DEEP and DEEP-ER

Estela Suarez
Jülich Supercomputing Centre

10.05.2016

The research leading to these results has received funding from the European Community's Seventh Framework Programme (*FP7/2007-2013*) under *Grant Agreement* n° 287530 and n° 610476

Topics

DEEP

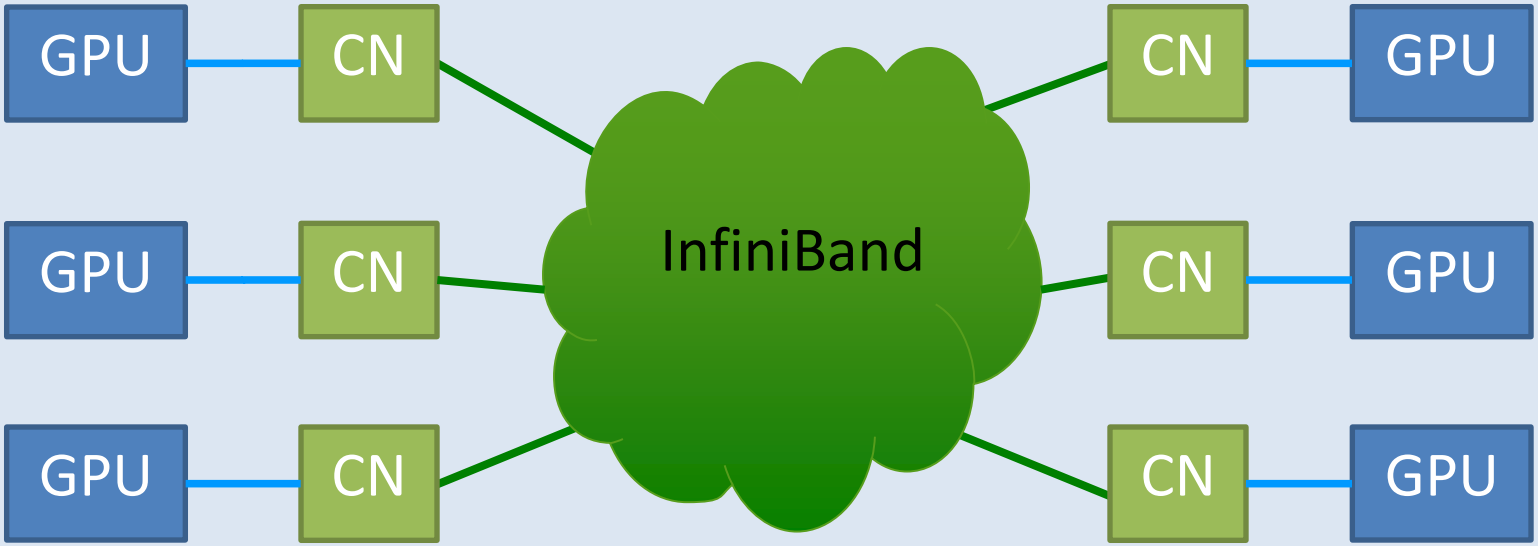
- **Cluster-Booster archit.**
- Software stack
- Programming environ.
- Energy efficiency
- Applications:
  - Co-design
  - Evaluation/demonstration
  - Code modernisation

DEEP-ER

- Extend memory hierarchy
- High-performance **I/O**
- Scalable **resiliency**
- Applications:
  - Co-design
  - Evaluation/demonstration
  - Code modernisation

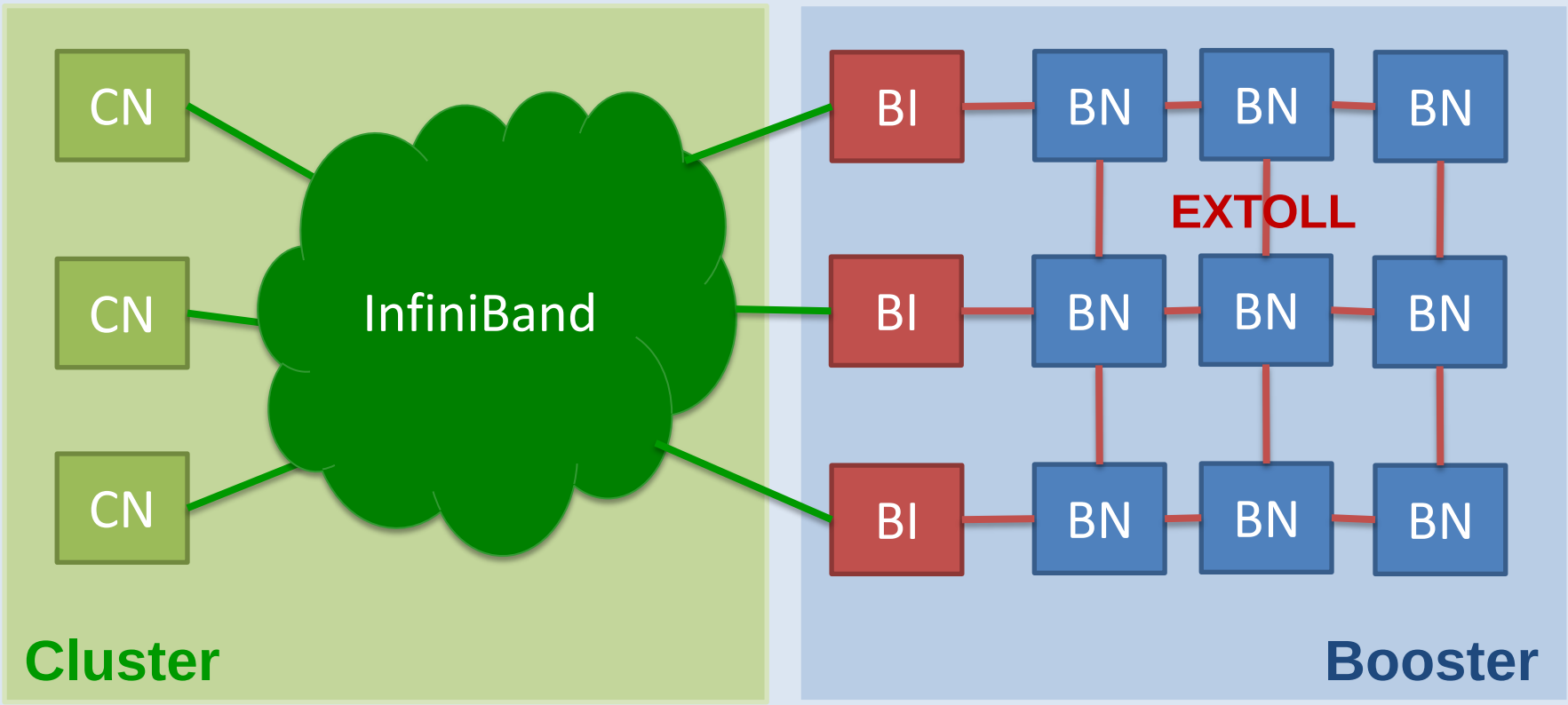
CLUSTER-BOOSTER ARCHITECTURE

“Standard” heterogeneity



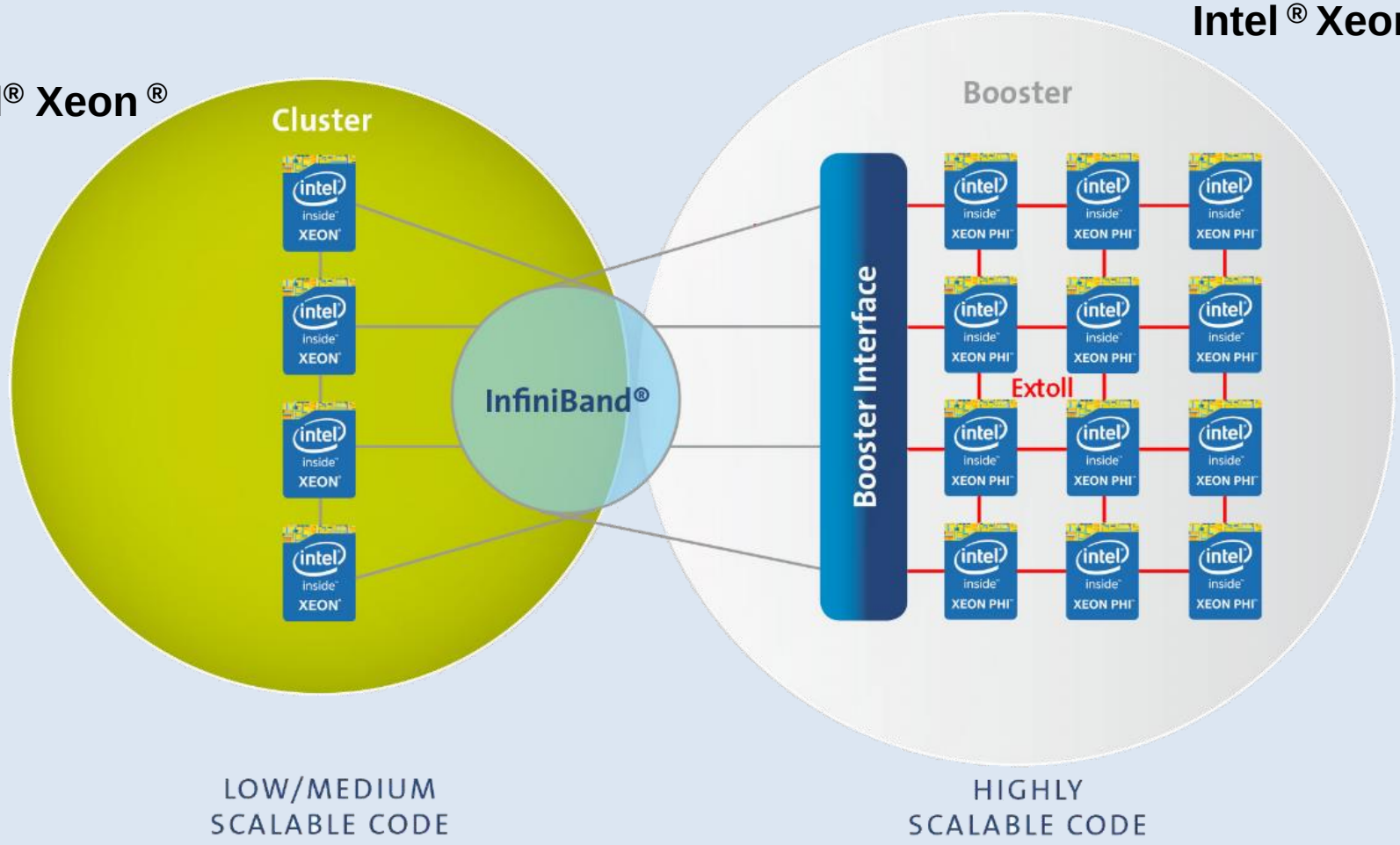
- Flat topology
- Simple management of resources
- Static assignment of accelerators to CPUs
- Accelerators cannot act autonomously

Cluster-Booster architecture



Flexible assignment of resources (CPUs, accelerators)
Direct communication between accelerators
“Offload” of large and complex parts of applications

DEEP Architecture



DEEP System

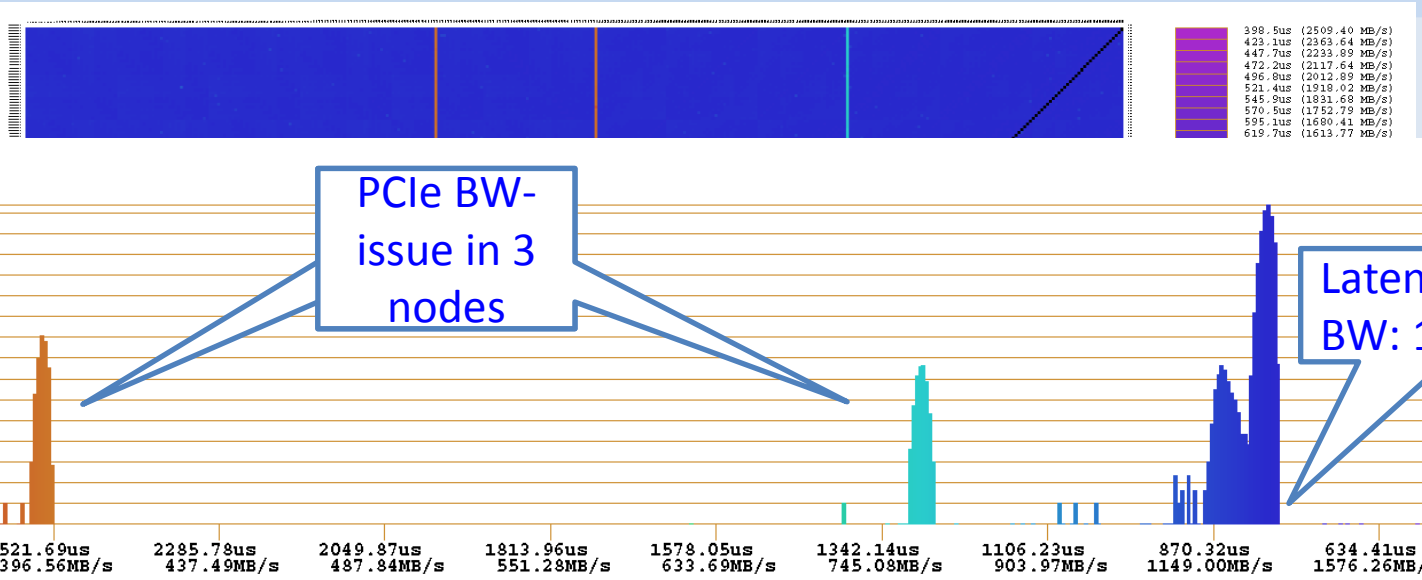
- Installed at JSC
- 1,5 racks
- 500 TFlop/s peak perf.
- 3.5 GFlop/s/W
- Water cooled

Cluster (128 Xeon)

Booster (384 Xeon Phi KNC)

Booster measurements

MPI Linktest: ping-pong



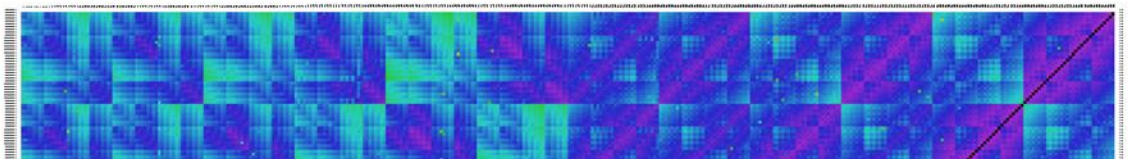
length_of_message:	1048576 bytes (1024.00 KBytes)	number_of_tasks:	384
number_of_messages:	25	Execution order:	Serial
Alltoall:	1	Mixing PE rank:	No
Min Value:	398.5us (2509.40 MB/s)	Alltoall Min Value:	616.1us (1 Byte)
Max Value:	2757.6us (362.63 MB/s)	Alltoall Max Value:	5038.0us (1 Byte)
Avg Value:	814.9us (1227.12 MB/s)	Alltoall Avg Value:	795.1us (1 Byte)

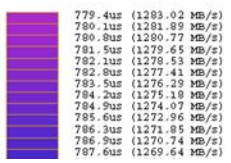
Report generated by FZJ Linktest Result Analyzer, Forschungszentrum Juelich GmbH

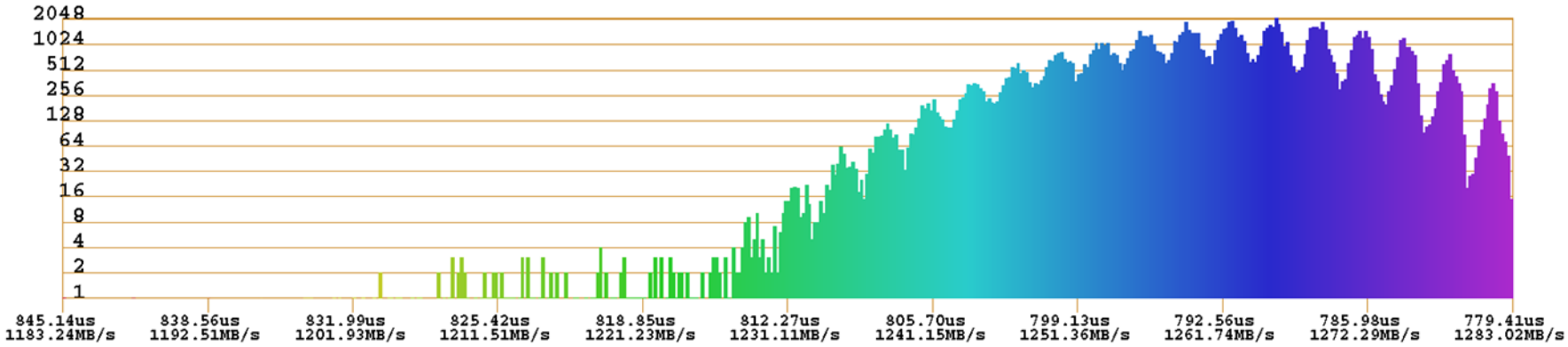
Booster Nodes (KNC) from 1 to 384

Booster measurements

MPI Linktest: ping-pong



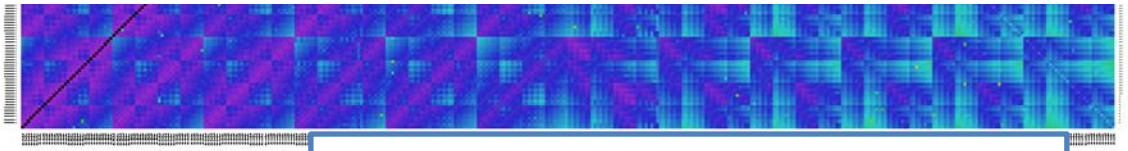


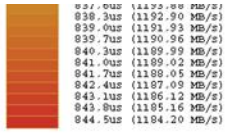


length_of_message:	1048576 bytes (1024.00 KBytes)	number_of_tasks:	384
number_of_messages:	40	Execution order:	Serial
Alltoall:	0	Mixing PE rank:	No
Min Value:	779.4us (1283.02 MB/s)		
Max Value:	845.1us (1183.24 MB/s)		
Avg Value:	792.7us (1261.47 MB/s)		

Stddev < 10%

Report generated by FZJ Linktest Result Analyzer, Forschungszentrum Juelich GmbH





Booster Nodes (KNC) from 1 to 384

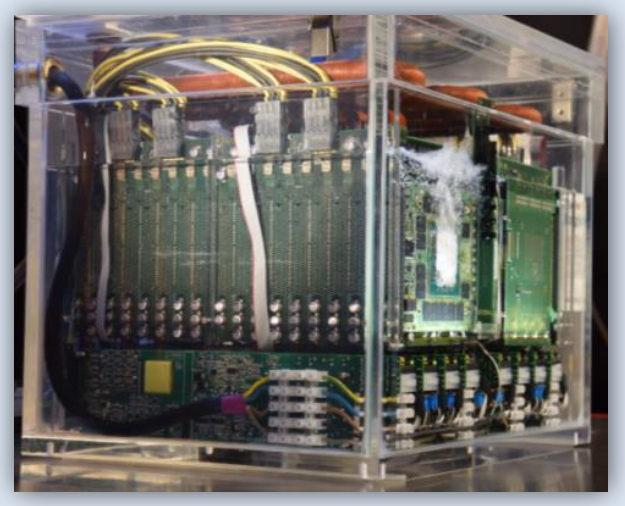
GreenICE system

Alternative Booster implementation

- Interconnect EXTOLL ASIC “Tourmalet”
- 32 KNC-node system
- Implement 4×4×2 topology, with Z dimension open

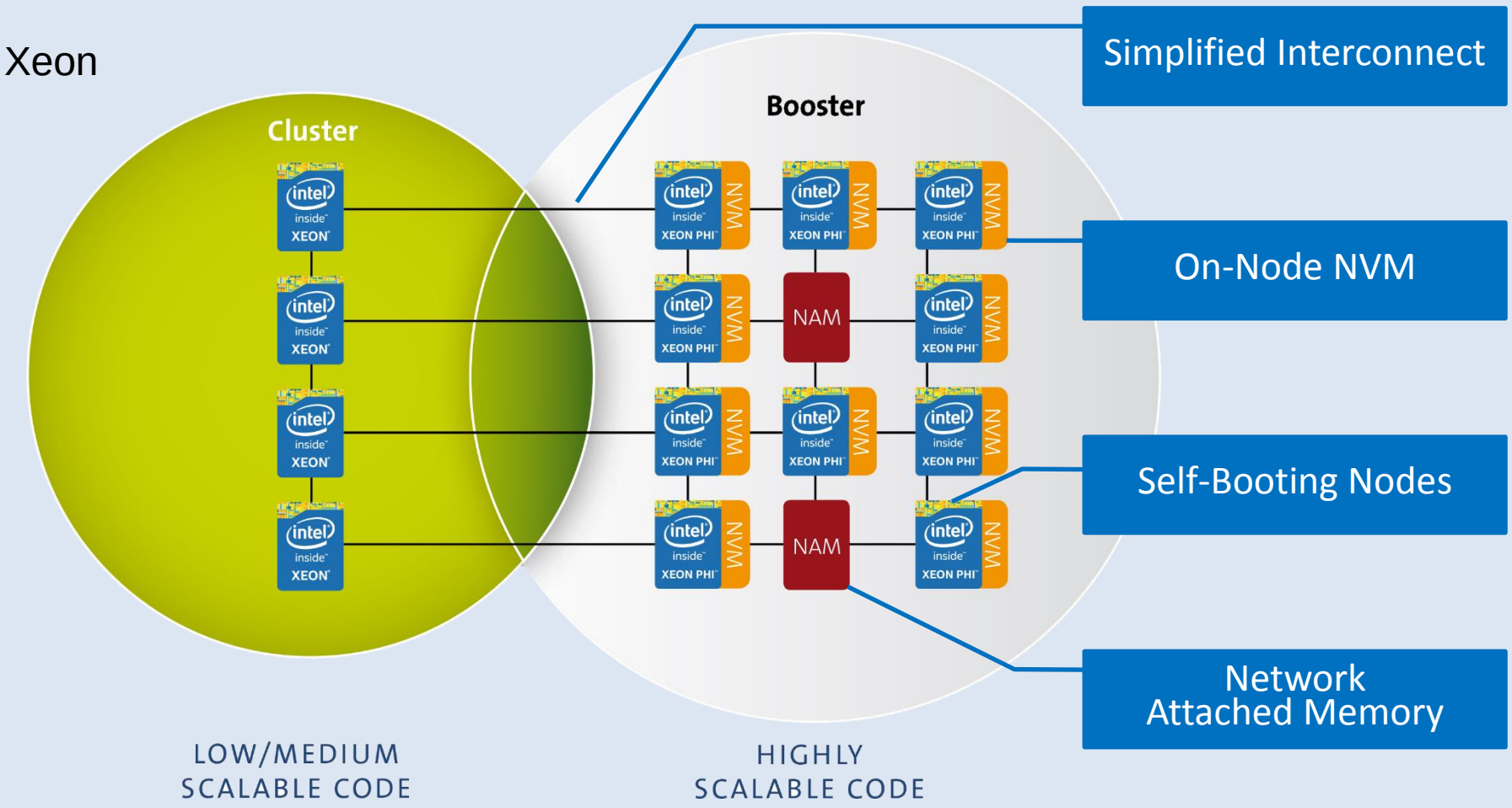
Experiment 2-phase immersion cooling

- NOVEC liquid from 3M
- Evaporates at about 50 degrees
- Condensates again in a water cooling pipe
- Allows very high-density integration



GreenICE Booster

DEEP-ER Architecture
Innovation



DEEP-ER Aurora Blade prototype

Eurotech's Aurora technology

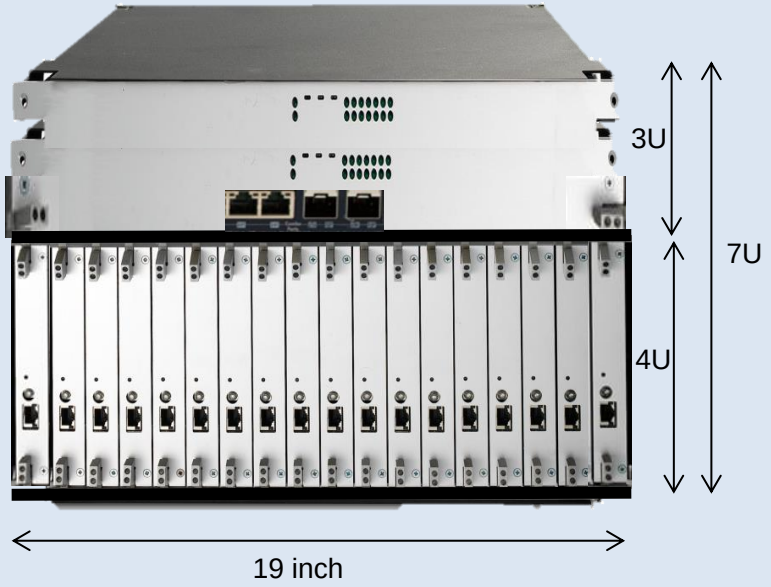
Direct water cooled, high density

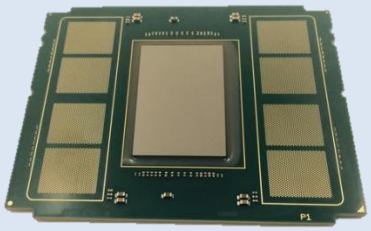
Aurora Blade DEEP-ER Booster
(in construction)

Aurora Blade Chassis

Rootcard
-18x EXTOLL
-18x NVMe

Chassis:
-18x KNL
-94GB Mem
-1x backplane





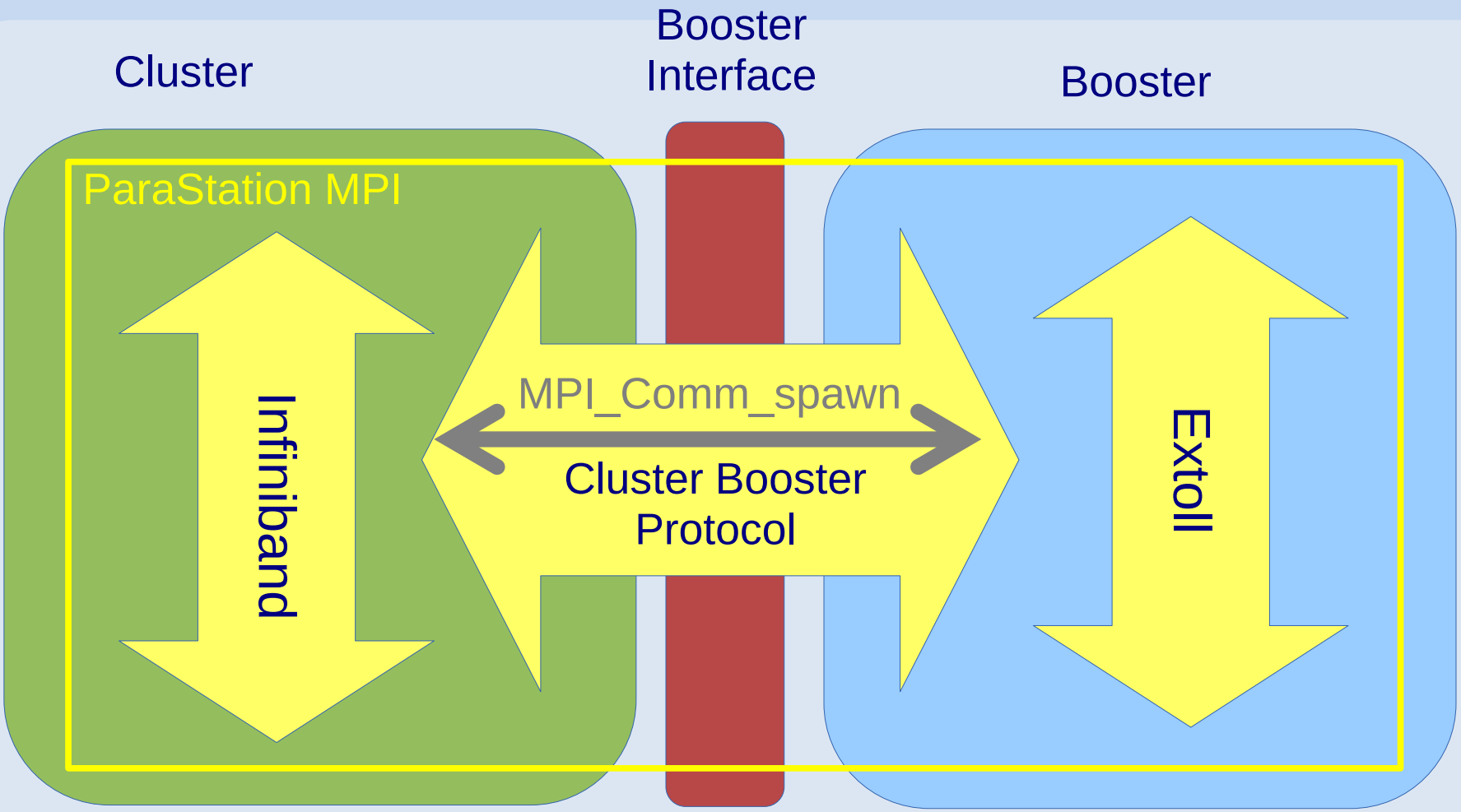
Intel Xeon Phi (KNL)

NVMe

EXTOLL Tourmalet

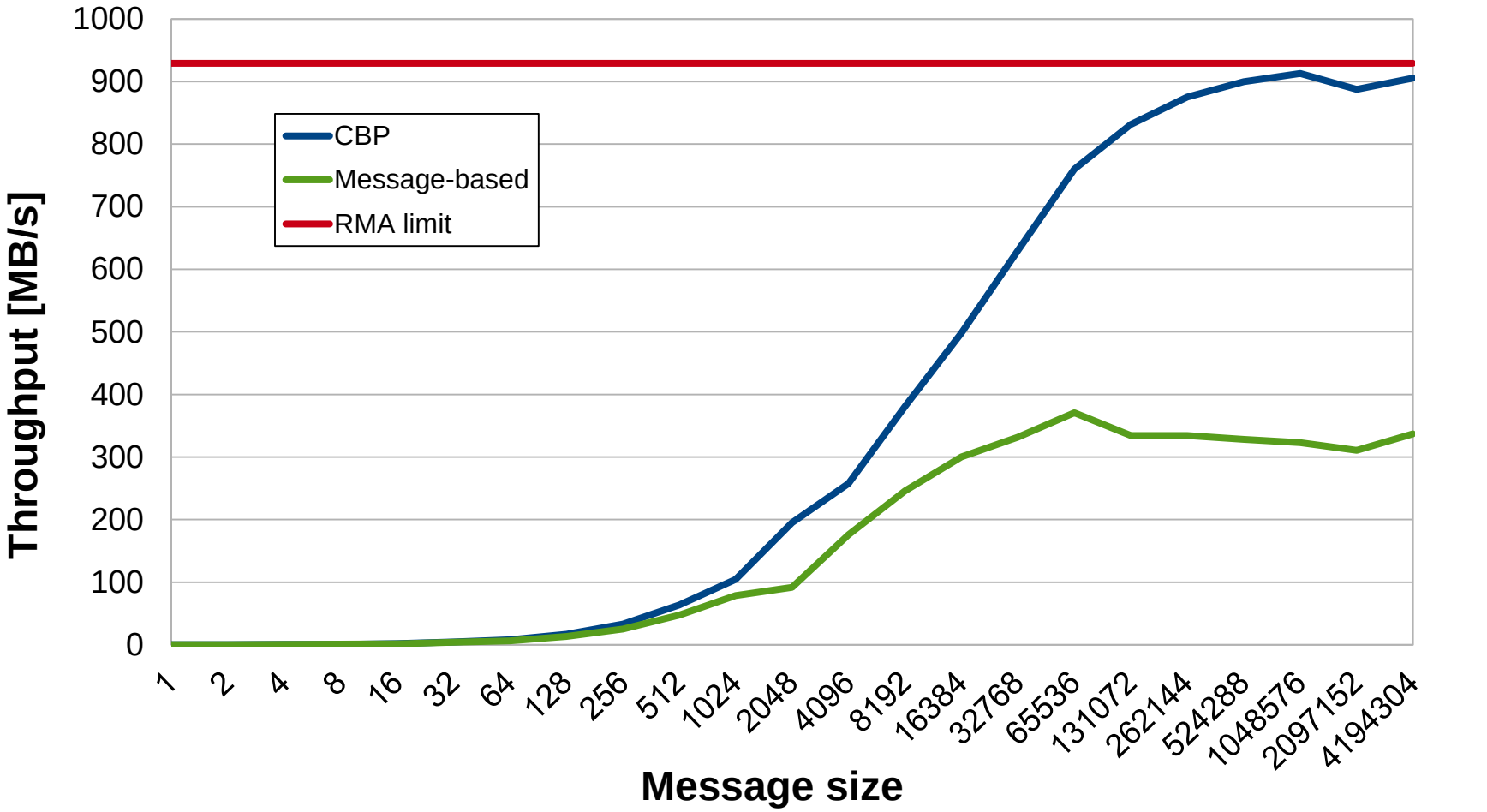
SOFTWARE

Programming environment



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

Cluster-Booster Protocol



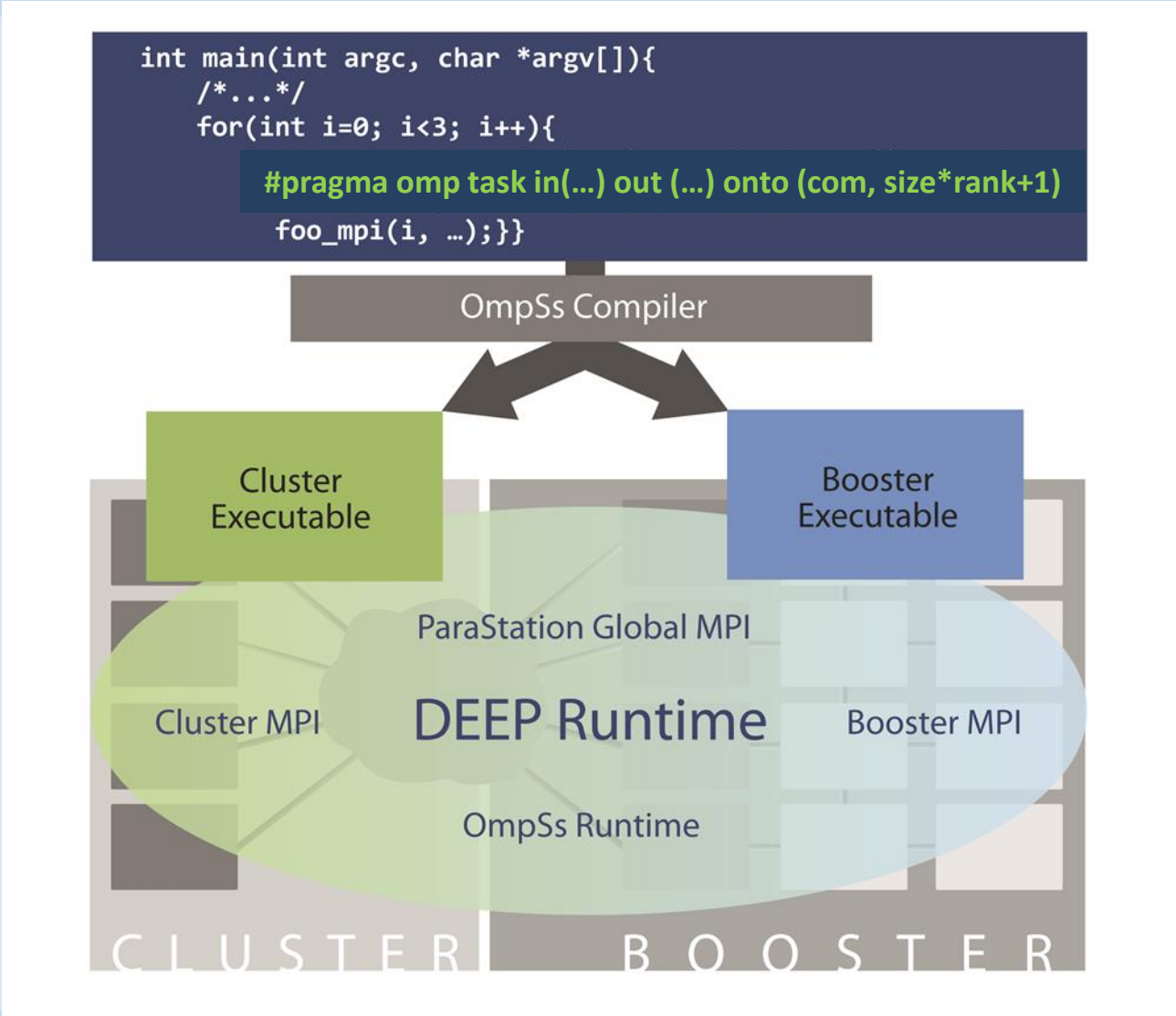
Application running on DEEP

Source code

Compiler

Application binaries

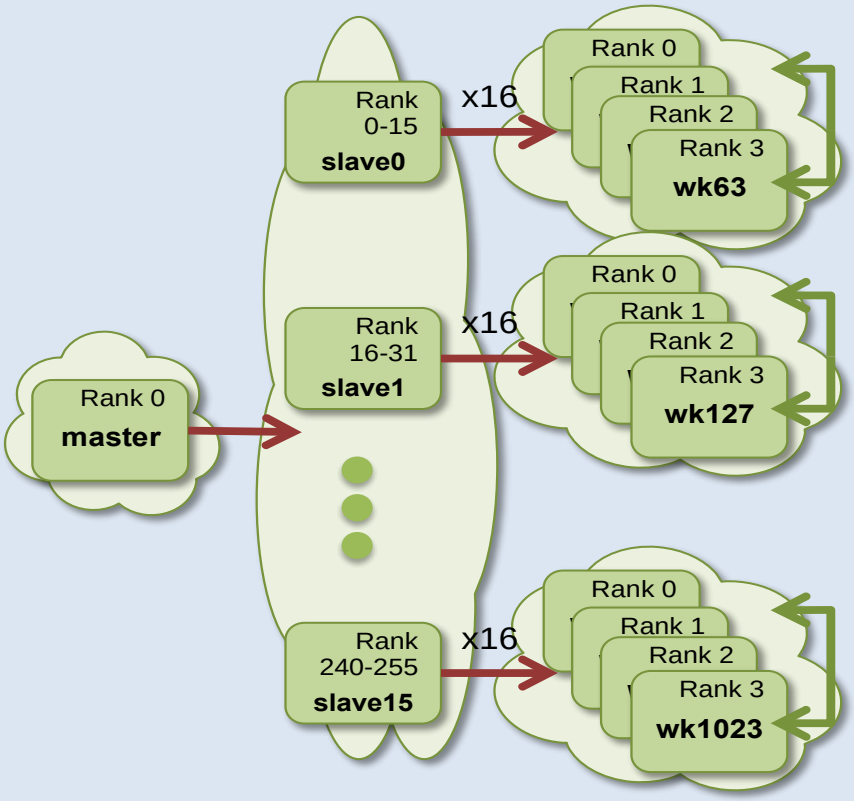
DEEP Runtime

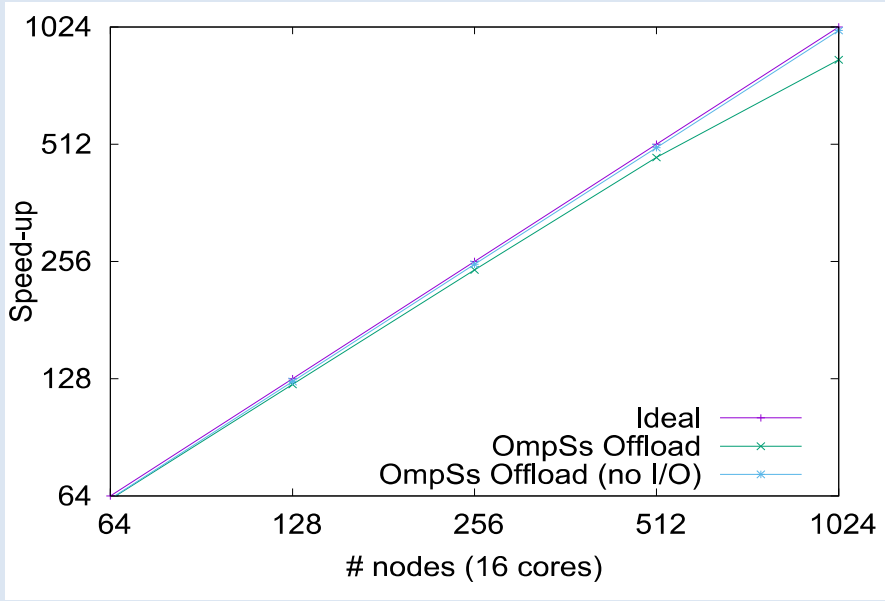


DEEP Offload (with OmpSs)

Performance & Scalability evaluation

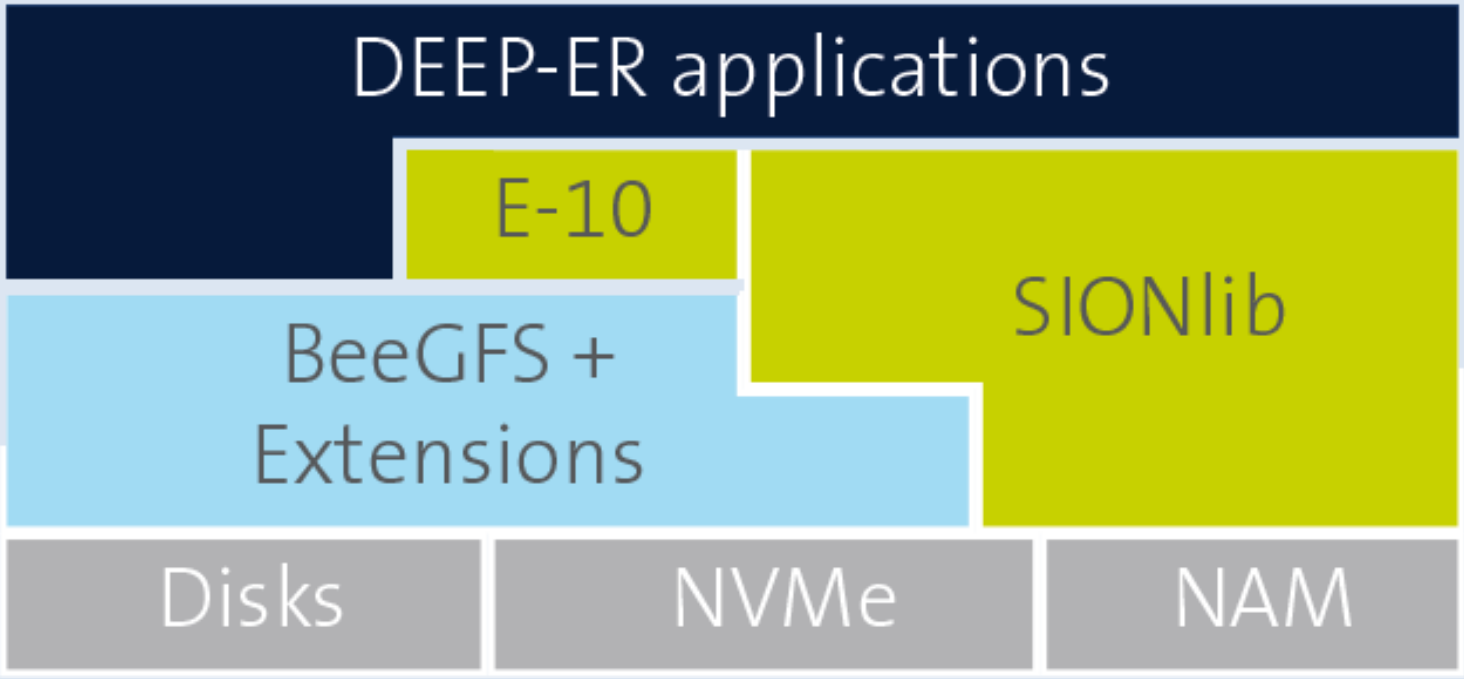
FWI (full wave inversion) code

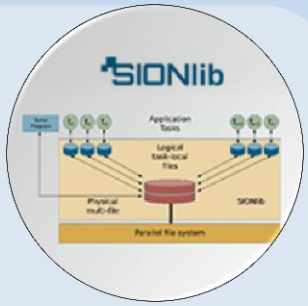




Published in*: "Collective Offload for Heterogeneous Clusters"*, HiPC 2015

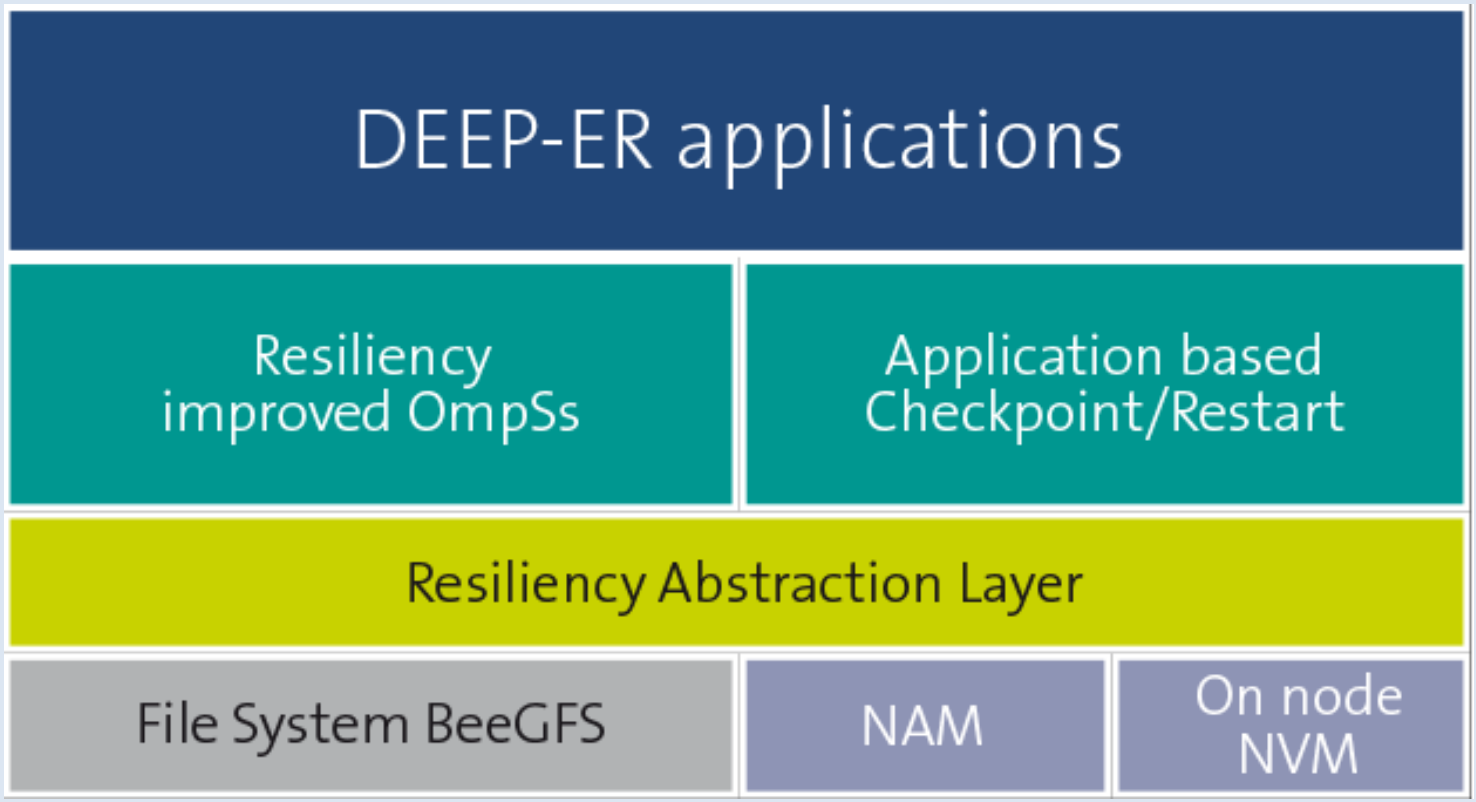
Scalable I/O





- Improve I/O scalability on all usage-levels
- Used also for checkpointing

Resiliency

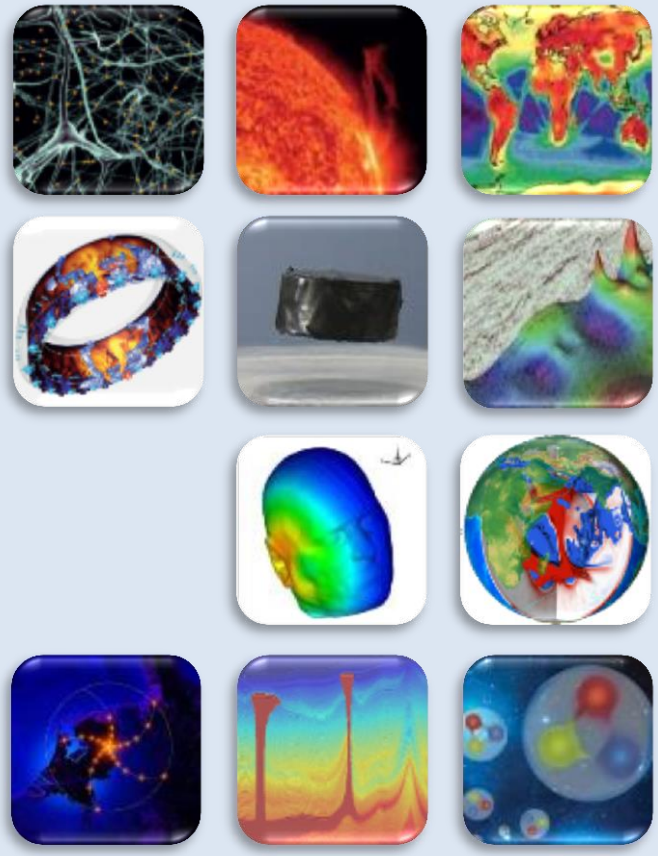


- Develop a hierarchical, distributed checkpoint/restart mechanism leveraging DEEP-ER architecture

APPLICATIONS

Application-driven approach

- **DEEP+DEEP-ER applications:**
  - Brain simulation (EPFL)
  - Space weather simulation (KULeuven)
  - Climate simulation (Cyprus Institute)
  - Computational fluid engineering (CERFACS)
  - High temperature superconductivity (CINECA)
  - Seismic imaging (CGG)
  - Human exposure to electromagnetic fields (INRIA)
  - Geoscience (LRZ Munich)
  - Radio astronomy (Astron)
  - Oil exploration (BSC)
  - Lattice QCD (University of Regensburg)



- **Goals:**
  - Co-design and evaluation of architecture and its programmability
  - Analysis of the I/O and resiliency requirements of HPC codes

Cluster-Booster Advantages

- More **flexible** than a standard architecture

  → This enables different use models:

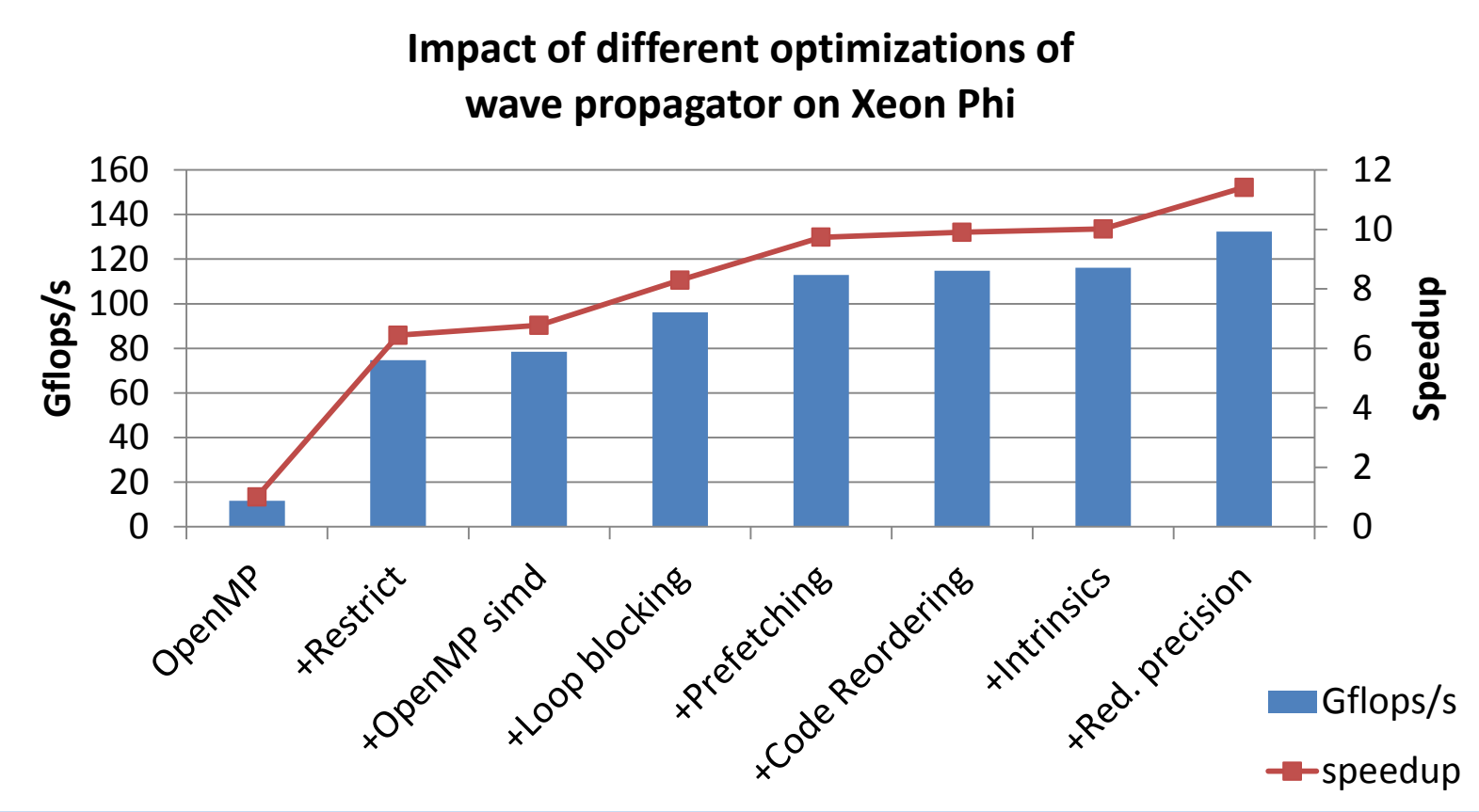
  1. Dynamic ratio of processors/coprocessors
  2. Use Booster as pool of accelerators (globally shared)
  3. Discrete use of the Booster
  4. Discrete use + I/O offload
  5. Specialized symmetric mode

- Enables a **more efficient use of system resources**
  - Only resources actually needed are blocked by applications
  - Dynamic allocation further increases system utilization

Code optimisations

BSC: Enhancing Oil Exploration (FWI, wave propagator)

1 XeonPhi (60 cores), 180 OpenMP threads

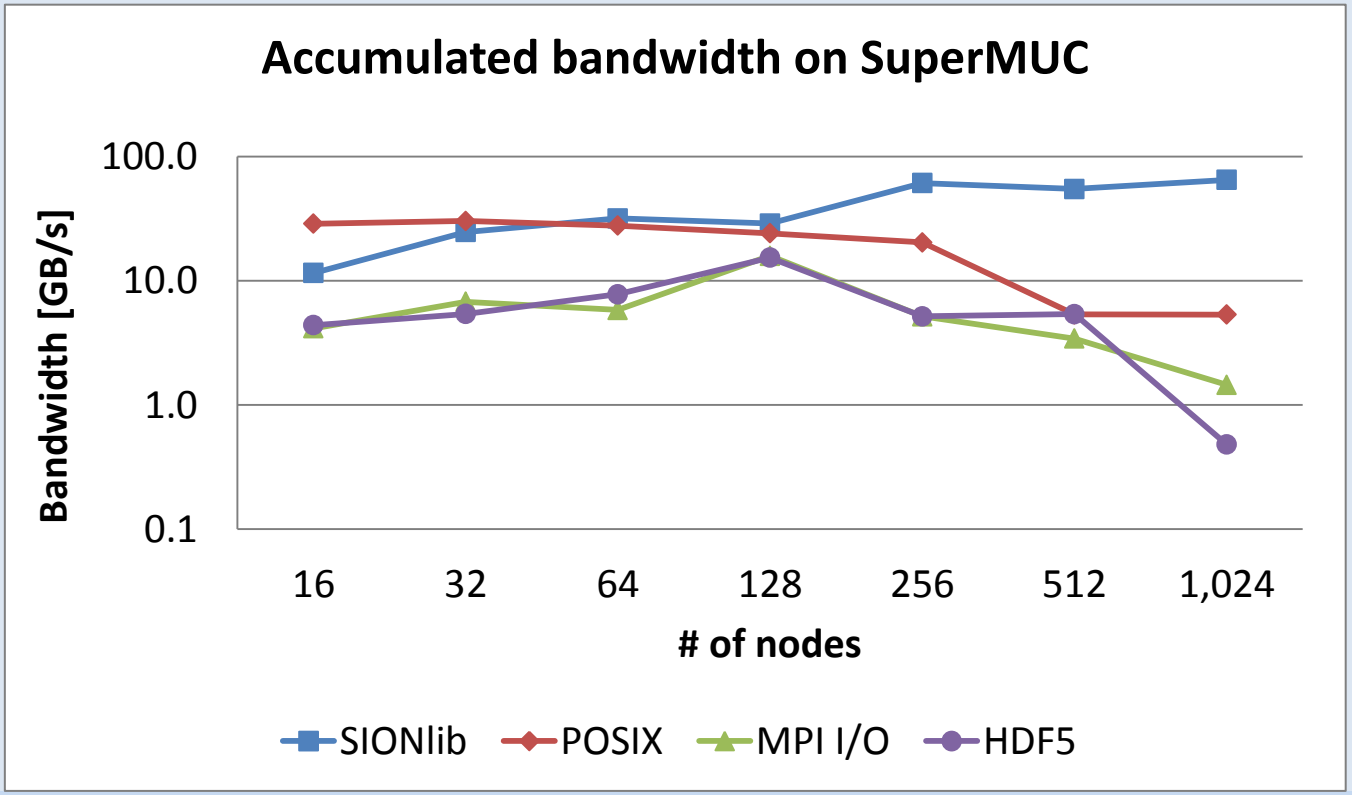


Using SIONlib

LRZ: Rapid crustal deformation & earthquake source equation (Seisol)

1 process per node, 16 threads per process

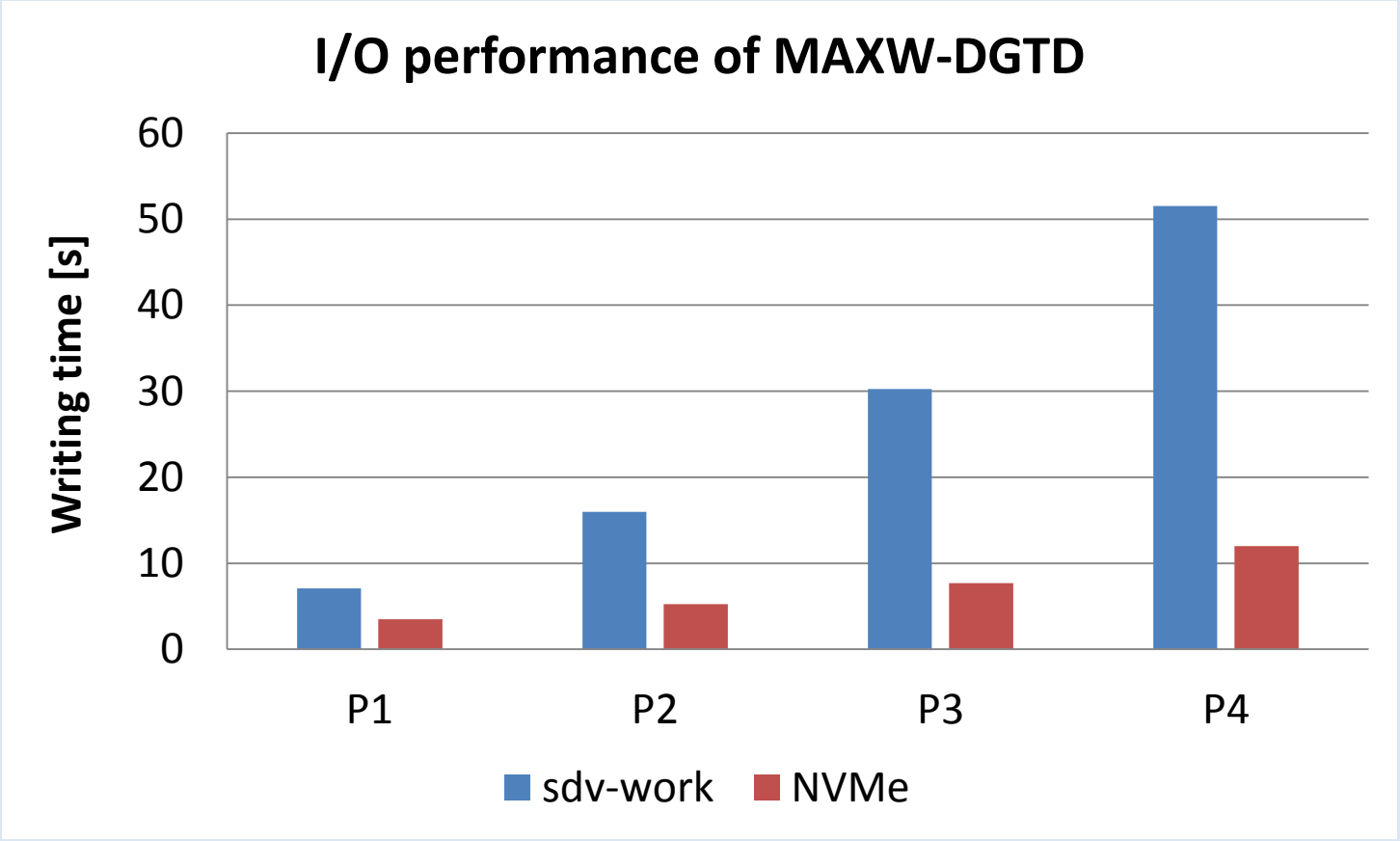
writing 20 checkpoint files (4GB/checkpoint)



Using NVMe

Inria: Assessment of Human exposure to EM fields

24 MPI processes, 1 thread per process



Increasing model precision P1<P2<P3<P4

Summary

- DEEP:
  - Holistic project (HW+SW +App.) in strong co-design
  - Showed the potential of an alternative approach to heterogeneous computing
  - Prototype available also to external users
    - If interested, please contact pmt@deep-project.eu

- DEEP-ER:
  - HW: Prototype under development (impacted by KNL delay)
  - SW + APPs: Very good progress in software and applications
  - APPs: Benefit from new memory technologies shown.
    Strong co-design within the project.

DEEP/-ER contribution

- **To SRA:**
  - HPC System Architecture and component (Cluster-Booster concept, EXTOLL network, memory, energy efficiency, etc.)
  - System Software management: (Interconnect management, ParaStation cluster management, etc.)
  - Programming environment (Heterogeneous programming model (MPI + OmpSs), optimisations for resiliency, etc.)
  - Balance compute, I/O and storage performance (BeeGFS and I/O software to efficiently manage multilevel memory hierarchy, etc.)
  - Energy and resiliency: (Warm water cooling, RAS monitoring, DEEP-ER resiliency, etc.)
  - Application code modernisation
- **To international collaboration:**
  - European Exascale Projects (EEP) initiative

DEEP and DEEP-ER

EU-Exascale projects
20 partners
Total budget: 28,3 M€
EU-funding: 14,5 M€
Nov 2011 – Mar 2017

Visit us @
ISC'16, Frankfurt (Germany)
20.-22.06.2016

-Booth
-BoF
-Workshop

www.deep-project.eu

www.deep-er.eu