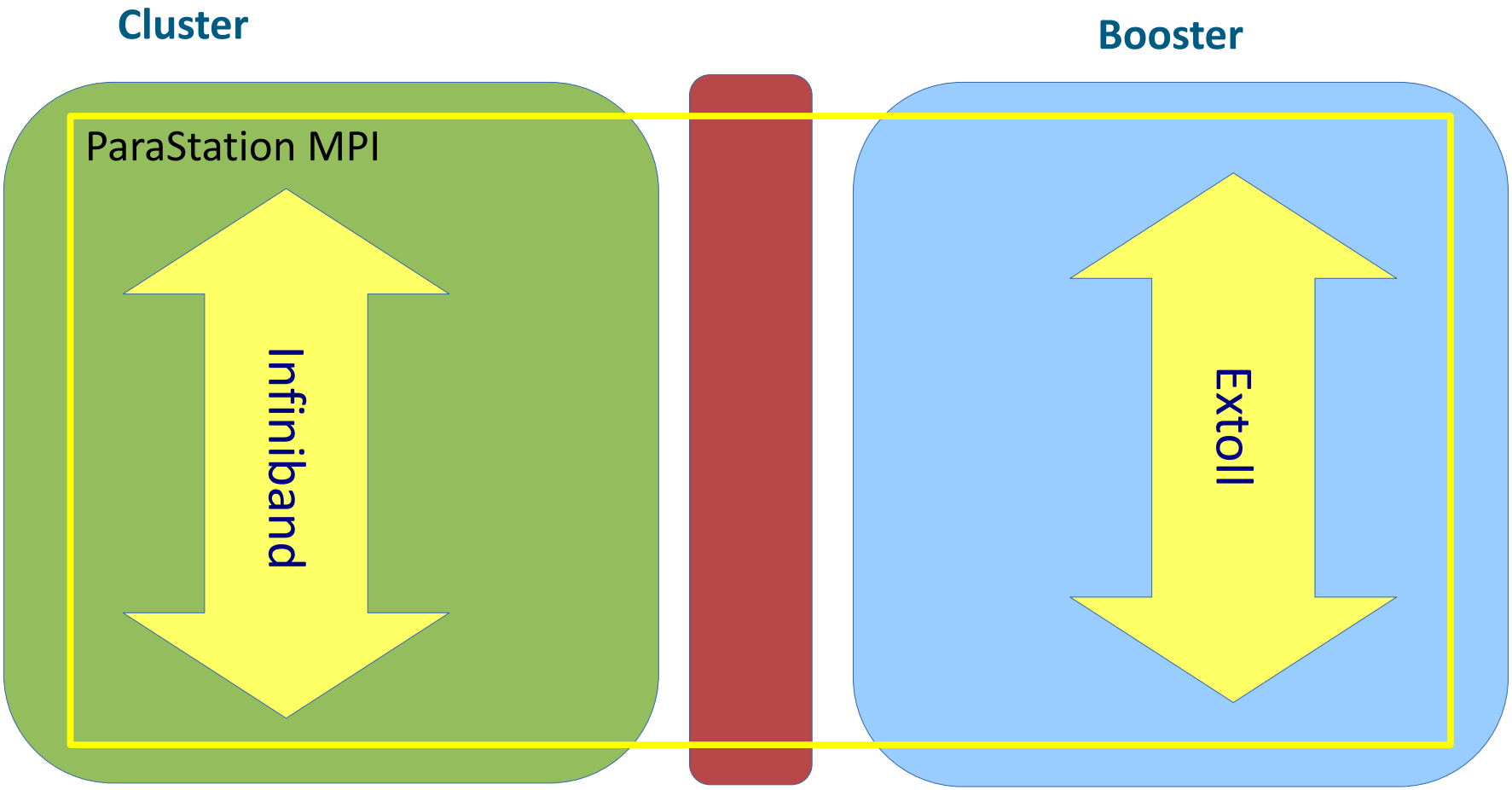
CO-DESIGN

Programming Environment



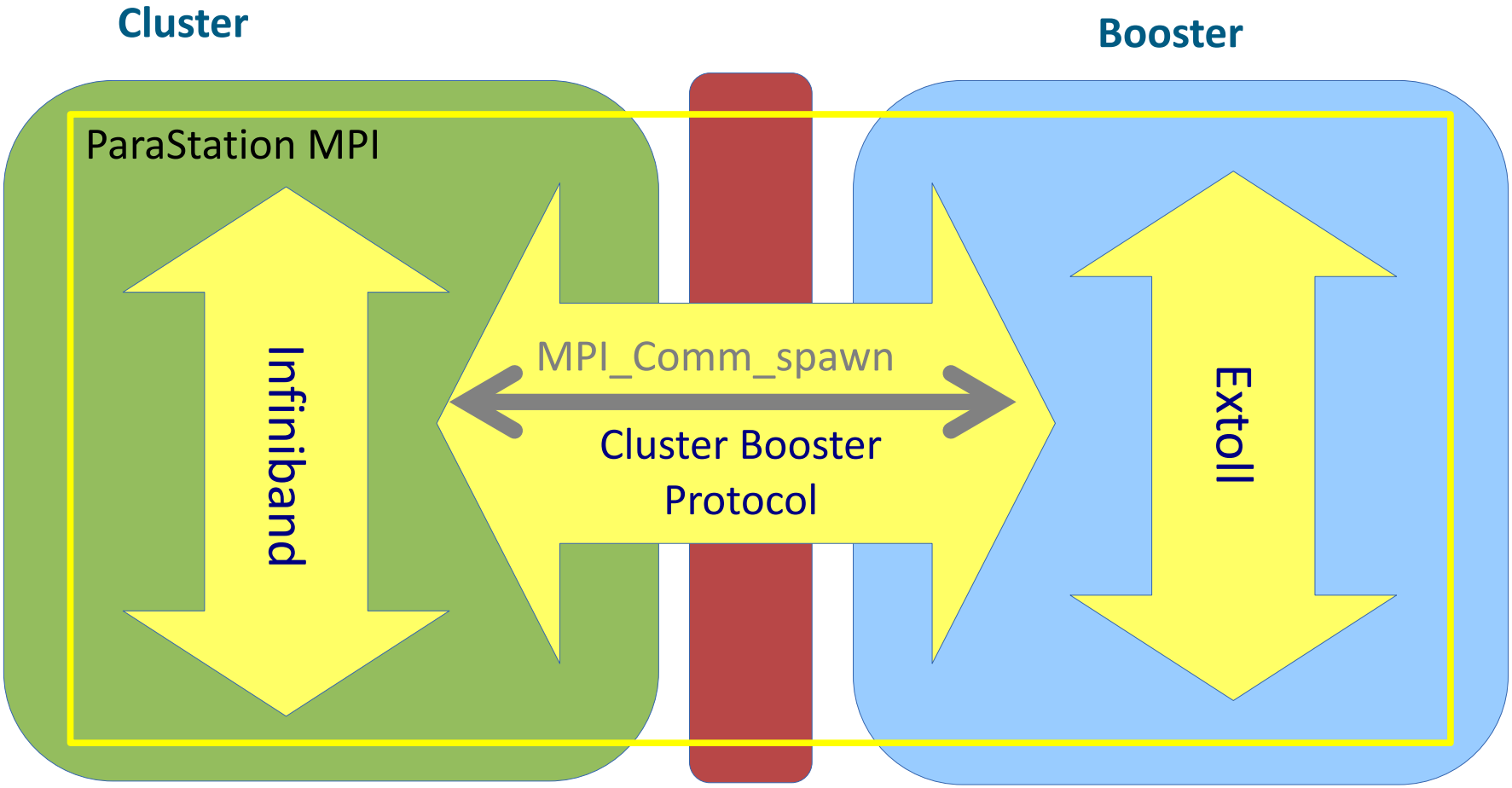
OmpSs on top of MPI provides pragmas to ease the offload process

Programming Environment



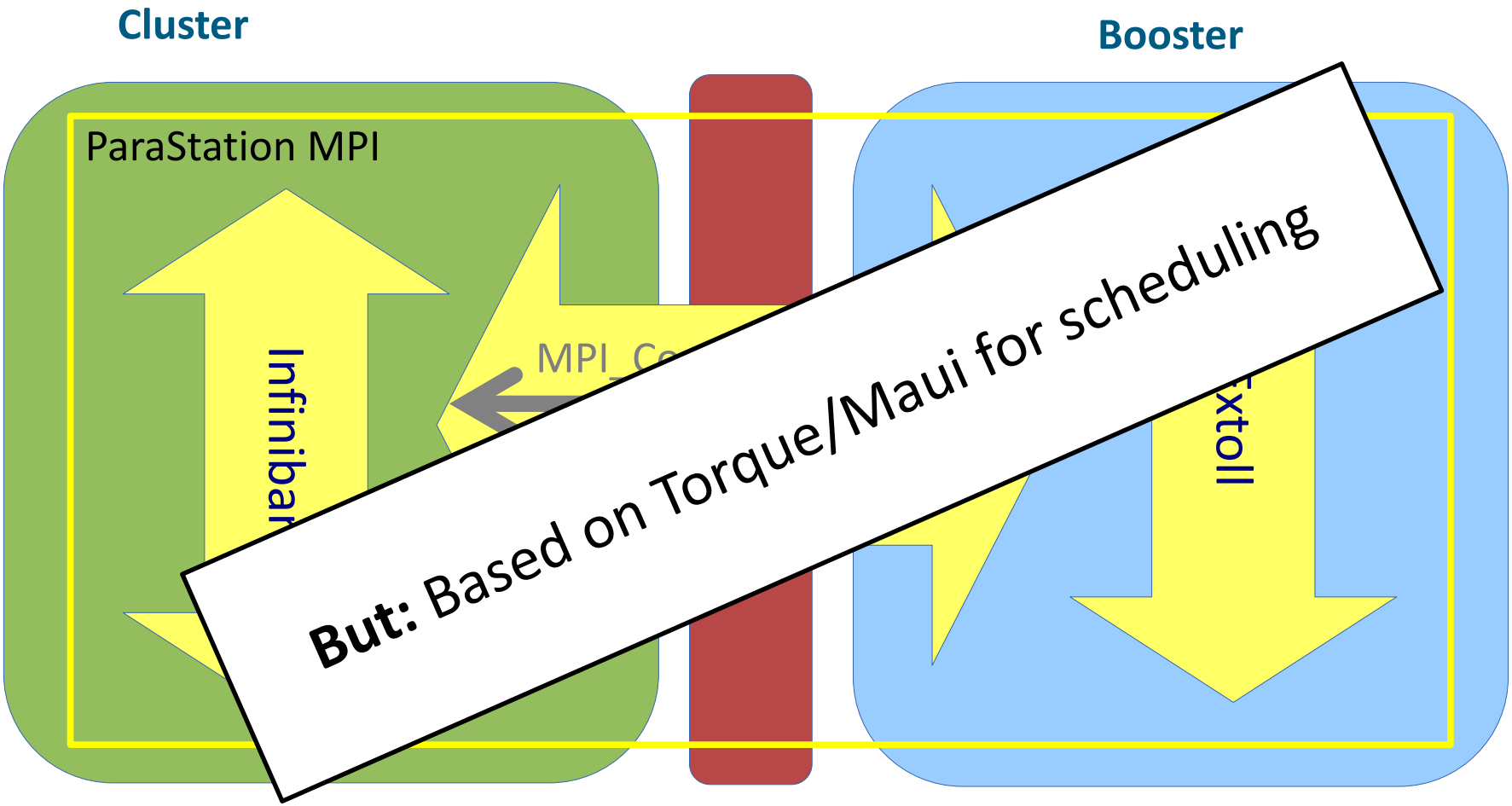
OmpSs on top of MPI provides pragmas to ease the offload process

Programming Environment



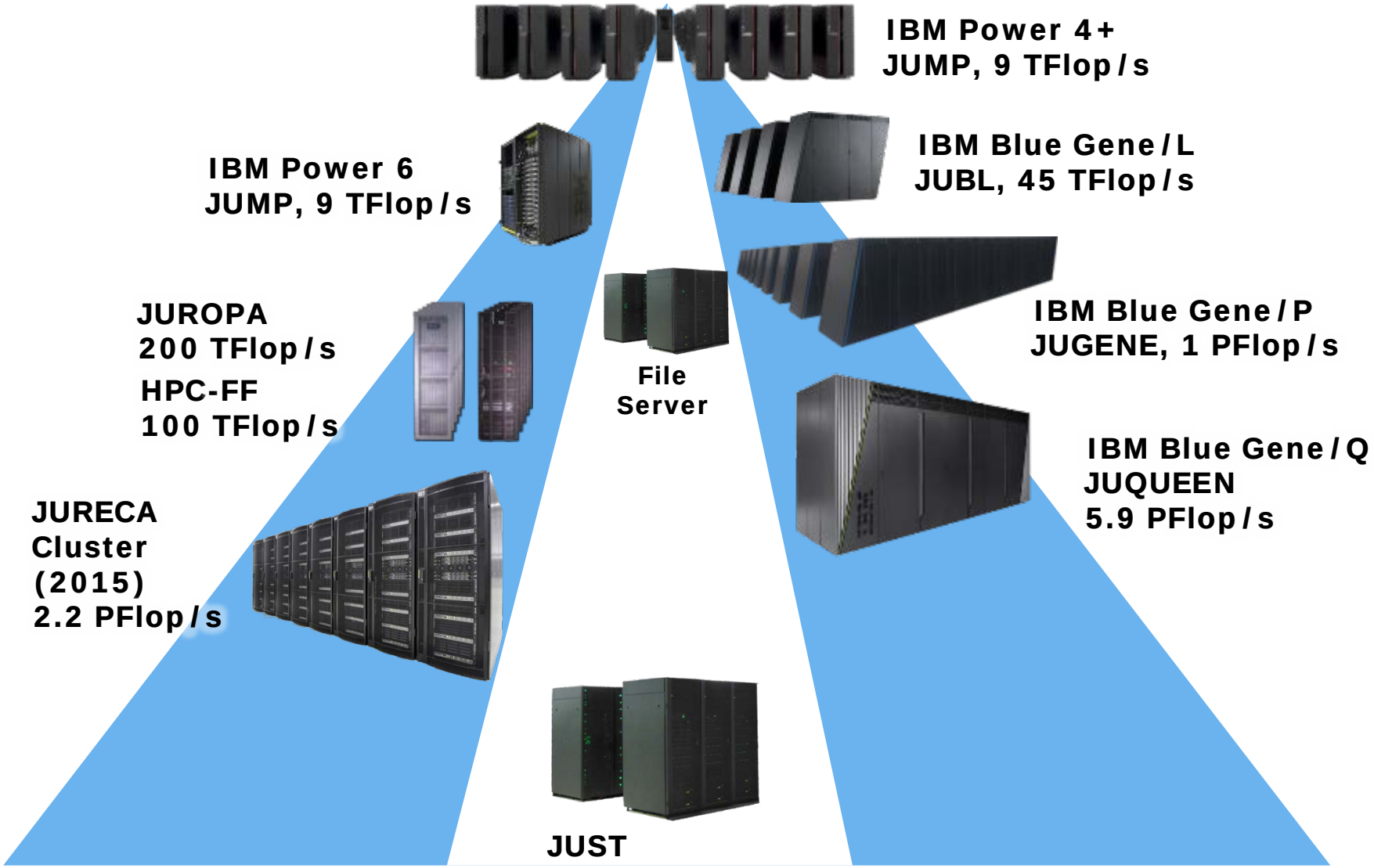
OmpSs on top of MPI provides pragmas to ease the offload process

Programming Environment

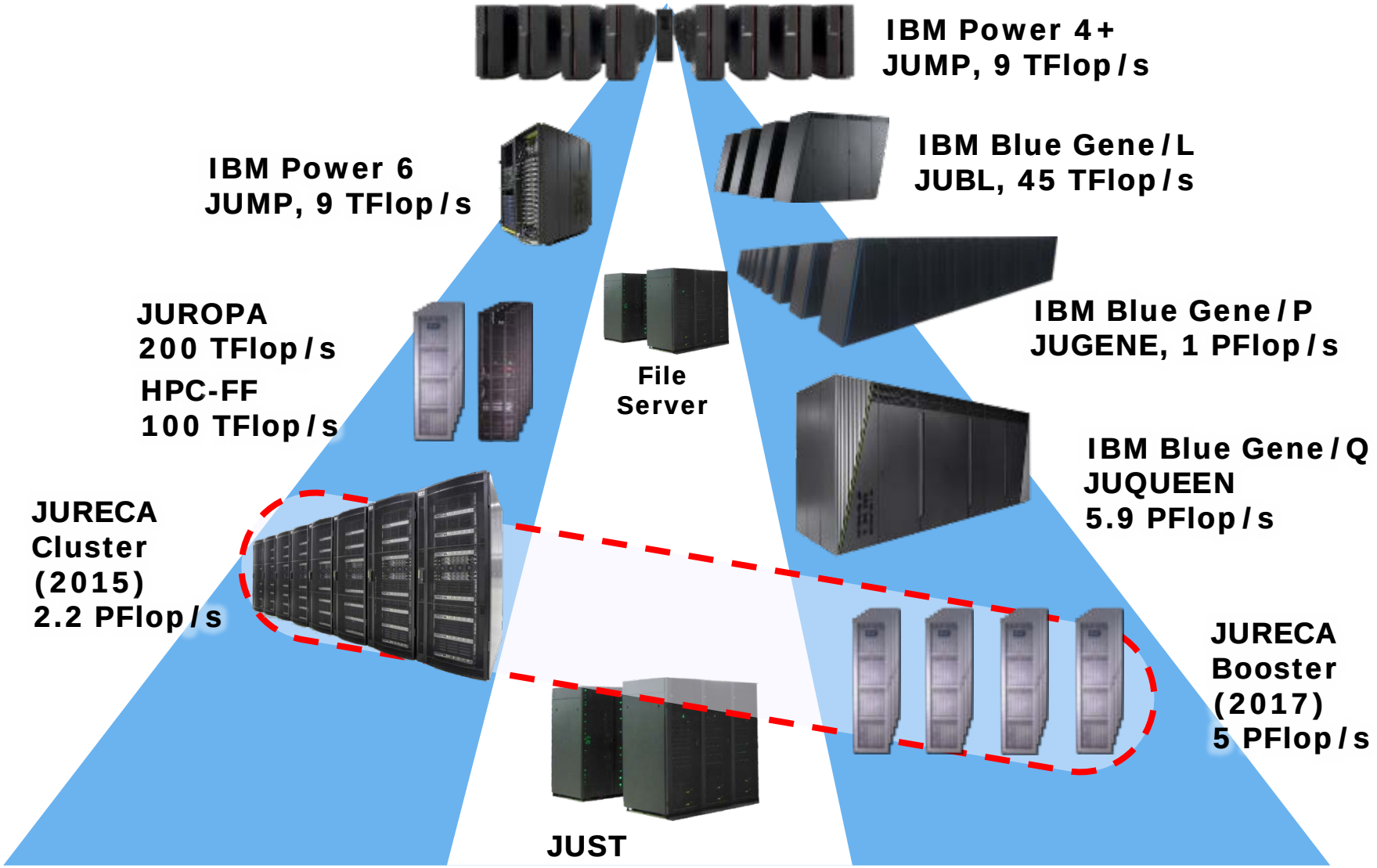


OmpSs on top of MPI provides pragmas to ease the offload process

Dual-Architecture Supercomputing Facility



Dual-Architecture Supercomputing Facility



JURECA Cluster-Booster System

JURECA Cluster

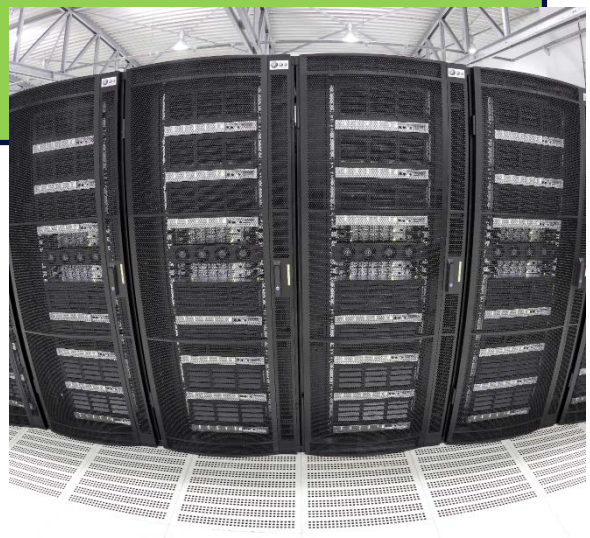
1884 nodes
Dual-Socket Xeon
Mellanox EDR IB
2.2 PF/s peak

Bridging: 20 Tb/s

Booster

1640 nodes
Xeon Phi 7250-F
Intel OPA full fat-tree
5 PF/s peak

198 bridge nodes



JURECA Cluster-Booster System

JURECA Cluster

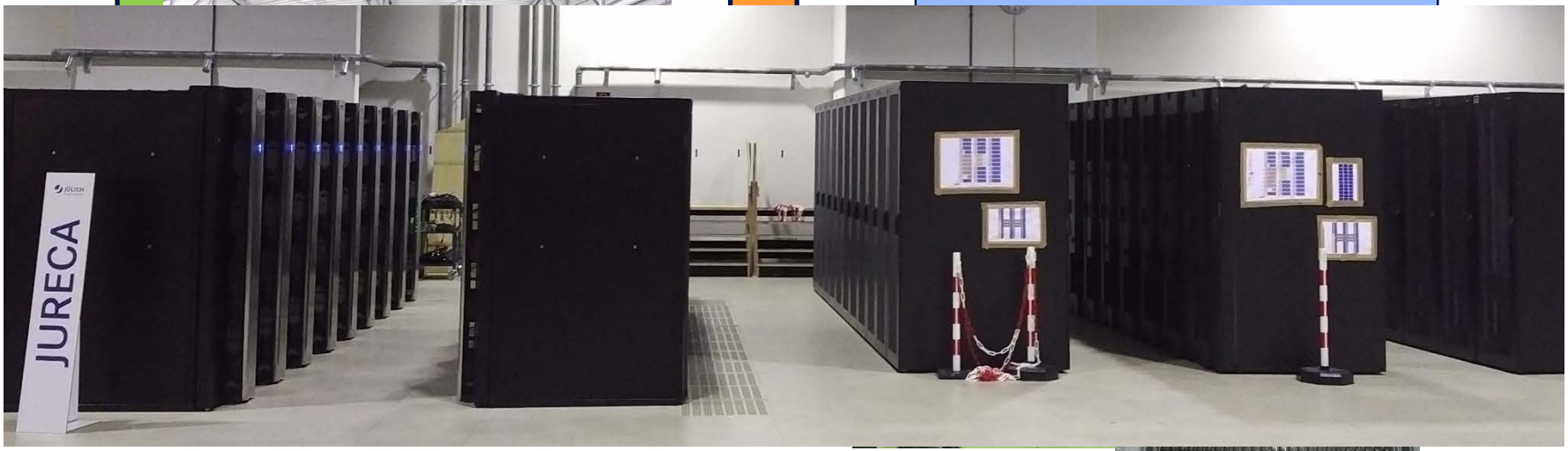
1884 nodes
Dual-Socket Xeon
Mellanox EDR IB
2.2 PF/s peak

Bridging: 20 Tb/s

Booster

1640 nodes
Xeon Phi 7250-F
Intel OPA full fat-tree
5 PF/s peak

198 bridge nodes



JURECA Cluster-Booster System

JURECA Cluster

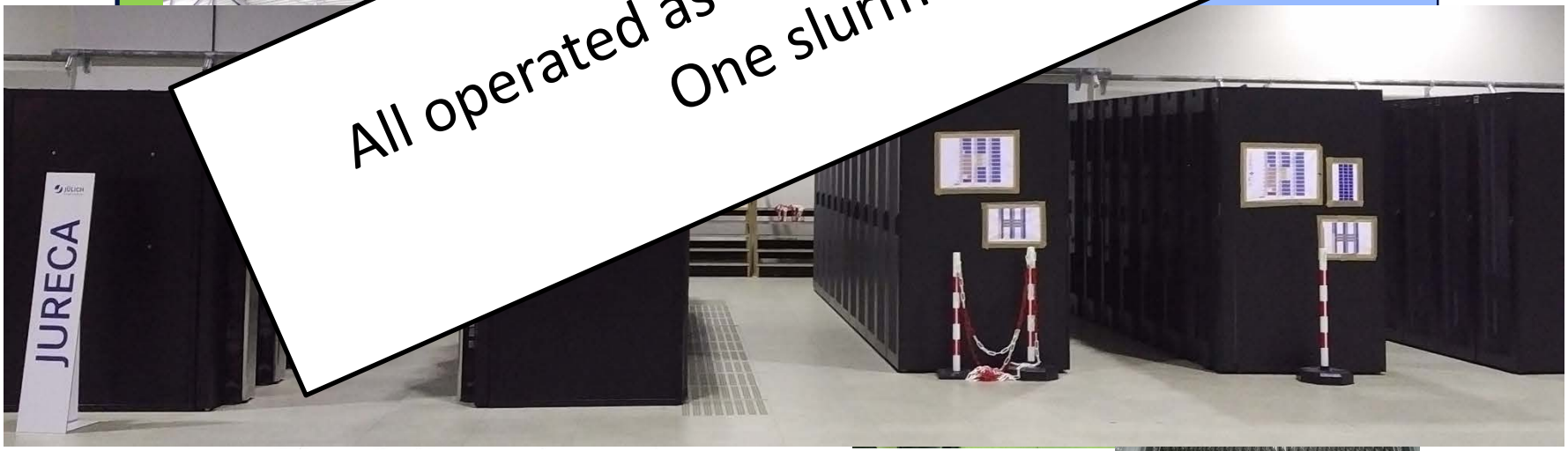
1884 nodes
Dual-Socket Xeon
Mellanox EDR IB
2.2 PF/s peak

Booste

1640 n

20 Tb/s

All operated as one system with Slurm:
One slurmctld!



JURECA Cluster-Booster System

JÜLICH FORSCHUNGSZENTRUM

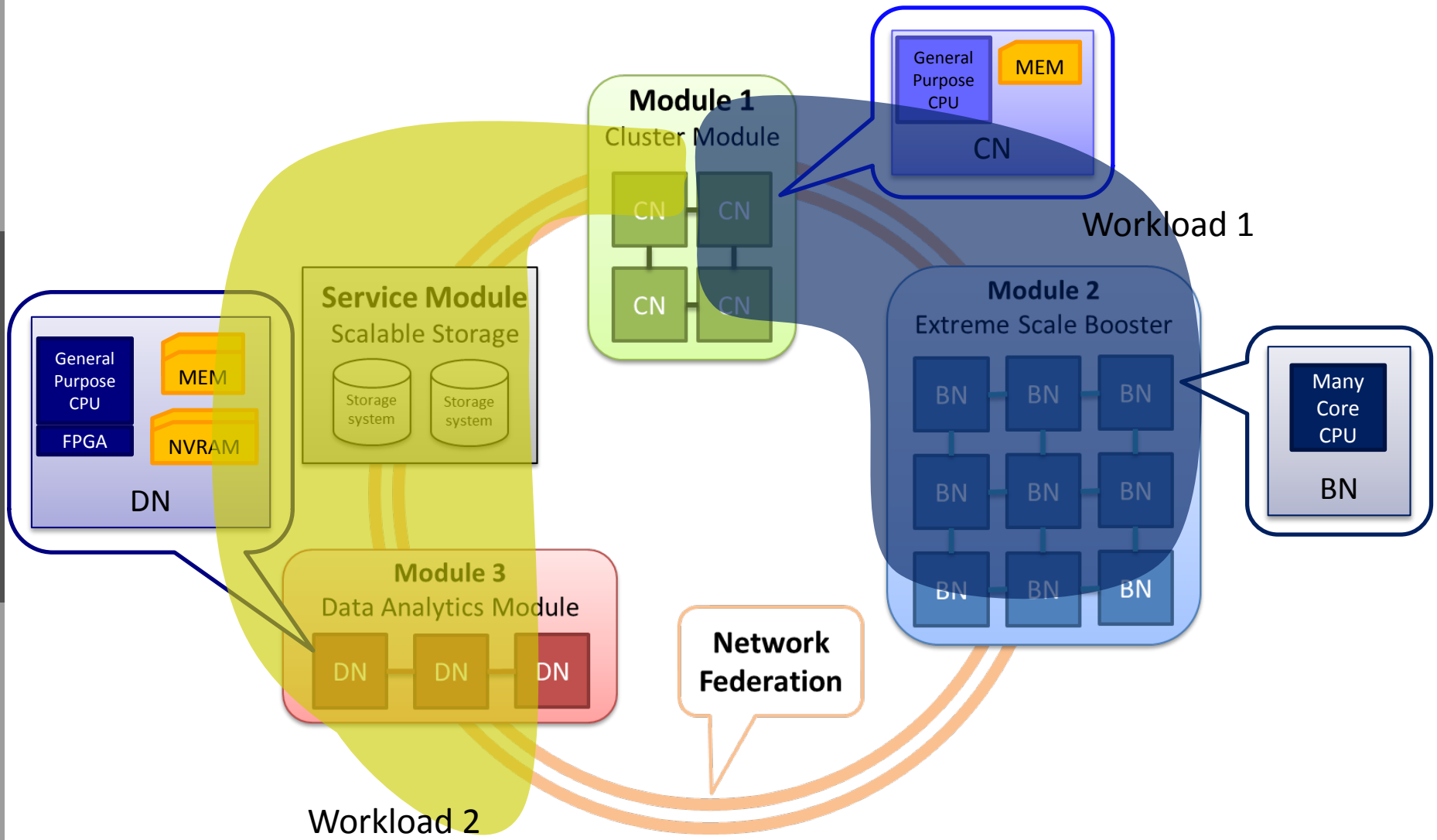
Building blocks

Cluster Hardware	✓
Cluster Software	✓
Booster Hardware	✓
Booster Software	✓
Cluster-Booster Bridging for MPI	✓
Workload Management	✓

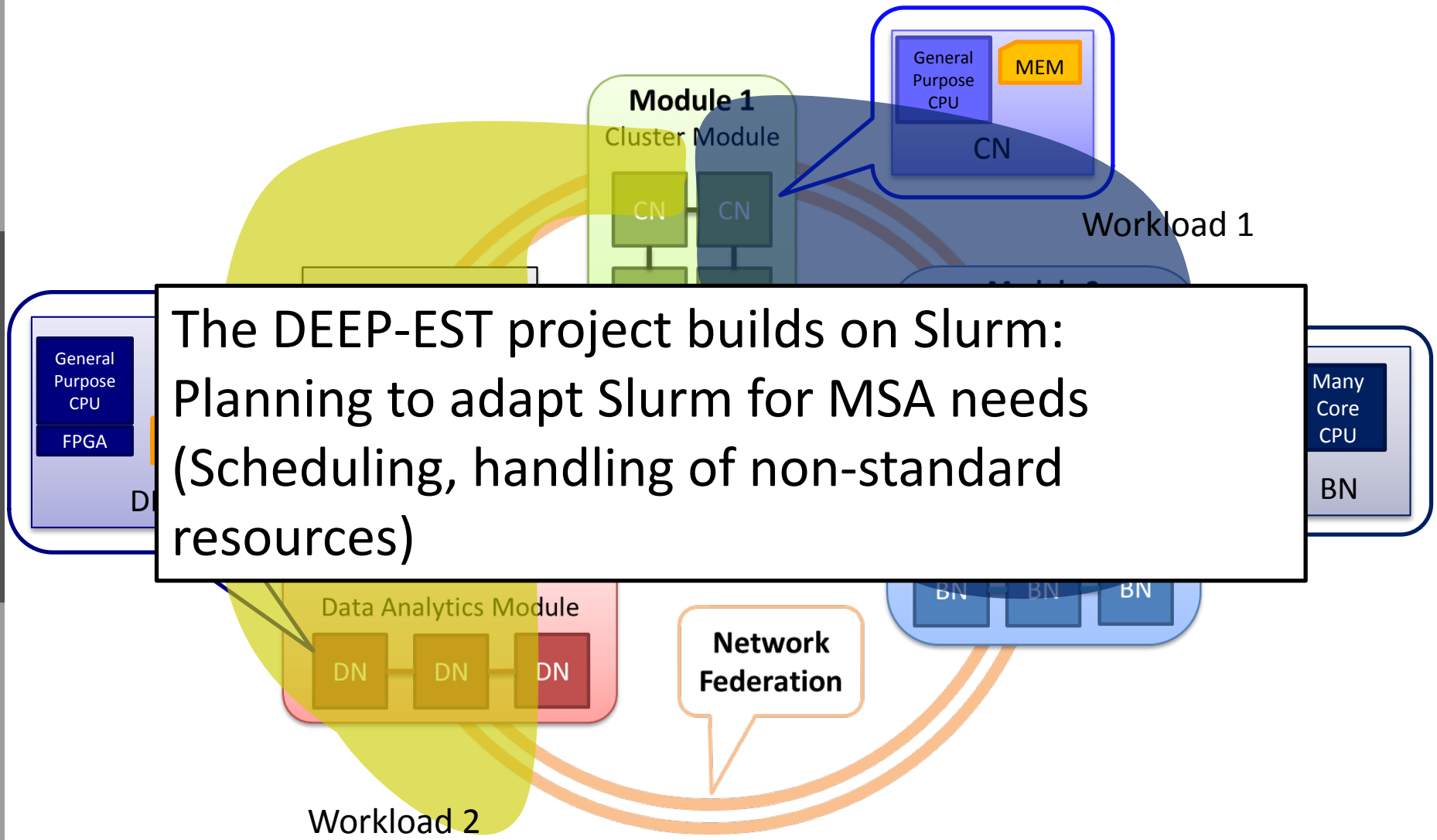
Slurm requirements

- Support for heterogeneity, QoS per partition ✓
- Support for flexible heterogeneous jobs across partitions

Next step: Modular Supercomputing



Next step: Modular Supercomputing



The DEEP-EST project builds on Slurm: Planning to adapt Slurm for MSA needs (Scheduling, handling of non-standard resources)

Summary

- Cluster-Booster Architecture combines multi- and many-core (or other accelerator) technologies in one system and enables a flexible resource selection.
- JURECA is currently extended to realize the Cluster-Booster Architecture in a 7.2 PF/s production system.
- Slurm is a key building block of our software stack. The community and SchedMD help us achieve our goal.
- Working towards Modular Supercomputing Architecture in DEEP-EST using Slurm. We hope to be able to contribute back.