The DEEP/-ER architecture: a modular approach to extreme-scale computing

Estela Suarez
Jülich Supercomputing Centre (JSC)
Germany

06.07.2017

The research leading to these results has received funding from the European Community (FP7/2007-2013 and H2020) under Grant Agreements #287530, #610476, and #754304
The DEEP projects
DEEP, DEEP-ER and DEEP-EST

Both combine:
- Hardware
- Software
- Applications
in a strong co-design

DEEP:
Cluster-Booster Architecture + software environment

DEEP-ER:
I/O + resiliency

DEEP-EST:
Modular Supercomputing

EU-Exascale projects
27 partners
Total budget: 44 M
EU-funding: 30 M
Nov 2011 – Jun 2020

DEEP-projects.eu
Homogeneous cluster

- Cluster Nodes: **general purpose** (multi-core) processor technology
  - Same processor characteristics in all nodes
- Single high-speed network connecting them all
- Good concept but limited efficiency for selected HPC applications
“Standard” heterogeneity

Flat topology
Simple management of resources

Static assignment of accelerators to CPUs
Accelerators cannot act autonomously
Flexible assignment of resources (CPUs, accelerators)
Direct communication between accelerators
“Offload” of large and complex parts of applications
DEEP Architecture
DEEP Prototype

- 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)
DEEP-ER Architecture
Innovation

Self-Booting Nodes
On-Node NVM
Network attached memory
Simplified Interconnect

Cluster
Booster
Extoll
**DEEP-ER prototype**

**Cluster**
- 16 dual-socket Intel Xeon E5-2680v3 (Haswell)
- 128 GB DRAM
- 400 GB NVMe
- EXTOLL Tourmalet

**Booster**
- 8 Intel Xeon Phi (KNL) 7210X nodes (16+96GB)
- 400 GB NVMe
- EXTOLL Tourmalet (ASIC)
  - 100 Gb/s per link
- 2x NAM devices
DEEP vs. DEEP-ER
Application performance comparison

GERShWIN (Inria): Single node walltime comparison of the time loop

Setup:
- 20 iterations of the time-loop / No checkpointing
- 1853832 cells
- 1 node with best process-threads combination

Walltime [s]

2.9x 1.6x 1.7x 2.3x 1.3x 1.7x 2.8x 1.8x

Lagrange order

PNL (SDV)  KNC (miclogin)  Haswell (SDV)  Sandy Bridge (DEEP Cluster)
SOFTWARE ENVIRONMENT
Software environment

- **Scheduler**: Torque/Maui → future moving to SLURM
- **Filesystem**: BeeGFS
- **Compilers**: Intel, gcc, PGI
- **Debuggers**: Intel Inspector (threading, memory), TotalView (source code, memory debugger)
- **Programming**: ParaStation MPI (mpivich), OpenMP, OmpSs
- **Performance analysis tools**: Extrae/Paraver, Scalasca, Intel Advisor, Intel, VTune
- **Libraries**: SIONlib, SCR, E10, HDF5, netcdf, PETSc …
OmpSs on top of MPI provides pragmas to ease the offload process
Application running on DEEP

Source code

Compiler

Application binaries

DEEP Runtime

```
int main(int argc, char *argv[]){
    /*...*/
    for(int i=0; i<3; i++){
        #pragma omp task in(...) out (...) onto (com, size*rank+1)
        foo_mpi(i, ...);
    }
}
```
DEEP-ER I/O and resiliency

- **I/O Software architecture**
  - DEEP-ER applications
    - E-10
    - BeeGFS + Extensions
    - SIONlib
  - Disks
  - NVMe
  - NAM

- **Resiliency SW architecture**
  - SCR (checkpointing handling)
  - ParaStation MPI (process CP)
  - OmpSs (task checkpointing)

Combination of SW packages provides new functionality and exploits HW
APPLICATIONS
Application-driven approach

**DEEP projects applications (15):**

- Brain simulation (EPFL + NMBU)
- Space weather simulation (KULeuven)
- 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)
• **Full user flexibility** – many different use modes
  – Dynamic ratio of processors/coprocessors
  – Use Booster as pool of accelerators (globally shared)
  – Discrete use of the Booster
  – Discrete use + I/O offload
  – Specialized symmetric mode

• **More efficient use of system resources**
  – Only resources really needed are blocked by applications
  – Dynamic allocation further increases system utilization

• Better I/O performance and resiliency
DEEP-EST and JURECA

MODULAR SUPERCOMPUTING ARCHITECTURE
Generalization of the Cluster-Booster concept

Module 1: Storage
- Disk

Module 2: Cluster
- CN
- CN

Module 3: Many core Booster
- BN
- BN
- BN

Module 4: Memory Booster
- NAM

Module 5: Data Analytics
- DN

Module 6: Graphics Booster
- GN

Modular Supercomputing
Modular Supercomputing
Modular Supercomputing
DEEP-EST prototype
Going production
The DEEP projects bring a new view to heterogeneity
- Modular Supercomputing architecture
- Software environment fully supporting system design
- Programming environment based on standards
- Hardware, software and applications jointly developed
- Strongly co-design driven
- Cluster + Booster going in production: JURECA system

Next step: DEEP-EST
- Three modules
- Address HPDA + HPC

Want to try out? →

www.deep-projects.eu
@DEEPProjects
pmt@deep-projects.eu