DEEP and DEEP-ER

Innovative Exascale Architectures – In the Light of User Requirements

Estela Suarez, Jülich Supercomputing Centre
Mark Tchiboukdjian, CGG
Gabriel Staffelbach, CERFACS

How to Address the Exascale Challenges?

- Power consumption
- Resiliency
- Heterogeneity
- Huge levels of parallelism
- Programmability
- Scalability
- Exploding data requirements
- Algorithms and application readiness

DEEP/-ER Way

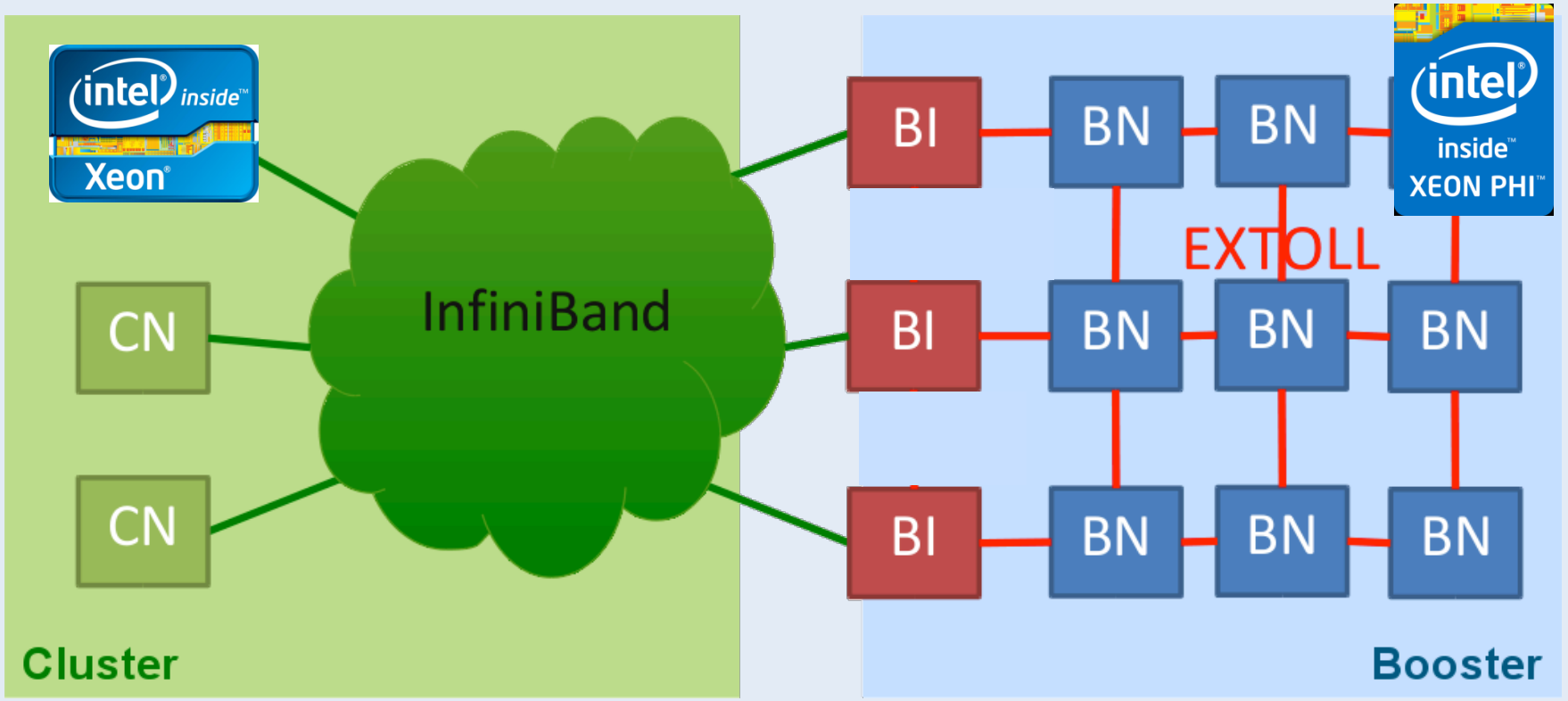
Develop an Exascale architecture tailored to application requirements

- Match HW characteristics with application scalability patterns
- Exploit benefits of processor heterogeneity
- Profit from new memory technologies
- In an overall energy efficient envelope

DEEP-ER Cluster-Booster Architecture

128 Xeon (Sandy Bridge)

384 Xeon Phi (KNC)



Low/Medium scalable code parts

Highly scalable code parts

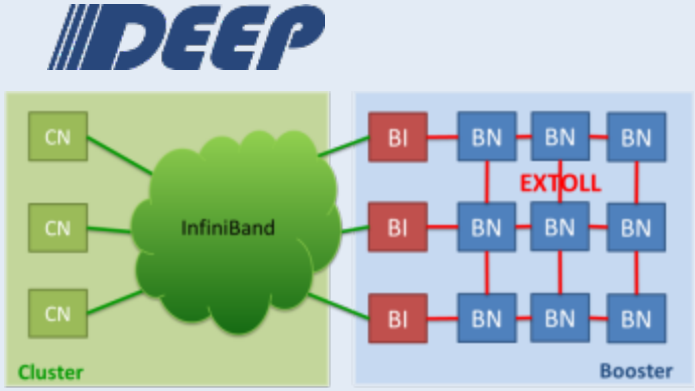
DEEP Hardware

Cluster (128 SB)

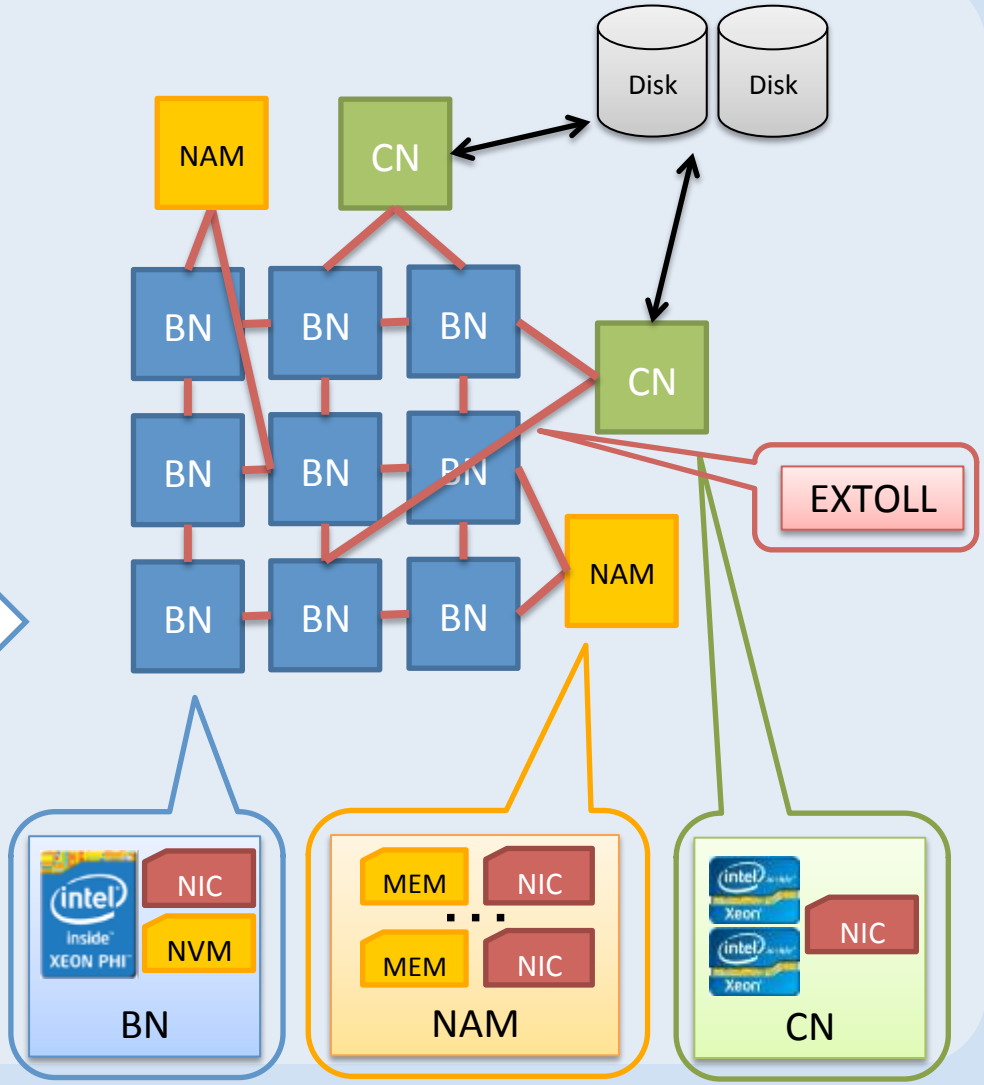
Booster (384 KNCs)

ASIC Evaluator (32 KNCs)

DEEP-ER Enhance DEEP Architecture



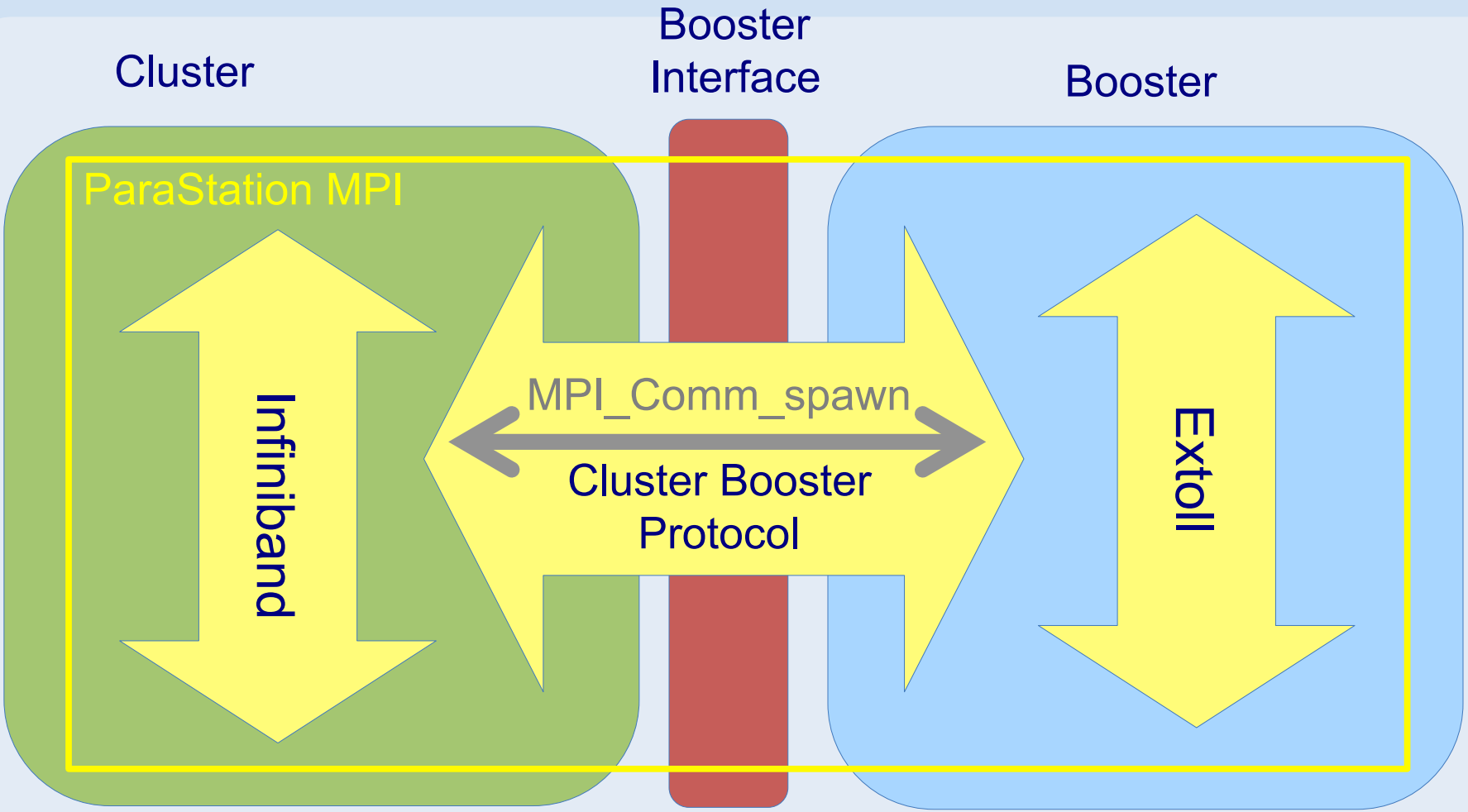
DEEP-ER



Legen

CN:	Cluster Node
BN:	Booster Node
BI:	Booster Interface
NAM:	Network Attached Memory
NVM:	Non Volatile Memory

Programming Environment



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

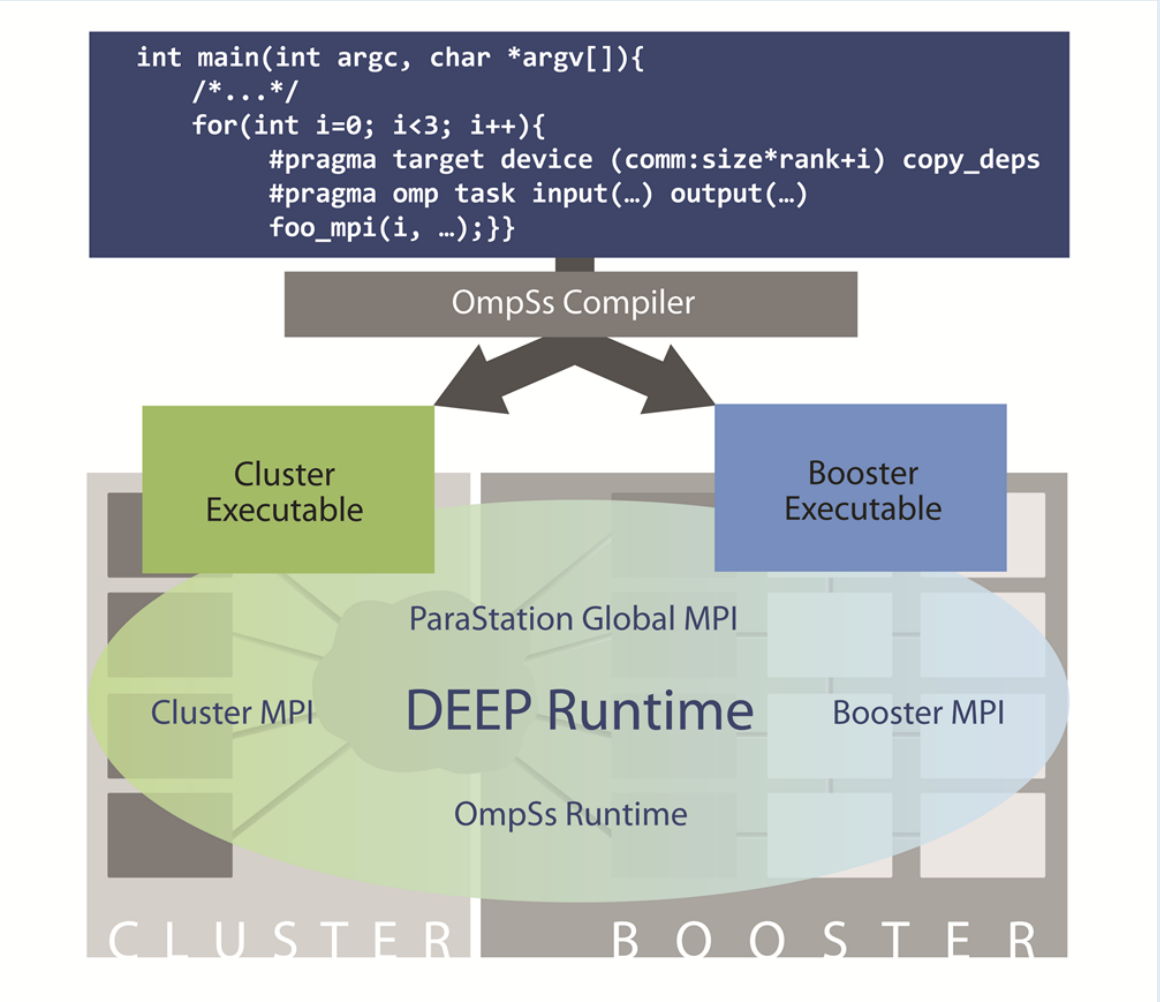
Application Running On DEEP / DEEP-ER

Source code

Compiler

Application binaries

DEEP Runtime



User Benefits

The DEEP/-ER systems offer:

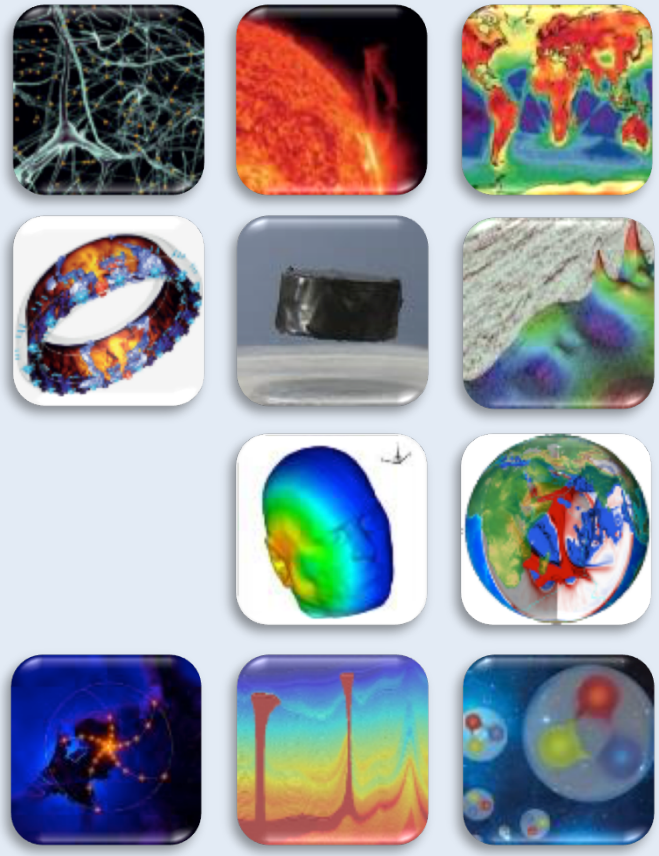
- Complete software stack based on standard components
- Hiding underlying hardware complexity
- A familiar programming environment
- Tools to analyse and optimise application performance
- I/O and resiliency capabilities for data-intensive apps

→ Reducing the burden of the application developer

General Purpose Platform

DEEP + DEEP-ER applications

- Brain simulation (EPFL)
- Space weather simulation (KULeuven)
- Climate simulation (CYI)
- Computational fluid engineering (CERFACS)
- High temperature superconductivity (CINECA)
- Seismic imaging (CGG)
- Human exposure to electromagnetic fields (INRIA)
- Geoscience (BADW-LRZ)
- Radio astronomy (Astron)
- Oil exploration (BSC)
- Lattice QCD (UREG)



Goals

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

Why DEEP/-ER For You

- You want to improve the scalability of your code but a part of it is only low/medium scalable and hinders the rest
- Your application simulates multi-physics, multi-scale phenomena with differentiated scalability characteristics
- You could profit from them but need an efficient (high bandwidth, low-latency) MPI communication between accelerators
- You want to test already a Xeon Phi cluster (for future KNL ones)
- You could profit from Xeon Phi but your code requires large memory at the nodes → NVM on node in DEEP-ER
- You need an efficient parallel I/O infrastructure on an hybrid system
- You worry about how to make your code resilient profiting from new memory technologies

Seismic Imaging

CGG: an integrated geoscience company

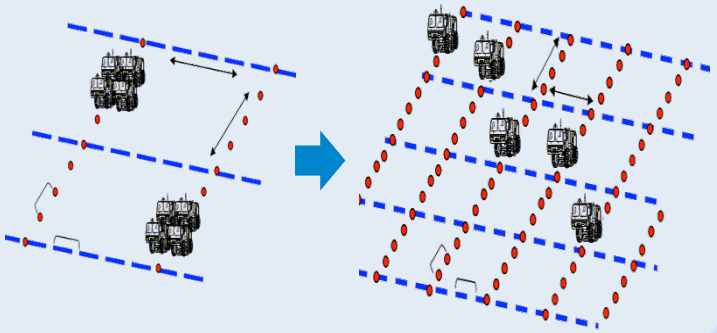
- Equipment
- Acquisition
- Geology, Geophysics, Reservoir

Seismic processing

- Imaging algorithms: RTM and others
- But also many other algorithms: noise removal, multiples removal, velocity model building, …

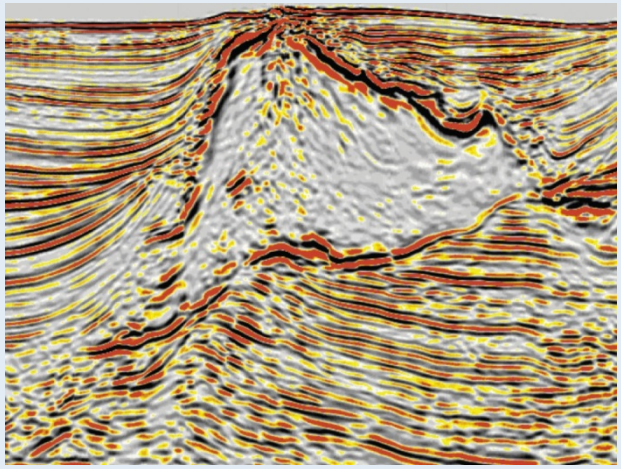
Two factors driving increase of computing resources

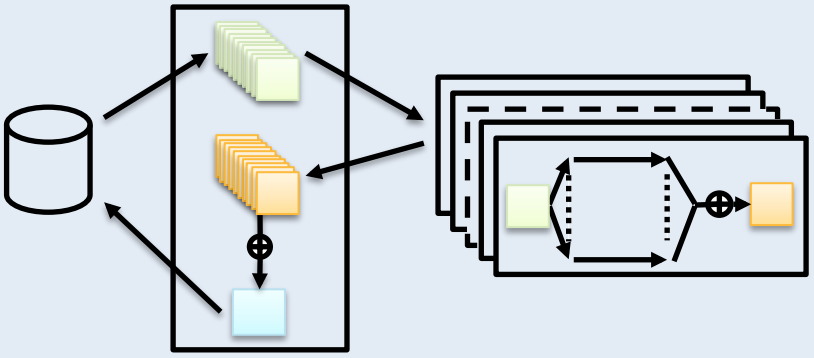
- Better seismic acquisition
  - Denser acquisitions (more sensors per km^2)
  - Increased frequency content
- Better seismic processing algorithms
  - Iterative algorithms (FWI, LS-RTM)
  - Better wave-equation approximations (elastic)



Seismic Imaging Applications

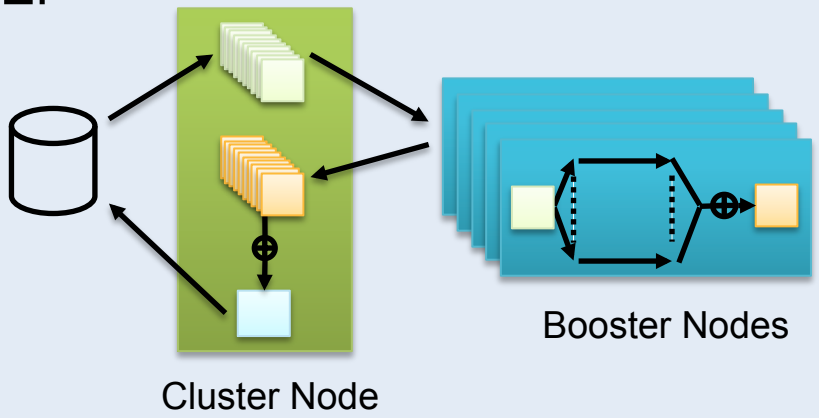
- Two seismic imaging applications: SRMIP & RTM
  - One-way vs two-way wave equation
  - Image each shot independently then stack
- Sketch of implementation
  - Master-worker scheme in MPI
  - Master
    - I/O from/to distributed storage
    - Shot distribution & load balancing
    - Fault tolerance
  - Workers
    - Process shots as efficiently as possible





Mapping SRMIP & RTM on the DEEP architecture

- Master-worker is a natural fit for DEEP
  - MPI Master(s) on Cluster node(s)
  - MPI Workers on Booster nodes
- Efficient worker implementation
  - OpenMP parallel loops
  - Explicit vectorization with pragmas

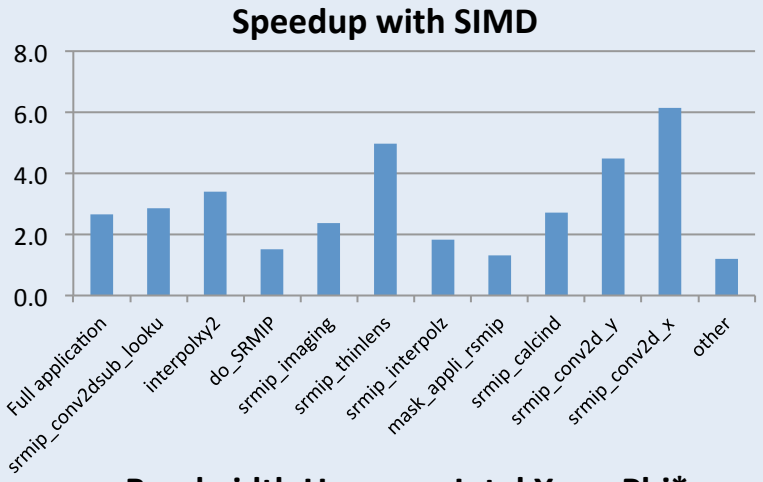


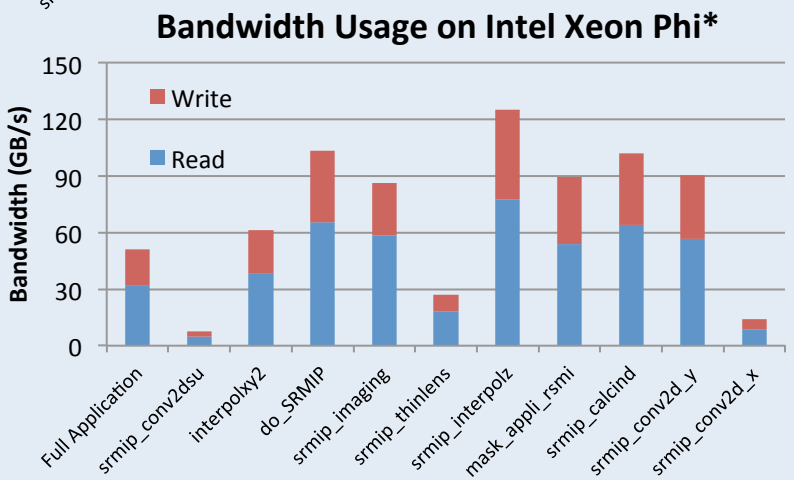
- Different from GPU offload
  - Only kernels are offloaded to the GPU
    vs
  - Complete worker is 'offloaded' on the booster
- Performance evaluation: focus on memory bandwidth
  - Most seismic kernels are limited by memory bandwidth
  - SP flop/byte is between ~.5 and ~5

Porting and Tuning SRMIP

- Porting and optimization on KNC
  - Offload on KNC with MPI
  - Vectorization of kernels
  - Optimized reduction on arrays
  - Optimized number of MPI workers per KNC

- Performance measurements
  - Vectorization: 2.7x speedup for the full application
  - Bandwidth: 25% of the peak bandwidth for the full application
  - Master-Worker scheme: 90% parallel efficiency

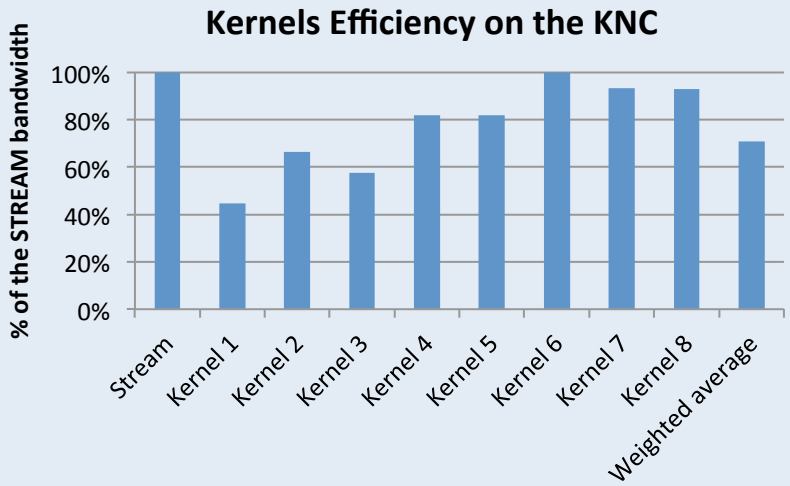


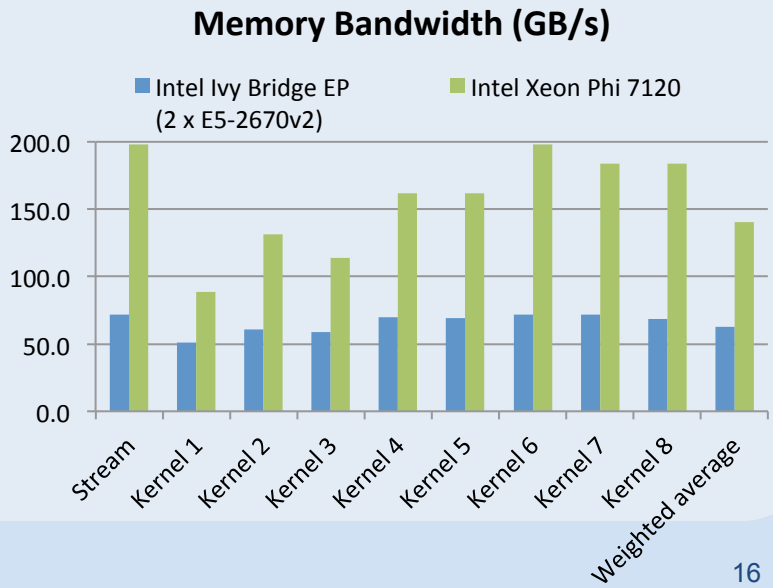


(*) Profile obtained from Xeon Phi Core Events scaled with Xeon Phi Uncore Events following http://software.intel.com/en-us/articles/optimization-and-performance-tuning-for-intel-xeon-phi-coprocessors-part-2-understanding

Porting and Tuning RTM

- Porting the RTM code to the KNC
  - CGG RTM code is written in C + CUDA
  - Porting to the KNC in 5 steps
    - Rewrite CUDA kernels in C + OpenMP
    - Remove GPU allocations and data transfers between CPU and GPU
    - Validate results
    - Optimize kernels for the KNC
    - Validate results
- Focusing on the performance of the modeling part
  - Modeling is the most compute intensive part of RTM (forward and backward wave propagation)
  - RTM has large I/O needs for checkpointing the wavefields, keeping the I/Os would limit performance on the DEEP prototype
  - Full RTM performance can be extrapolated from the modeling performance





Ongoing work

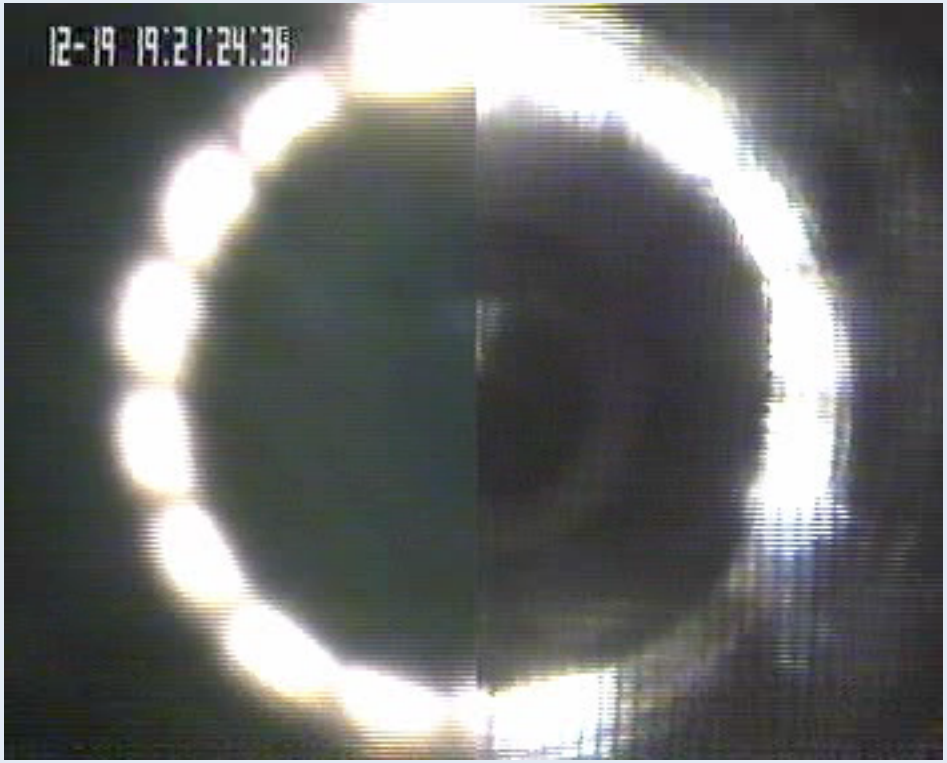
- Benchmarking on the final hardware and comparing performance with other architectures
  - Regular GPU-accelerated cluster
  - Regular Xeon Phi cluster
  - DEEP architecture
- Seismic processing on DEEP
  - MPI Master-Worker is easy to efficiently map to the DEEP architecture thanks to Parastation Global MPI
  - Availability of node-local NVM is a great addition for DEEP-ER
- Domain decomposition for the modelling
  - Growing grid size (longer offsets, higher frequencies)
  - Limited capacity of high bandwidth memory
  - → Will need efficient MPI communications between accelerators

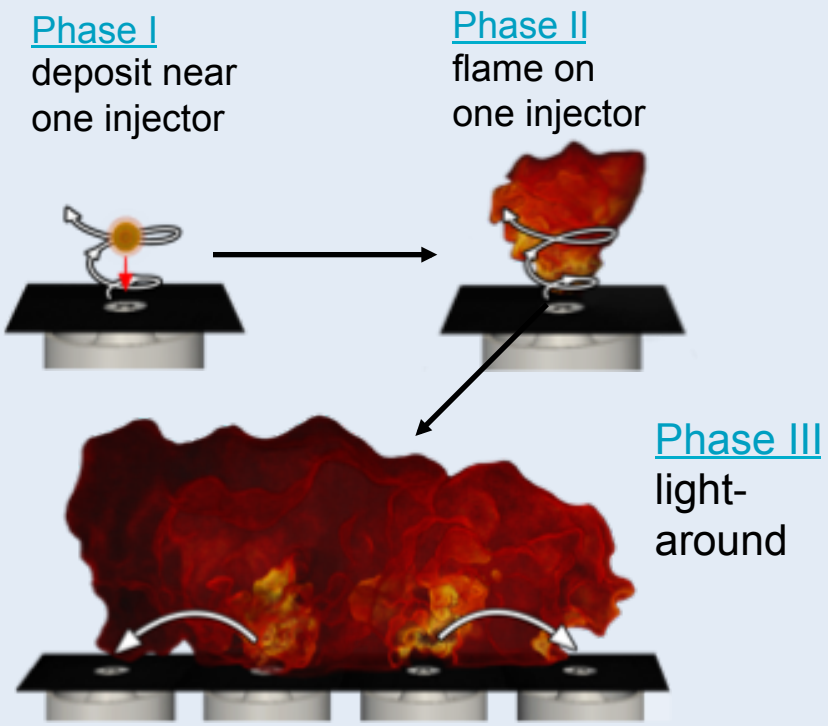
Combustion: An engineering science at the cross-road between *chemistry & fluid mechanics* with strong *technological/ industrial and societal* implications

Current Application & Challenges

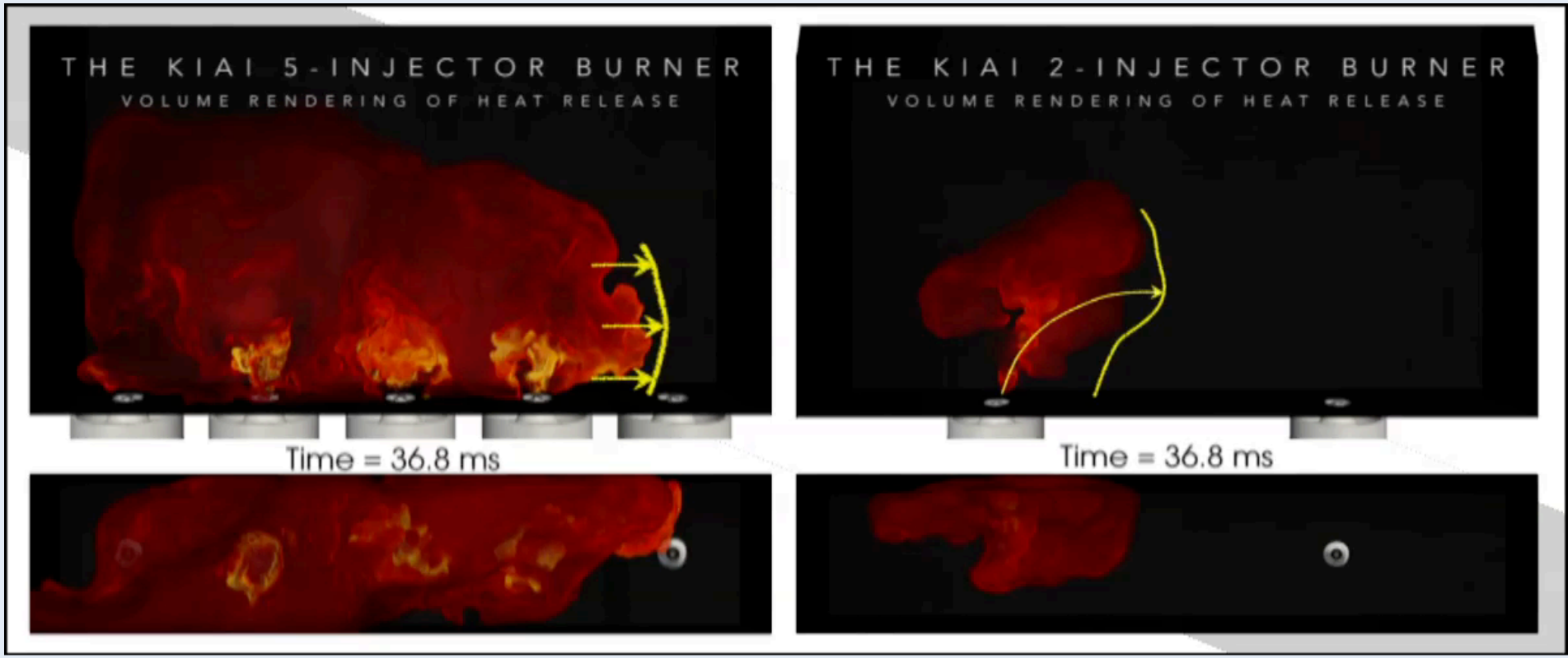
In the aeronautical context: ignition is of paramount importance

- Number of fuel injection systems which calibrate the effective cost and power of the engine
- Operability as well as security issue of the engine





Computational Fluid Dynamics



Computational Fluid Dynamics

- Faster and more affordable simulations using AVBP: “2013 awarded one of the most innovative HPC application for Industry in Europe” by PRACE.

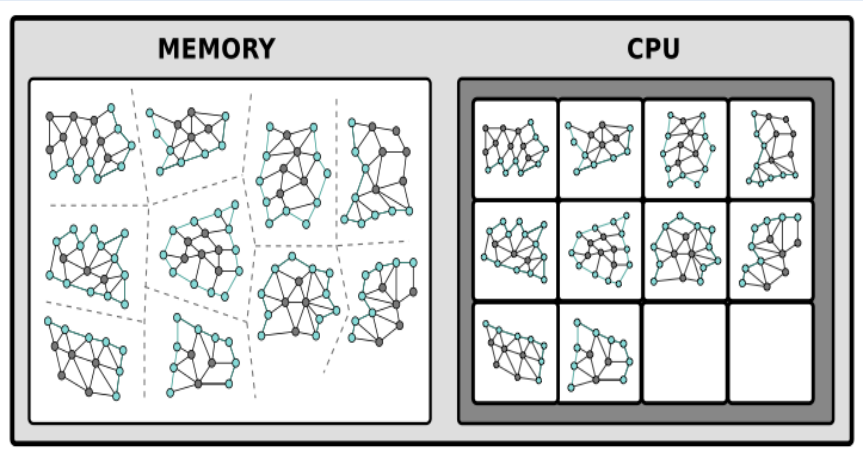
- Increased computational efficiency
  - Better performing architecture

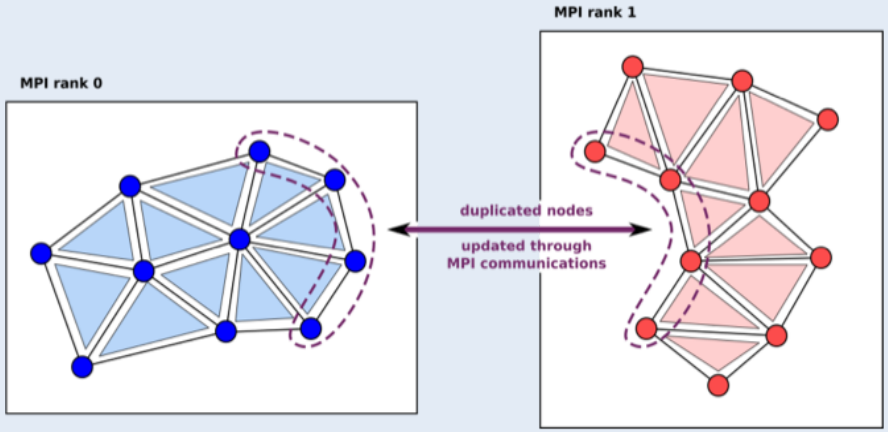
- Increased parallelization
  - More compact/powerful parallel architecture

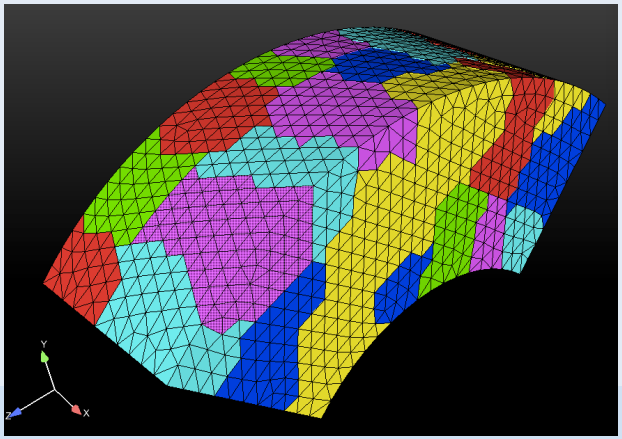
- Decreased computational cost (Flops/$)

Computational Fluid Dynamics

Increased Parallelism required to tackle 224 task chips instead of 16!

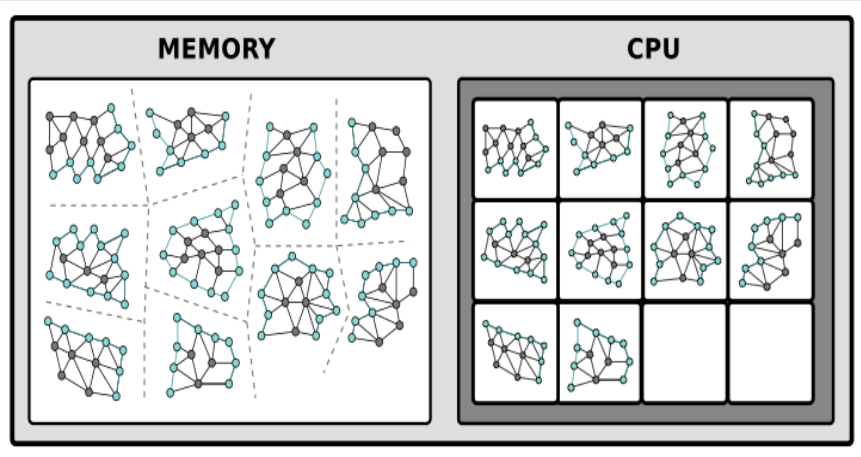


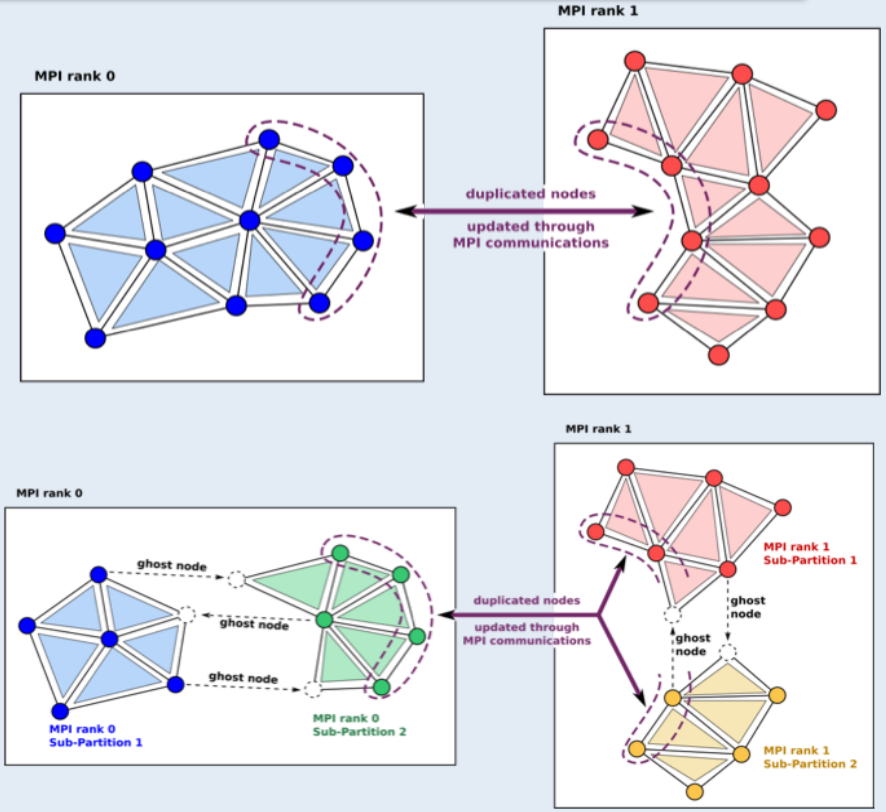




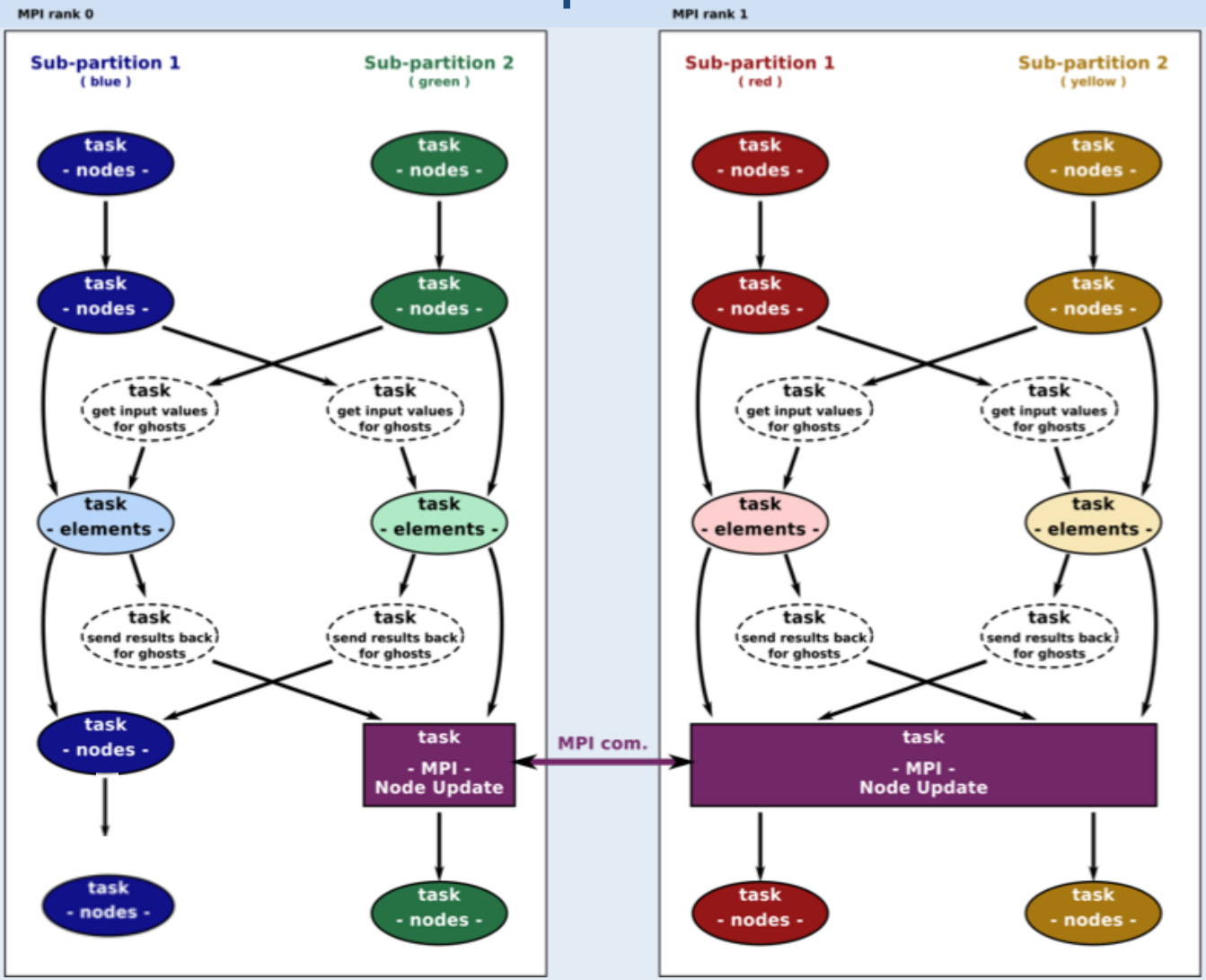
Computational Fluid Dynamics

Increased Parallelism required to tackle 224 task chips instead of 16!

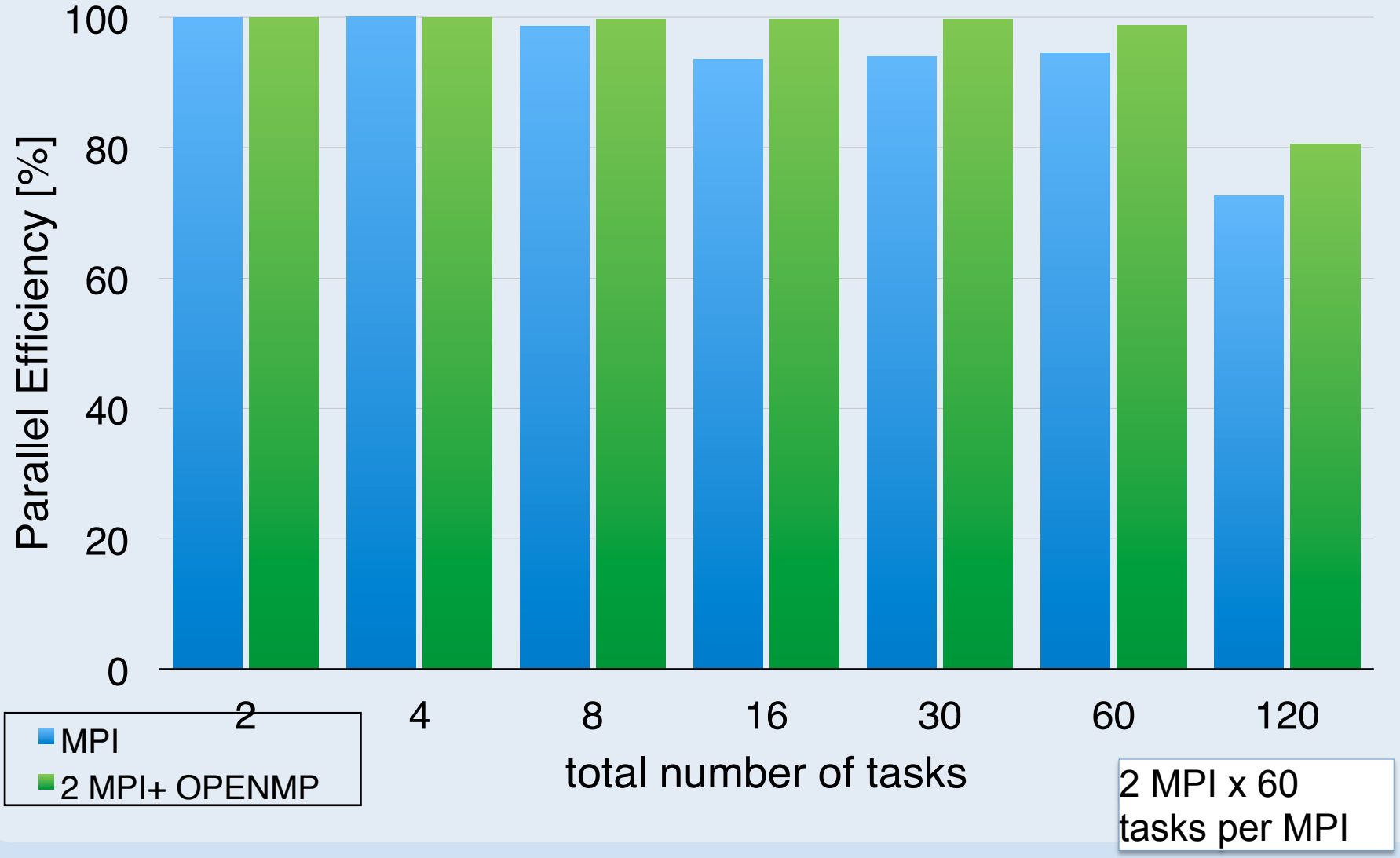




AVBP OpenMP 4 Asynchronous pattern



Parallel efficiency on a single Xeon with 60 cores on a small test case



Computational Fluid Dynamics

The DEEP architecture allows for faster and more scalable applications, a definite step forward on HPC applied for industrial applications.

Partners

Fraunhofer ITWM

CGGVERITAS

CERFACS

CINECA

The Cyprus Institute

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

ASTRON

EUROTECH
Imagine. Build. Succeed.

intel

German Research School for Simulation Sciences

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

ParTec Cluster Competence Center

KU LEUVEN

Mellanox TECHNOLOGIES

Universität Regensburg

SEAGATE

JÜLICH FORSCHUNGSZENTRUM

lrz

Ruprecht-Karls-Universität Heidelberg

Interested?

Contact us!

DEEP
pmt@deep-project.eu

DEEP-ER
pmt@deep-er.eu

LinkedIN
http://linkd.in/1KiBe3y

Twitter
@DEEPprojects

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