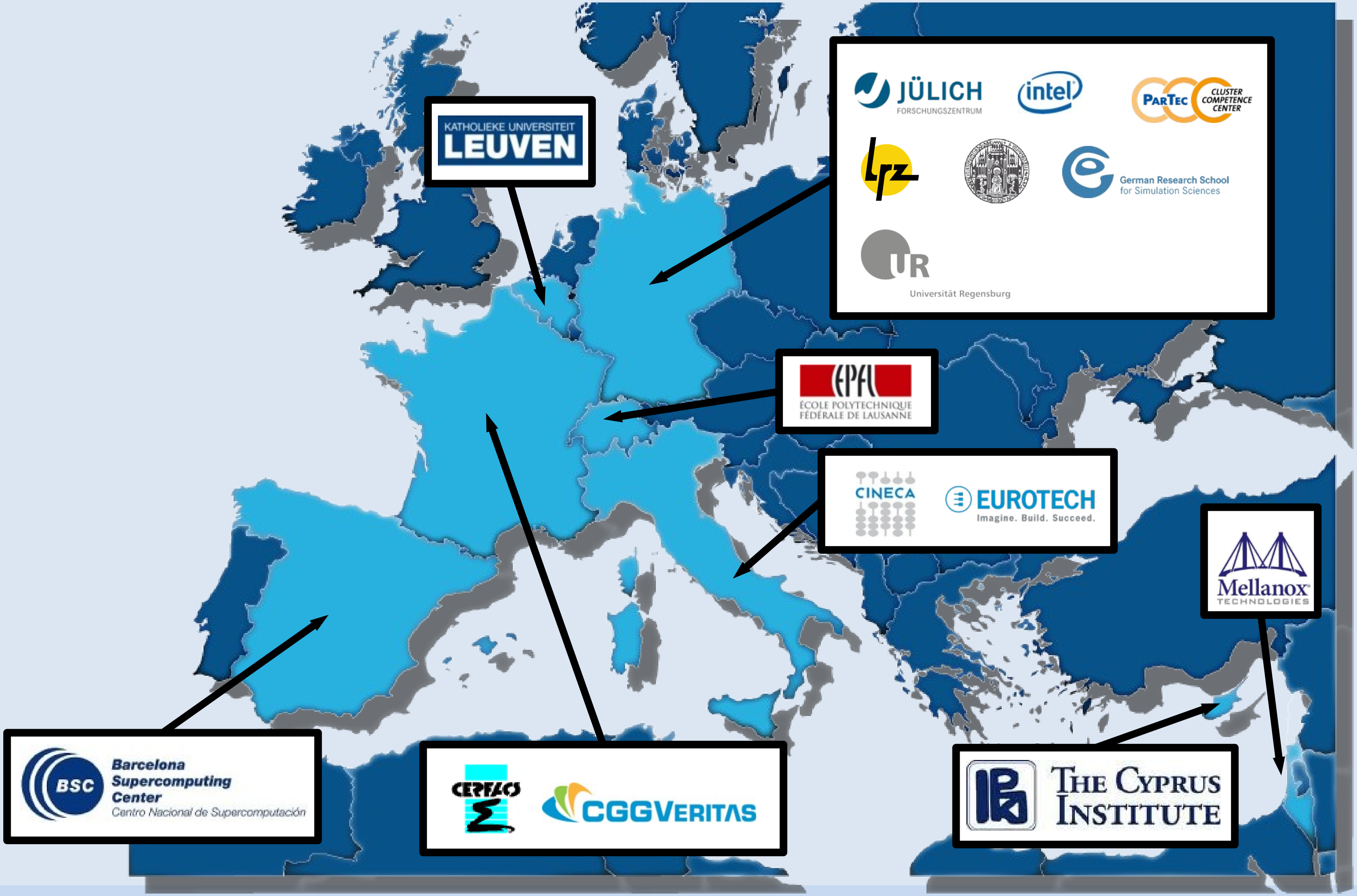
FP7-ICT-2011-7
Integrated Project No. 287530

Dynamical Exascale Entry Platform DEEP

16 partners from 8 countries:
- 3 PRACE Hosting Members
- 5 industry partners

Start: 1st Dec 2011
Duration: 3 years
Budget: 18.5 M€ (8.03 M€ funded by EU)

DEEP Partners


KATHOLIEKE UNIVERSITEIT LEUVEN
JÜLICH FORSCHUNGSZENTRUM
intel
ParTec CLUSTER COMPETENCE CENTER
lrz
German Research School for Simulation Sciences
UR Universität Regensburg
EPFL ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
CINECA
EUROTECH Imagine. Build. Succeed.
Mellanox TECHNOLOGIES
BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación
CERFACS
CGGVERITAS
The Cyprus Institute

Goals

- Design of an architecture leading towards Exascale
- Development of hardware:
  - Implementation of a Booster based on Intel MIC processors and EXTOLL interconnect
- Energy-aware integration of components:
  - Hot-water cooling
- Cluster-Booster Resource-Management System
- Programming environment, programming models
- Libraries and performance analysis tools
- Porting Applications

Application's Scalability

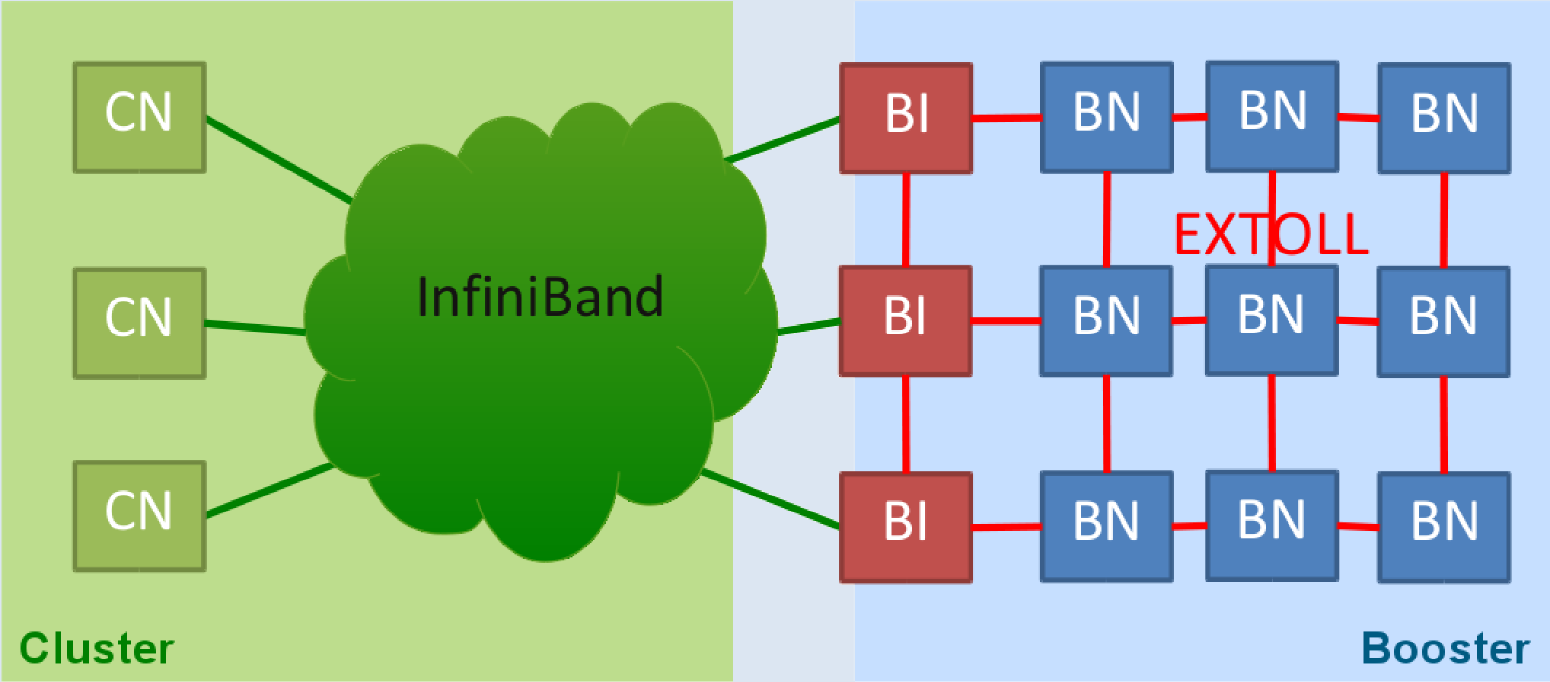
- Only few application capable to scale to O(300k) cores
  - Sparse matrix-vector codes
  - Highly regular communication patterns
  - Well suited for BG/P

- Most applications have more complex kernels
  - Complicated communication patterns
  - Less capable to exploit accelerators
- In fact:
  - Highly scalable apps dominated by highly scalable kernels
  - Less scalable apps dominated by less scalable kernels
    - But there might be highly scalable kernels, too!
    - How to improve their scalability?

Accelerated Cluster vs. Cluster of Accelerators

- Cluster with Accelerators
  - Each node has a classical host CPU
  - Accompanied by one or more Accelerators
  - Communication typically via main memory
  - PCIe bus turns out to be a bottleneck
- Cluster of Accelerators
  - Node consists of Accelerator directly connected to network
  - Impossible with (most) current accelerators
    - Accelerator requires host-CPU to boot
    - Unable to directly talk to the network
    - Accelerator not capable to run general purpose code (OS)

Proposed architecture

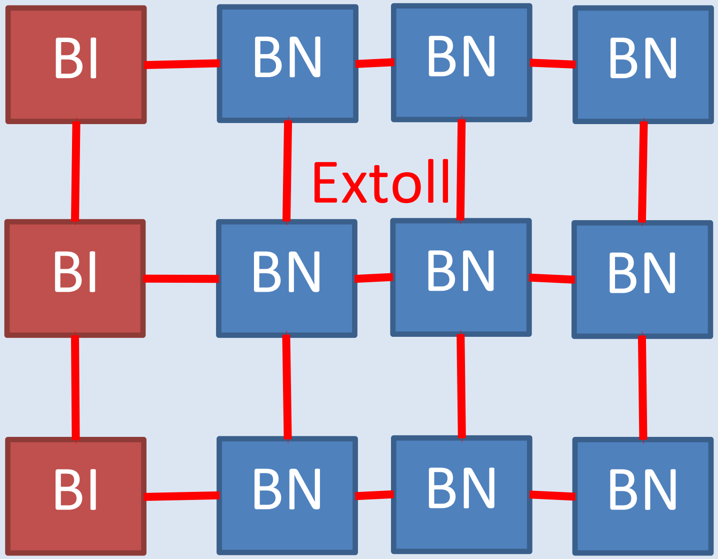


Keep flexibility due to IB between cluster-nodes and booster-nodes

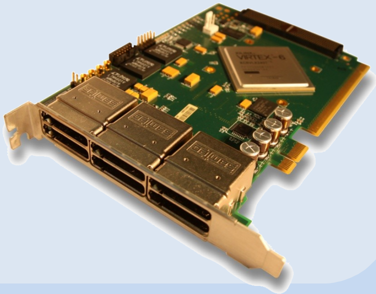
Complex kernels to be offloaded expected to have regular communication patterns

Kernels relieve pressure on CPU to Acc. communication

Extoll

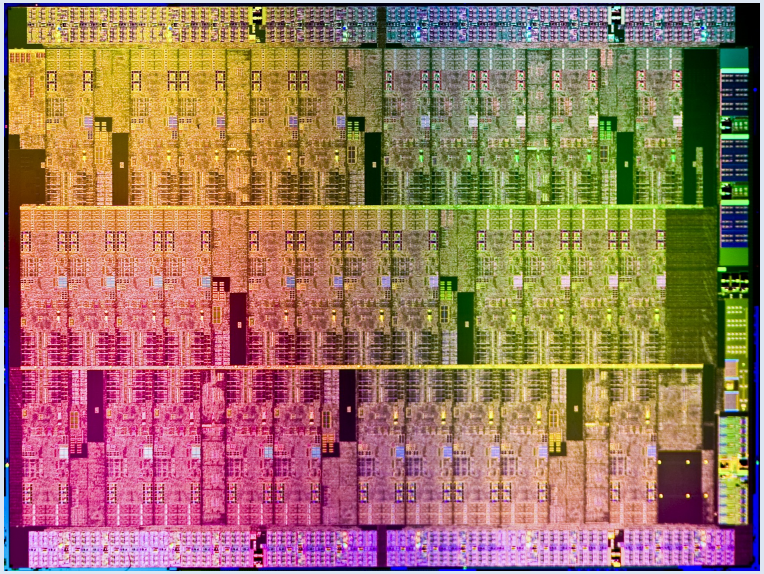


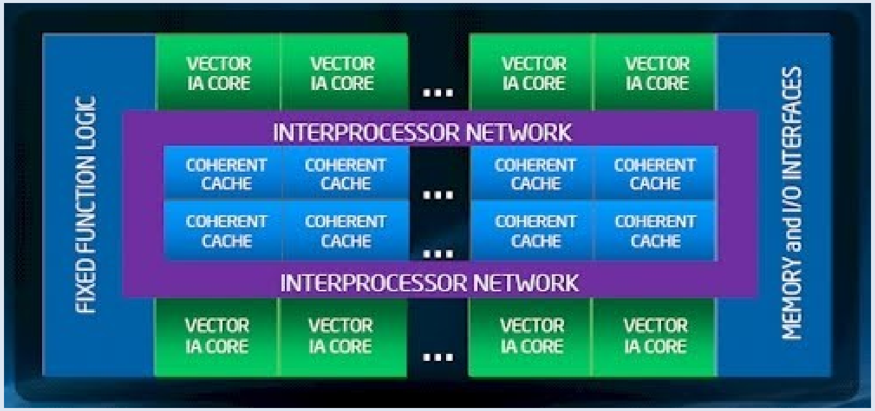
- Ultra low latency (<1 µsec)
- High bandwidth (32 Gbit/s)
- 3D-torus topology preferred
  - any topology possible
  - but very small atomic switches
  - currently just 8 ports
  - will require many layers
- Very scalable (currently limited to 64k due to addresses)
- But less flexible
  - many communication patterns might introduce congestion
  - latency depends heavily on distance
  - max. latency increased with diameter



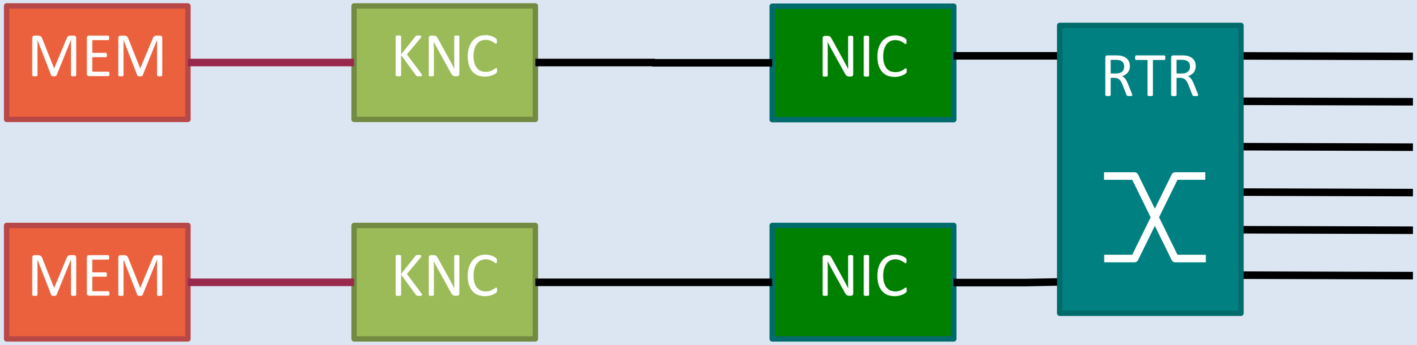
Intel MIC Architecture

- Knights Ferry Processor
  - 32 Core Server Chip
  - GDDR5 memory
  - Based on IA32 cores
  - In order architecture
  - 512 bit wide vector register
- Basically a SoC
- Knights Corner Processor
  - Next generation MIC
  - > 50 Cores
  - 22 nm process





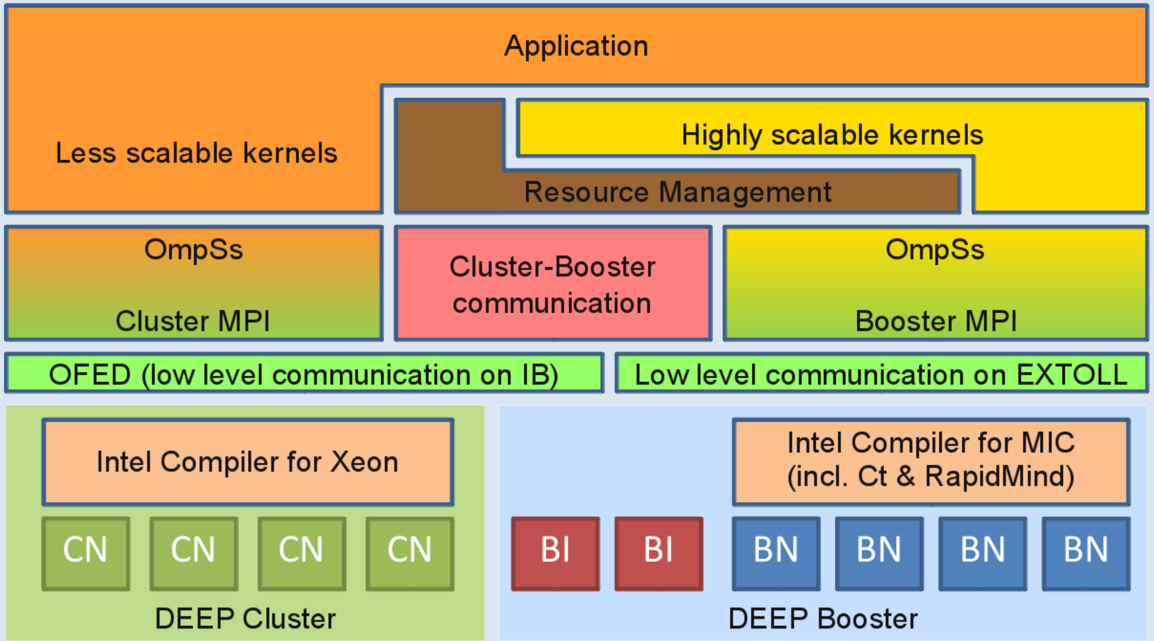
Booster node internal structure



- Booster nodes based on KNC
- Two nodes integrated on one physical PCB
- No extra CPU for bring-up, etc. (relocated to I/O node)
- PCIe root-complex to be integrated into Extoll NIC
- Each KNC shall be able to act autonomously
- Extoll communication to be initiated from within KNC

Software Architecture

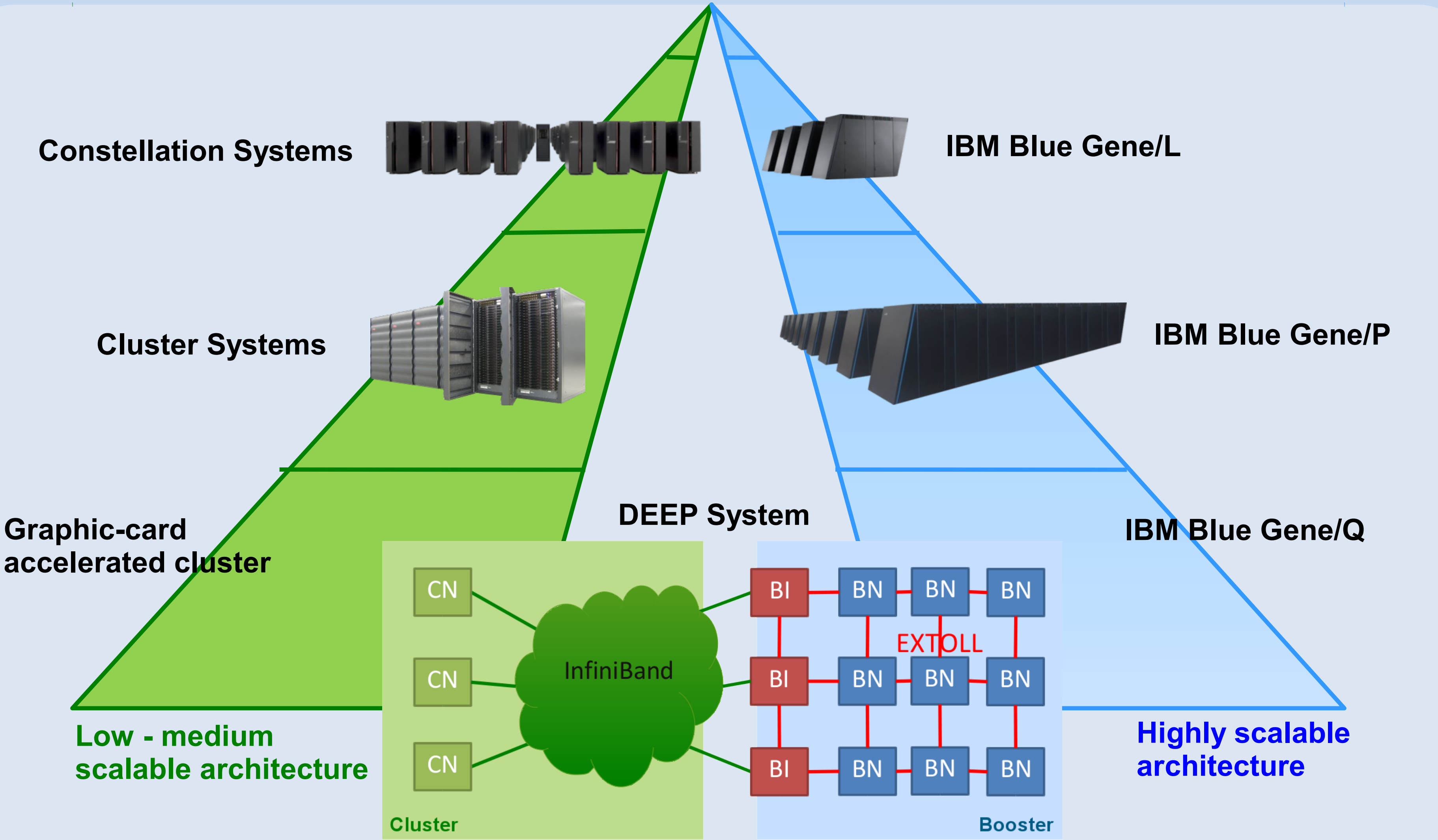
- Basic strategy to port applications:
  - Highly scalable kernels offloaded to the Booster part
  - Less scalable kernels executed on the Cluster part



Pilot Scientific Applications:

- Brain simulation (EPFL)
- Space weather simulation (KULeuven)
- Climate simulation (CYI)
- Computational fluid engineering (CERFACS)
- High temperature superconductivity (CINECA)
- Seismic imaging (CGGVS)

DEEP position



Constellation Systems

Cluster Systems

Graphic-card
accelerated cluster

DEEP System

IBM Blue Gene/L

IBM Blue Gene/P

IBM Blue Gene/Q

Low - medium
scalable architecture

Highly scalable
architecture