Mitglied der Helmholtz-Gemeinschaft

JÜLICH
FORSCHUNGSZENTRUM

Impacts of Current Hardware and Software Developments on Simulation Sciences

N. Attig

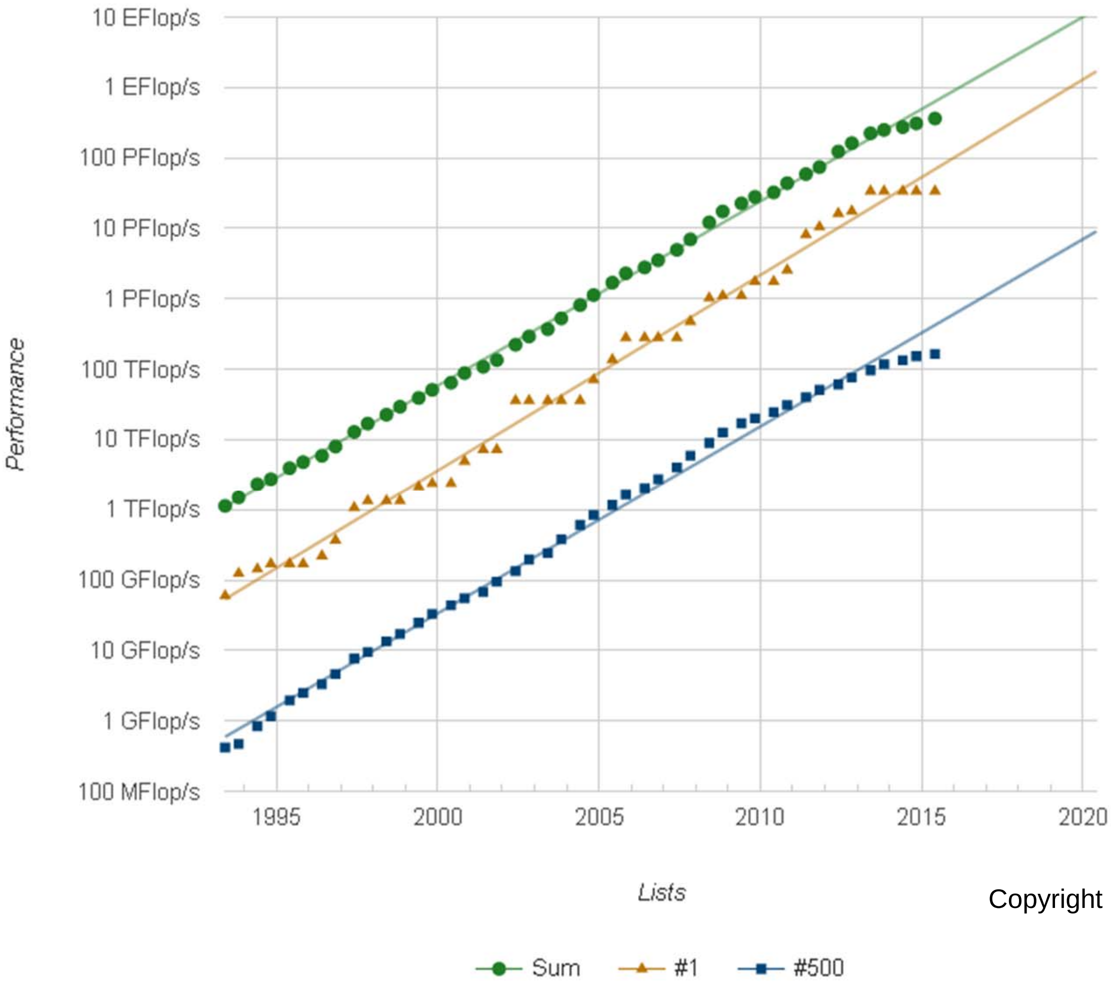
Jülich Supercomputing Centre (JSC), Forschungszentrum Jülich

Motivation

Introduce current and prominent trends in computer science

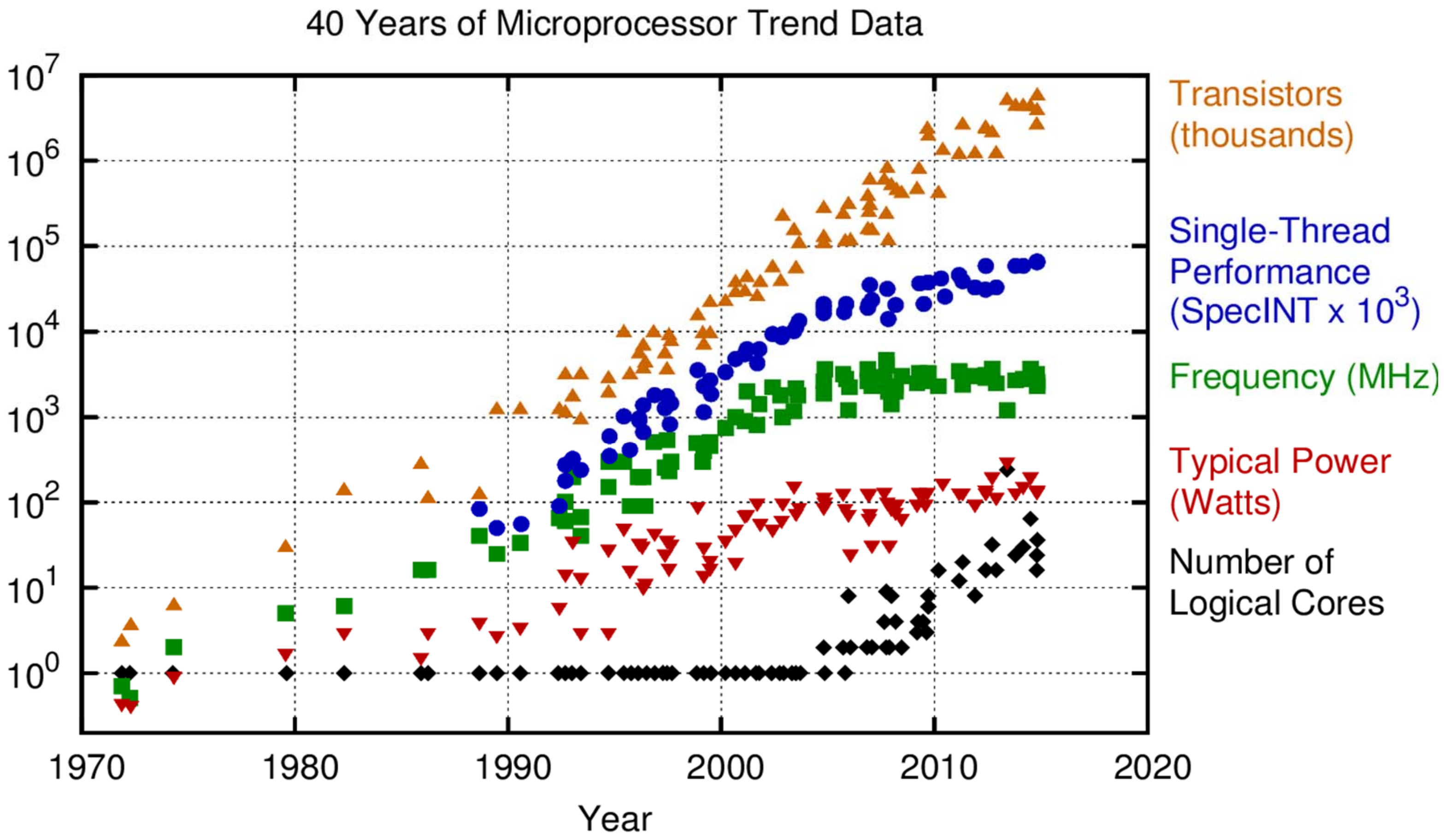
and discuss how an HPC/data centre can guide and support computational scientists to improve their simulation codes on current and future computer systems.

Performance Development



Copyright 1993-2015 TOP500.org

Processor Developments



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2015 by K. Rupp

Technology Trend: More Parallel Processors

Processor Parallelism

- Micro-architecture level:
  - *Data-parallel instructions (SIMD)*
  - *Number of instruction pipelines*
- Processor level: multi-core

Example: JUROPA/JURECA Clusters at JSC

	JUROPA [2009]	JURECA [2015]
SIMD width	2x64 bit	4x64 bit
No. of SIMD pipelines	1	2
Core/processor	4	12
Flop/cycle/processor	16	192
Core clock frequency [GHz]	2,93	2,5

Even More Parallel "Accelerators" …

Competing Technologies

- Graphics processing units (GPU)
- Xeon Phi

Processor level parallelism

	NVIDIA K40	Intel Xeon Phi 7120D
Flop/cycle/processor	1920	976
Core clock frequency [GHz]	0.75	1.24

Technology Trend: Deeper Memory Hierarchy

High memory capability and capacity requirements

- Increasing compute performance
  → Increase of memory bandwidth B_{mem}
- Applications ambition to solve large problems
  → Significant memory capacity C_{mem}

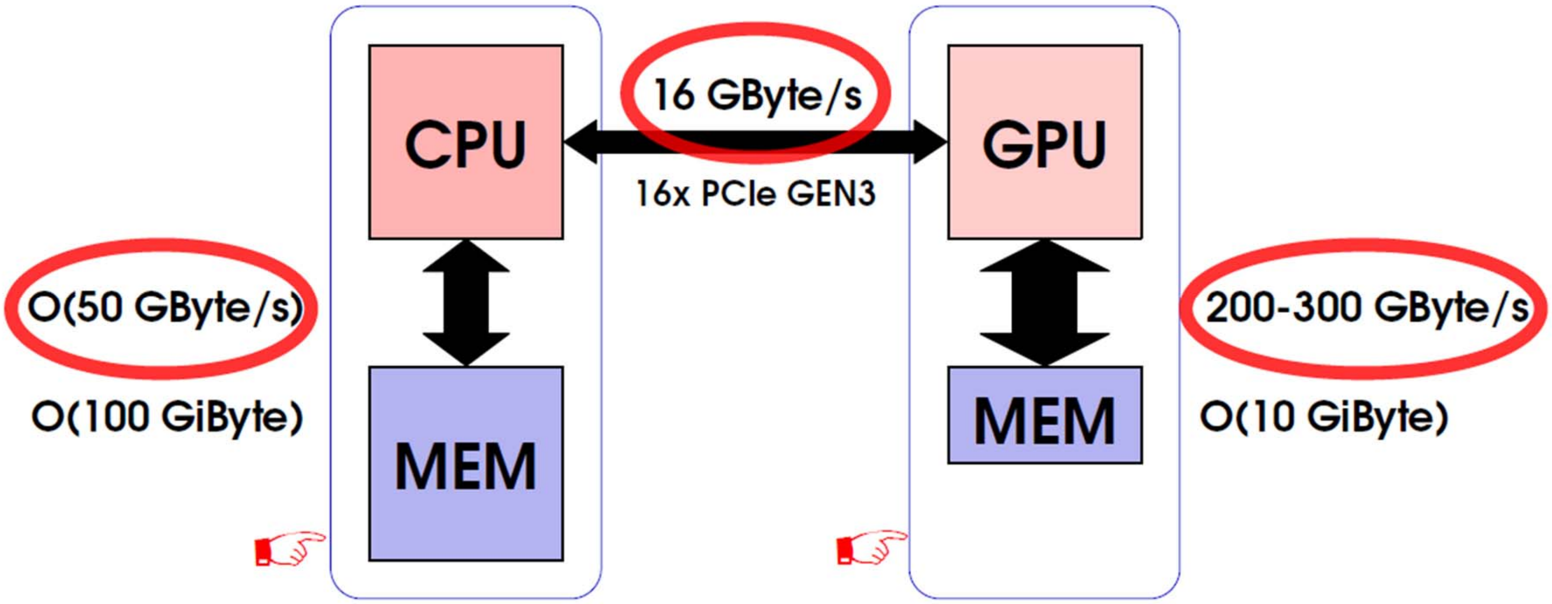
Costs challenge

- Faster memory = more expensive (larger GByte/EUR)

Solution: Memory hierarchy with more levels

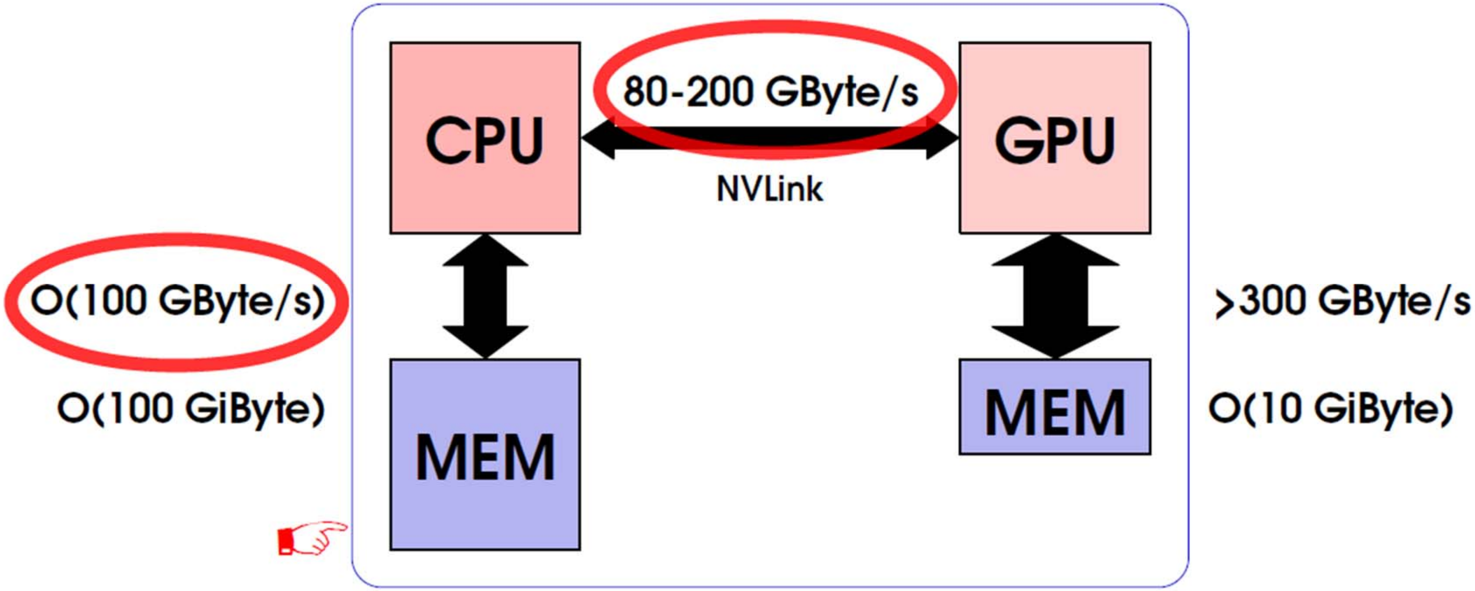
- Fast memory, smaller capacity
- Large capacity, slower memory

Accelerators Architectures Today



- Relatively small bandwidth between host and device
- Separate memory coherence domains

Future GPU Architectures



- Similar bandwidth host-device and host-memory
- Single memory coherence domains
- OpenPOWER is going down this road

Technology Trend: Energy Efficiency

Challenge: Common understanding of upper limit for a 1 Exaflop system: 20 MW

- Current #1 system (Tianhe-2): 55 Pflops, 17,8 MW
  → energy consumption has to be reduced by a factor of 20!!
- Latest optimistic estimates: 1 Eflop in 2020 needs 180-425 MW
  Peter M. Kogge, Lecture Notes in Computer Science **9137** (2015) 323-339

Options

- Develop energy-efficient processors → accelerators and other components
- Processor voltage optimization
- Optimize data centre cooling: avoid fans, free cooling

Technology Trend: Innovative Interconnects

	EXTOLL	Intel True Scale		Mellanox IBAN		PLX Technology
	Tourmalet	QDR	QDR 80	EDR	FDR	ExpressFabric®
Availability	Q3/2015	Now	Now	2015	Now	2015
Switches	None	IBAN	IBAN	IBAN	IBAN	PCIe switches req.
Topologies	≤7 direct connections	Switched, any, 1 rail	Switched, any 2 rails	Switched, any, 1-2 rails	Switched, any, 1-2 rails	Switched, any, 1 rail only
# Links per NIC	7	1 or 2	1 or 2	1 or 2	1 or 2	1-4 (for DEEP-ER)
Link BW	120 Gbit/s	40 Gbit/s	80 Gbit/s	103 Gbit/s	56 Gbit/s	32 (4 links) –128 (1 link) Gbit/s
Aggregate BW	940 Gbit/s	80 Gbit/s	160 Gbit/s	206 Gbit/s	112 Gbit/s	128 Gbit/s
# contexts	256	64	2*64			64
SR-IOV support	No	No	No	No	Yes	Yes
Drivers & Firmware	Adaptable	Available	Available	N/A	Available, KNL?	OSS
Driver I/F	VELO, SMFU, OFED	OFED, PSM	ODEF, PSM	OFED	OFED	OFED

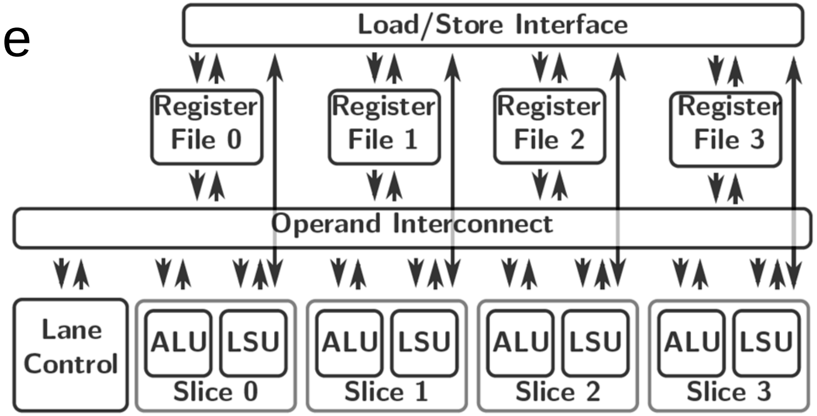
Technology Trend: Data Avoiding Architectures

Processing in Memory

Data are processed in memory without transferring them between memory and CPU

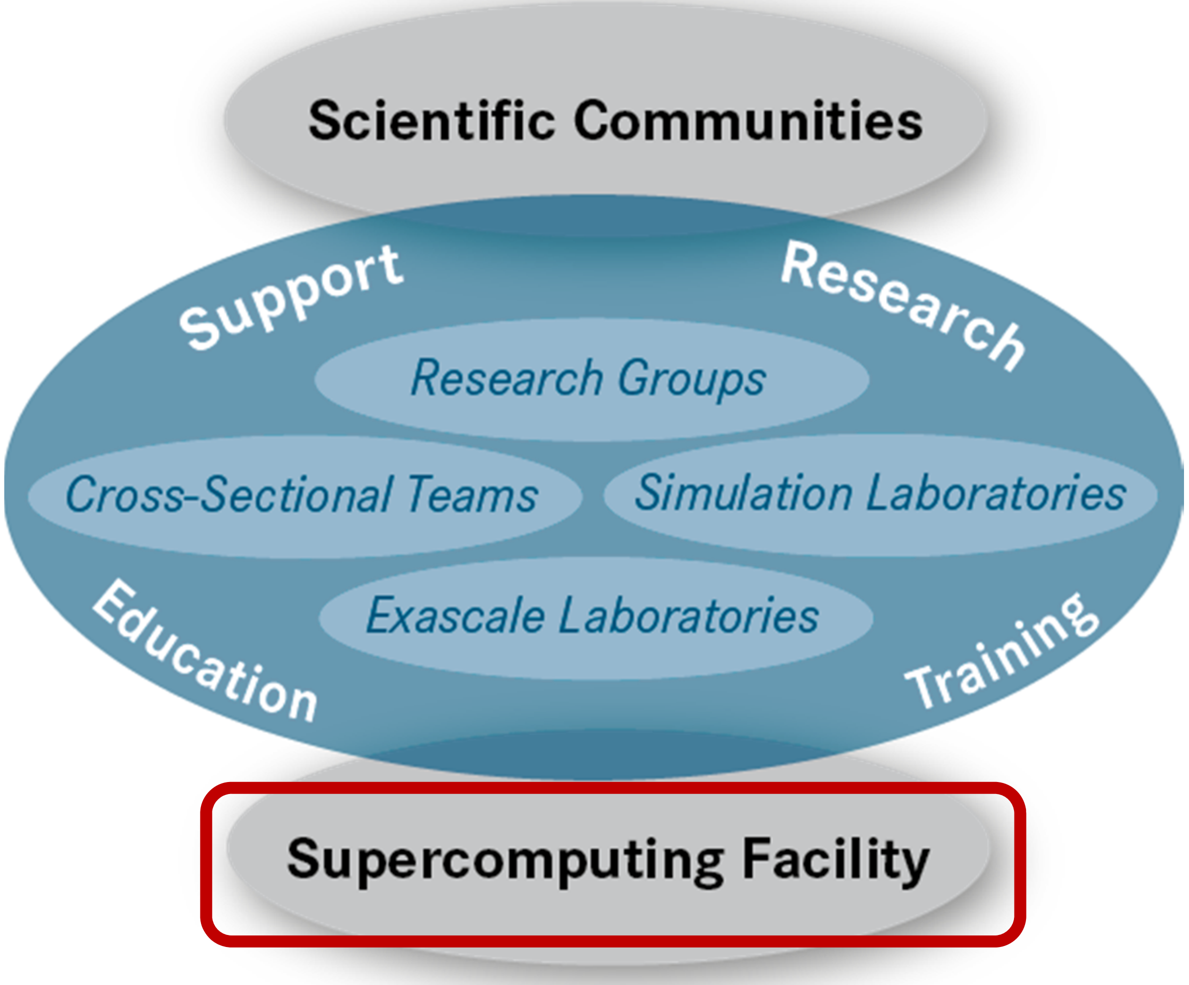
Example: Active Memory Cube (IBM)

- Processing-in-memory device based on Hybrid Memory Cube
- 3-dim stacked memory with Through-Silicon-Vias
- Logic layer with 32 compute elements
- Chainable for capacity and compute performance



Paul F. Baumeister, Hans Boettiger, José R. Brunheroto, Thorsten Hater, Thilo Maurer, Andrea Nobile, Dirk Pleiter, Lecture Notes in Computer Science **9137** (2015) 96-112

R&D and Application Support at the Jülich Supercomputing Centre



Scientific Communities

Support

Research

Research Groups

Cross-Sectional Teams

Simulation Laboratories

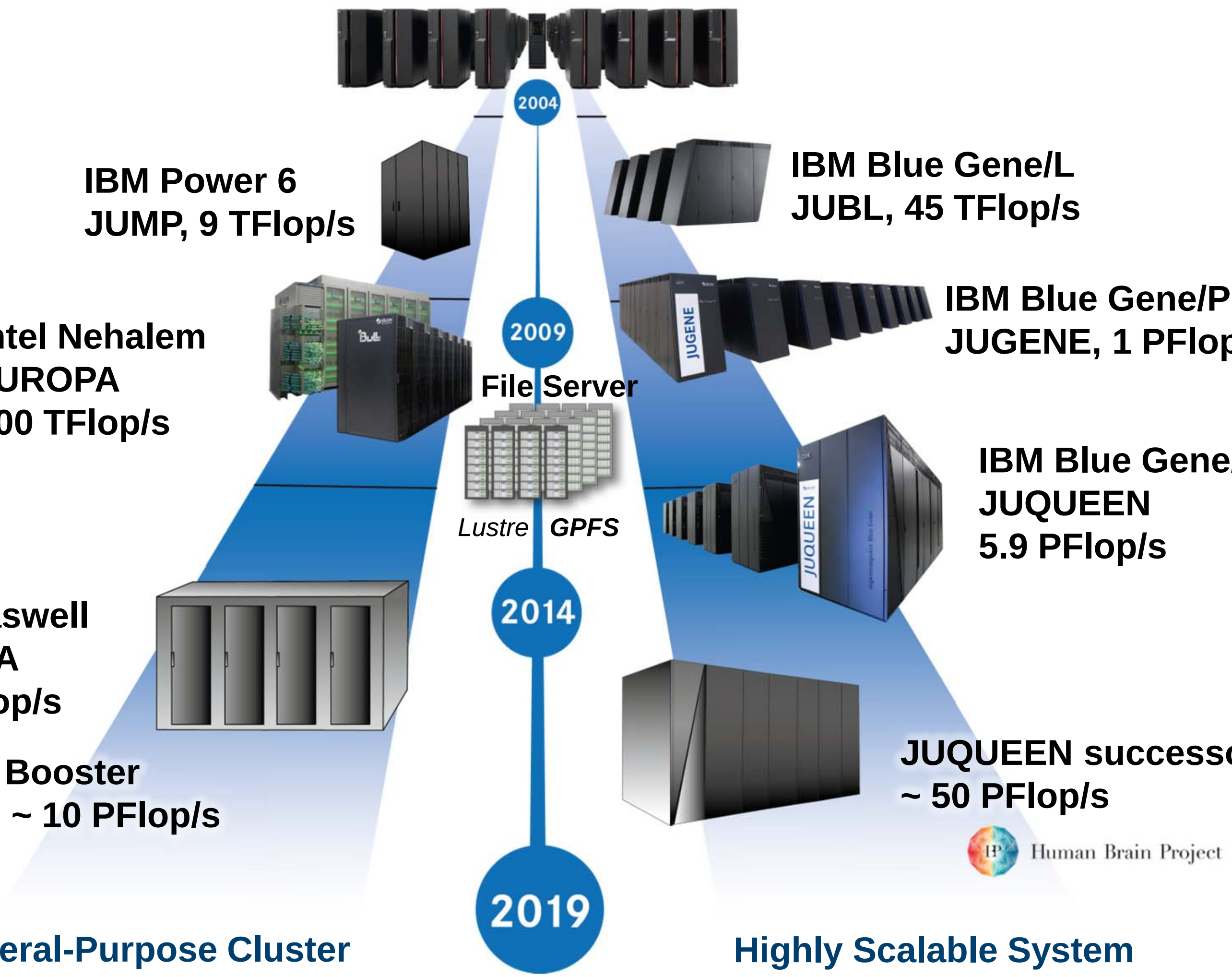
Exascale Laboratories

Education

Training

Supercomputing Facility

HPC Systems @ JSC: Dual Architecture Strategy



IBM Power 4+ JUMP, 9 TFlop/s

IBM Power 6
JUMP, 9 TFlop/s

IBM Blue Gene/L
JUBL, 45 TFlop/s

Intel Nehalem
JUROPA
300 TFlop/s

IBM Blue Gene/P
JUGENE, 1 PFlop/s

File Server

Lustre GPFS

IBM Blue Gene/Q
JUQUEEN
5.9 PFlop/s

Intel Haswell
JURECA
~ 2 PFlop/s

+ Booster
~ 10 PFlop/s

JUQUEEN successor
~ 50 PFlop/s

Human Brain Project

General-Purpose Cluster

Highly Scalable System

JURECA: Jülich Research on Exascale Cluster Architectures

System integrator:
T-Platforms, Russia

JURECA, an Intel-based cluster

- 2 Intel Haswell 12-core processors,
  2.5 GHz, SMT, 128 GB main memory
- 1,884 compute nodes or 45,216 cores, thereof
  75 nodes with 2 K80 NVIDIA graphics cards each and
  12 nodes with 512 GB main memory and 2 K40 NVIDIA graphics cards each for visualisation
- 2.245 Petaflop/s peak (including K80 graphics cards)
  ?? Petaflop/s Linpack
- 281 TByte memory
- Mellanox Infiniband EDR
- Connected to the GPFS file system on JUST

JUQUEEN: Jülich's Scalable Petaflop System

IBM Blue Gene/Q JUQUEEN

- IBM PowerPC® A2 1.6 GHz, 16 cores per node
- 28 racks, 458,752 cores
- 5,9 Petaflop/s peak
  5,0 Petaflop/s Linpack
- 448 TByte main memory
- connected to a Global Parallel File System (GPFS) with O(10) PByte online disk and O(50) PByte offline tape capacity
- 5D network
- Production start: Nov 5, 2012

Jun 2015:
#2 in Europe
#9 worldwide
#51 in Green500

Community-specific Systems and Services at JSC

Astrophysics, Neuroscience and Biomedicine

- JUDGE: Intel-based Linux cluster + NVIDIA GPUs
  206 nodes + GPUs, 240 Tflop/s, 20 TB disk
- Used internally by disciplinary researchers
  and members of DFG special research fields

ILDG and LOFAR

- dCache storage systems
- 240 TB disk + 3 PB tape capacity

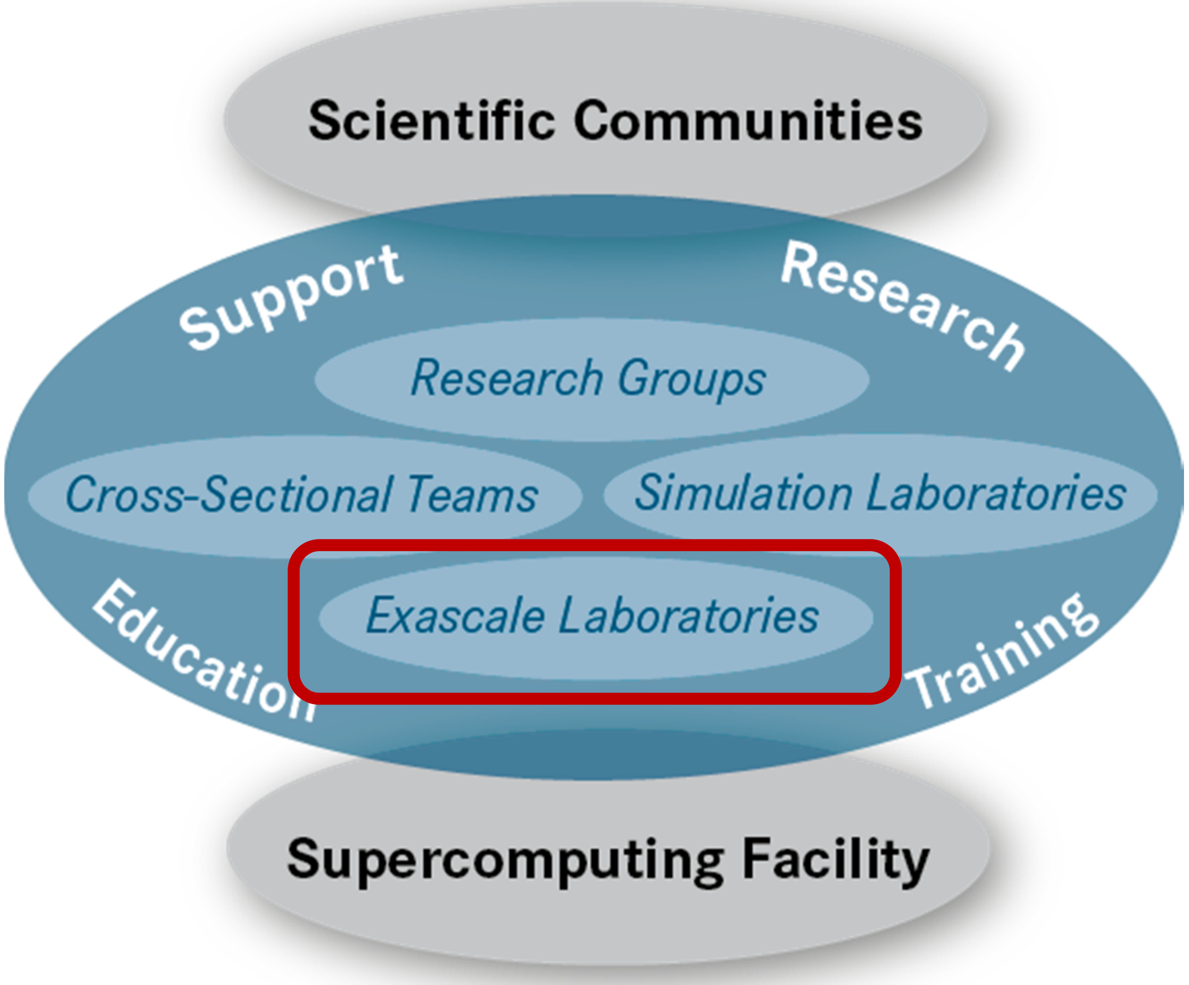
AMS - Cosmic-ray research on the ISS

- Compute and data facilities to support
  the data analysis of the AMS partner RWTH Aachen
- 140 Intel Haswell processors with 14 cores each,
  about 3 PB disks on GPFS@JUST

Prototype Systems at JSC

- QPACE (QCD Parallel Computing on the Cell)
  1,024 Power XCell 8i processors, 100 Teraflop/s, 4 TB memory
  - Innovative “cold plate cooling”; node card cooled by conduction
  - #1 in Green500 2009/2010
- **JUDGE** (Jülich Dedicated GPU Environment)
  206 nodes with 2 Intel Westmere 6-core 2.66 GHz processors each, 412 graphic processors (NVIDIA Fermi), 240 Teraflop/s, 20 TB memory
- DEEP (Dynamical Exascale Entry Platform)
  Cluster: 128 nodes with 2 Intel Sandy Bridge 8-core 2.7 GHz procs each,
  Booster: 384 Intel MICs (KNC) connected via Extoll interconnect
- **BGAS** (Blue Gene Active Storage)
  attached to JUQUEEN and an external storage system
  - Boosts I/O performance, facilitates interactive access to the data

R&D and Application Support at the Jülich Supercomputing Centre



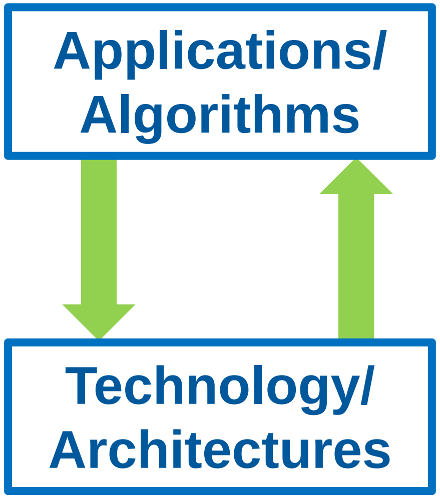
Exascale Research at JSC

Exascale challenges

- Drastically improve energy efficiency
- Preserve usability at tremendously increased level of parallelism
- Keep overall system balanced
- Address reliability and resilience

Co-design approach

- Scientific problem requirements influence architecture design and technology
- Architectural constraints impact formulation and design of algorithms and software



Co-design enabled through Exascale Labs

Exascale Research at JSC (cont.)

Established Exascale Labs

- Exascale Innovation Center (EIC) with IBM [2010]
- ExaCluster Lab (ECL) with Intel and ParTeC [2010]
- NVIDIA Application Lab [2012]
- Power Acceleration and Design Center (PADC) [2015] with IBM (Böblingen and Zürich) and NVIDIA

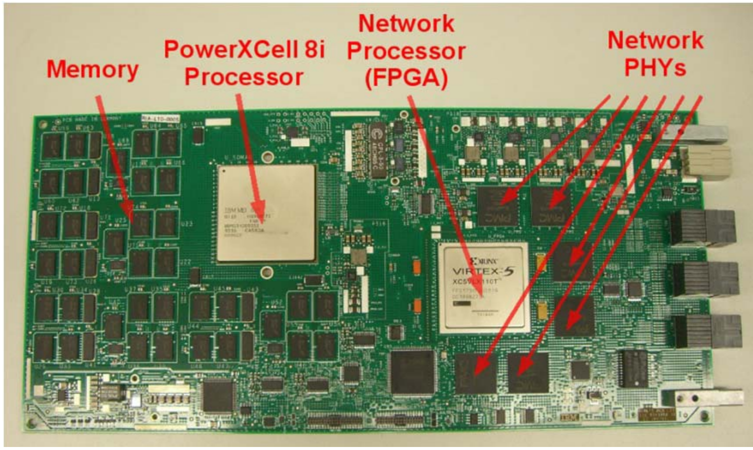
Topics addressed

- New architectural concept exploration
  - *Booster concept*
  - *Active storage architectures*
- Efficient and productive use of many-core architectures
- Richer memory hierarchies
- Scalability through new network technologies

QPACE (2009-today)

QCD Parallel Computing on the Cell

- Massively parallel architecture optimized for LQCD applications
- Developed by an academic-industrial team
  - Academic team: U Regensburg, U Wuppertal, U Ferrara/Milano, FZJ, DESY
  - Industrial partner: IBM
- Concept
  - Fast commodity processor = IBM PowerXCell 8i
  - Custom network → custom network processor
  - Custom system design

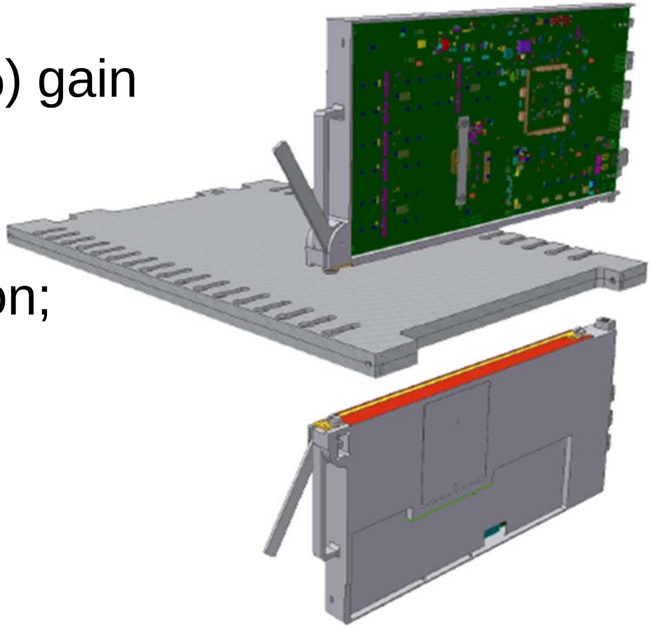


QPACE (2009-today)

QCD Parallel Computing on the Cell

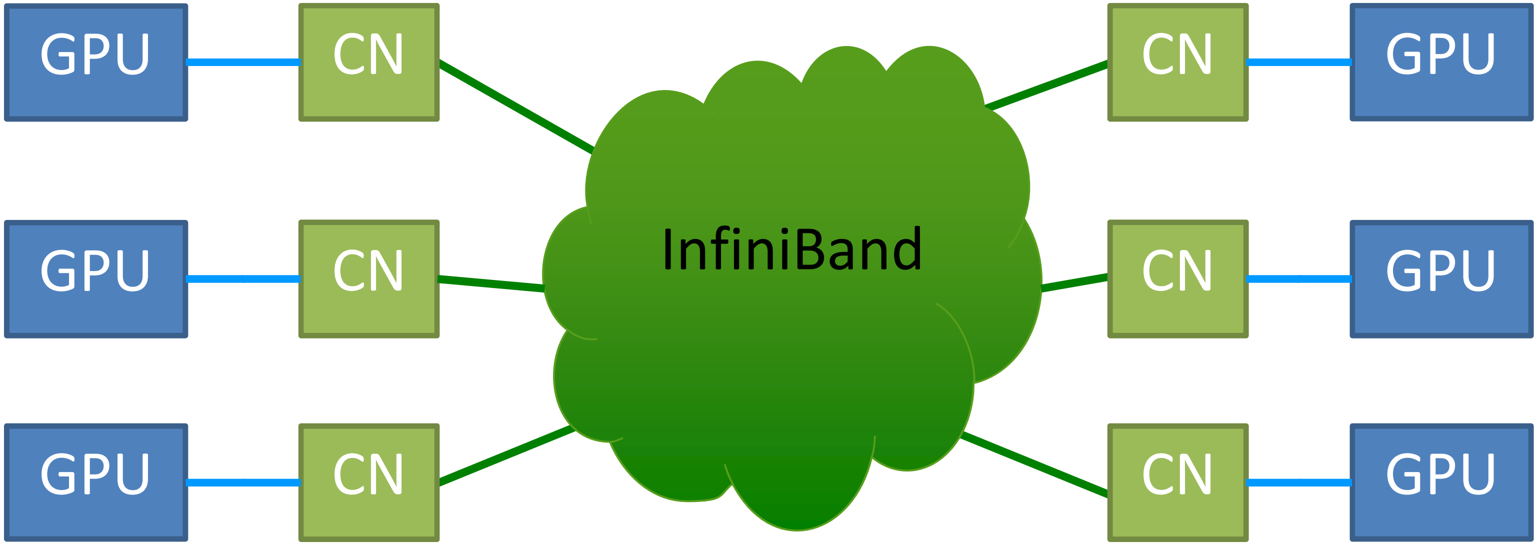
Goals

- Selection of power efficient components + Maximization of hardware utilization
- Component power tuning
  - Voltage tuning algorithm reaches O(10%) gain
- Energy-efficient cooling system
  - Avoid fans, air-conditioners as they are a significant source of power consumption; water cooling more energy efficient
  - Node-card cooled by conduction → dry connection
  - Water-cooled cold-plate



EU Exascale Project DEEP (12/2011-08/2015)

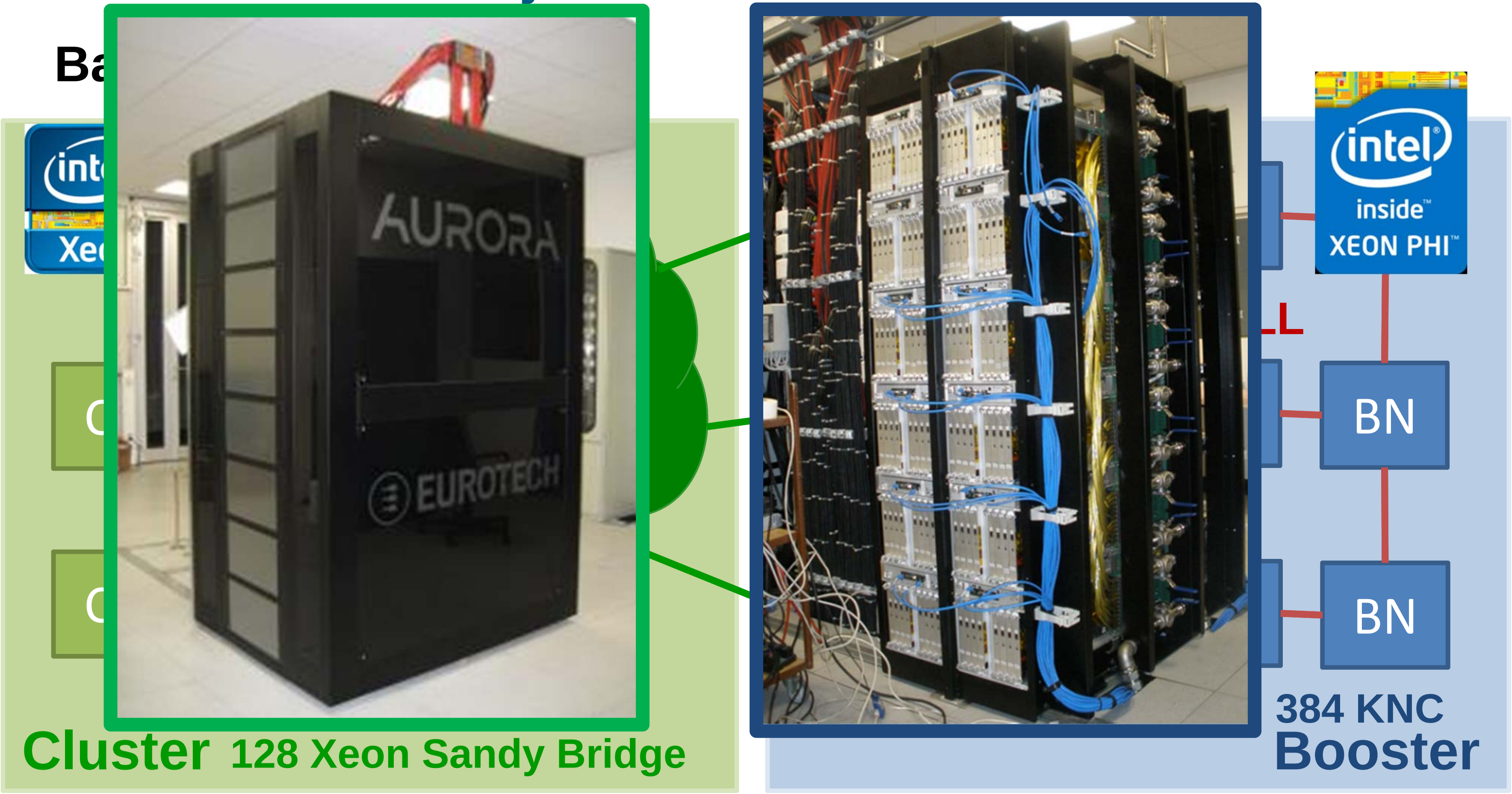
Starting point: Traditional cluster with GPUs



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

EU Exascale Project DEEP

Ba



Cluster 128 Xeon Sandy Bridge

384 KNC Booster

Low/Medium scalable code parts

Highly scalable code parts

EU Exascale Project DEEP

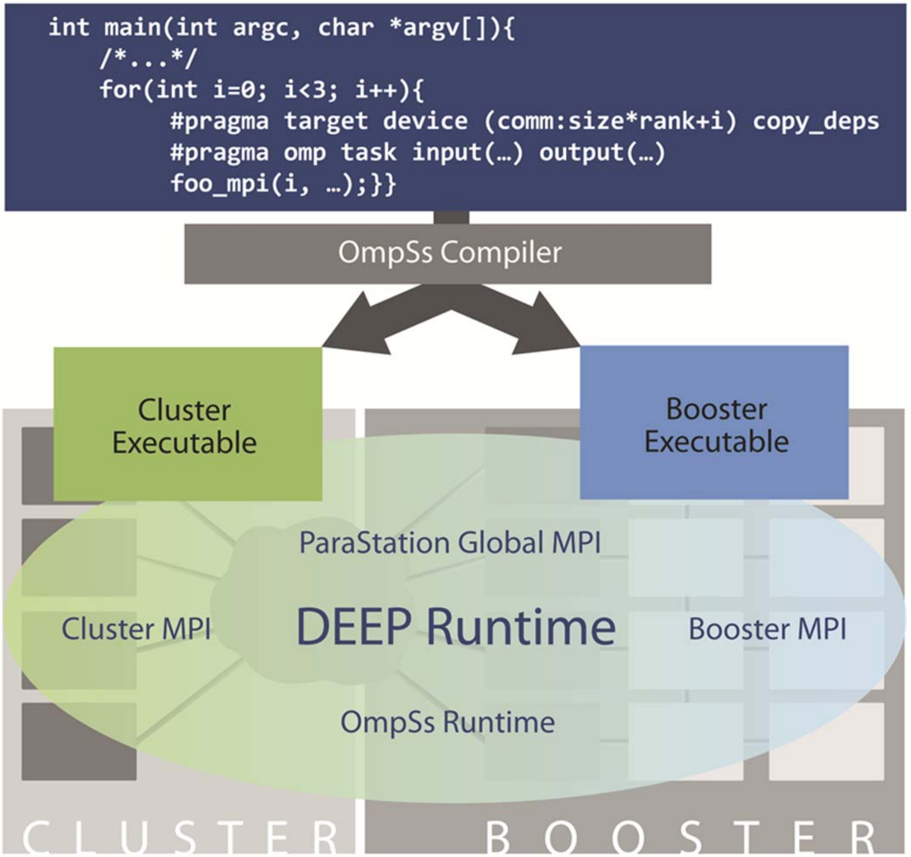
Application running on DEEP

Source code

Compiler

Application binaries

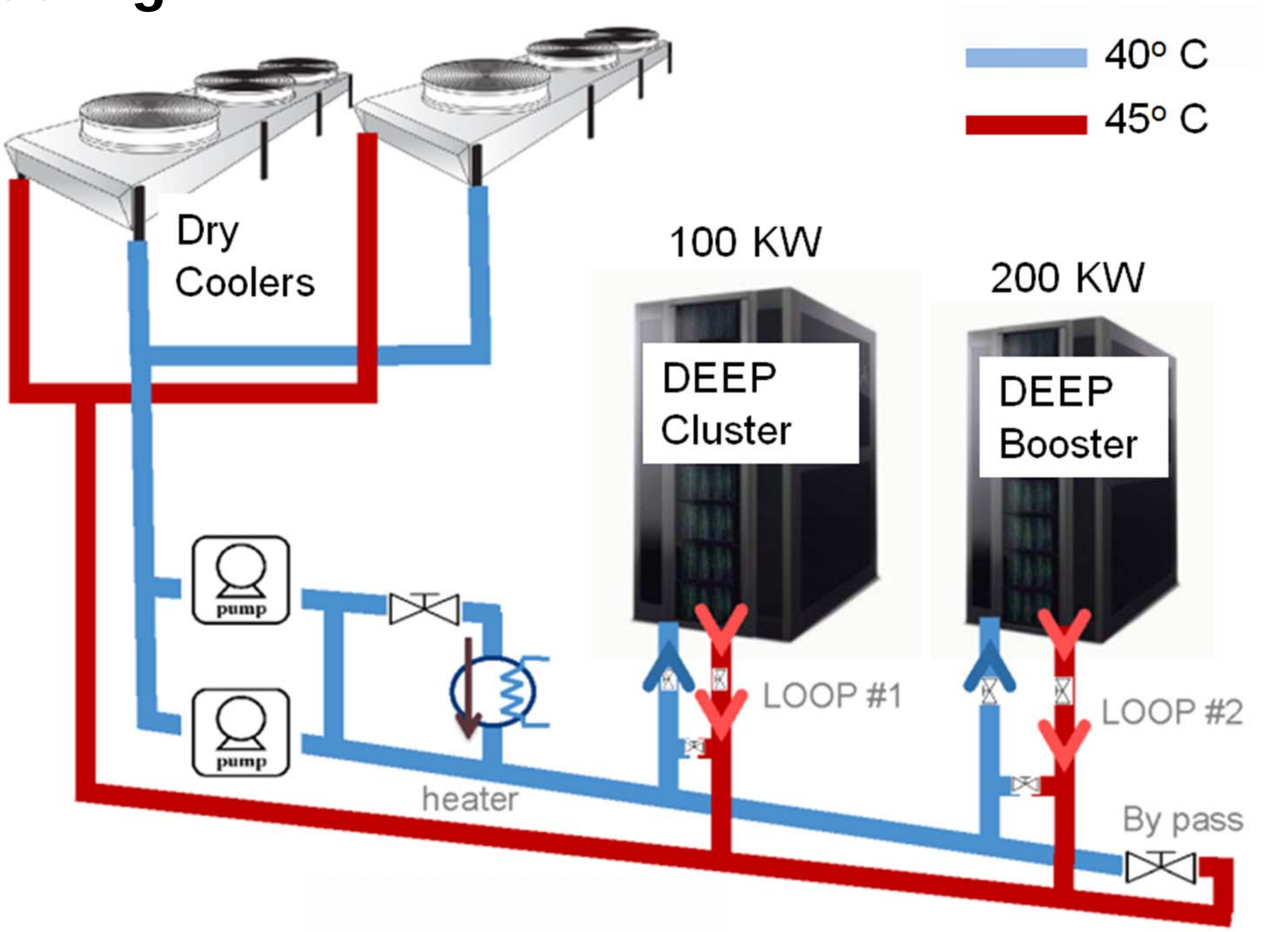
DEEP Runtime



```
int main(int argc, char *argv[]){
    /*...*/
    for(int i=0; i<3; i++){
        #pragma target device (comm:size*rank+i) copy_deps
        #pragma omp task input(…) output(…)
        foo_mpi(i, …);}}
```

EU Exascale Project DEEP

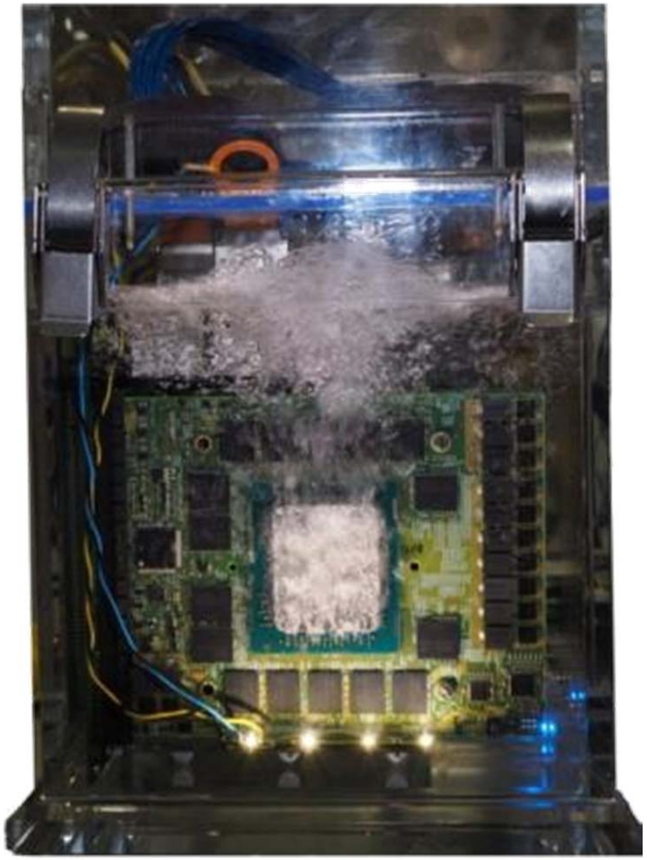
Free cooling



EU Exascale Project DEEP

GreenICE system

- Alternative Booster implementation
  - Interconnect EXTOLL ASIC “Tourmalet”
  - 32 KNC-nodes system
  - Implement 4x4x2 topology, with Z dimension open
- Experiment with immersion Cooling
  - 2-phase NOVEC liquid from 3M
  - Evaporates at about 50°C
  - Condensates again in a water cooling pipe



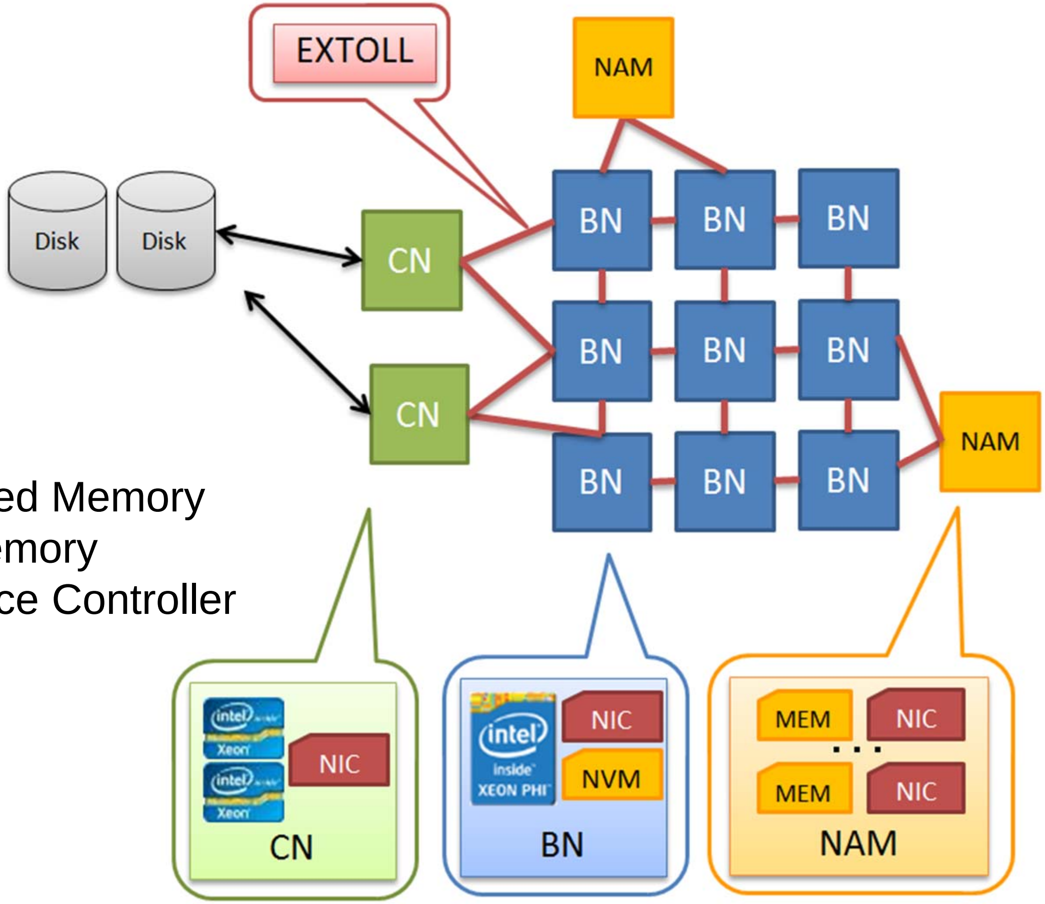
EU Exascale Project DEEP-ER (10/2013-03/2017)

Objectives

- DEEP-ER extends the Cluster-Booster architecture of DEEP by a highly scalable **I/O** system and a recover mechanism (**resiliency**) for applications that failed due to hardware errors
- Leverage new **memory technology and hierachies**
- Build a **prototype** based on Intel MIC (KNL)
- Develop a highly scalable and efficient **I/O subsystem**, based on FhGs BeeGFS and using the I/O middleware SIONlib and Exascale 10
- Extend the DEEP **Programming Model** based on OmpSs
- Seven important **HPC applications** are optimised demonstrating the usability, performance and resiliency of the DEEP-ER Prototype

EU Exascale Project DEEP-ER

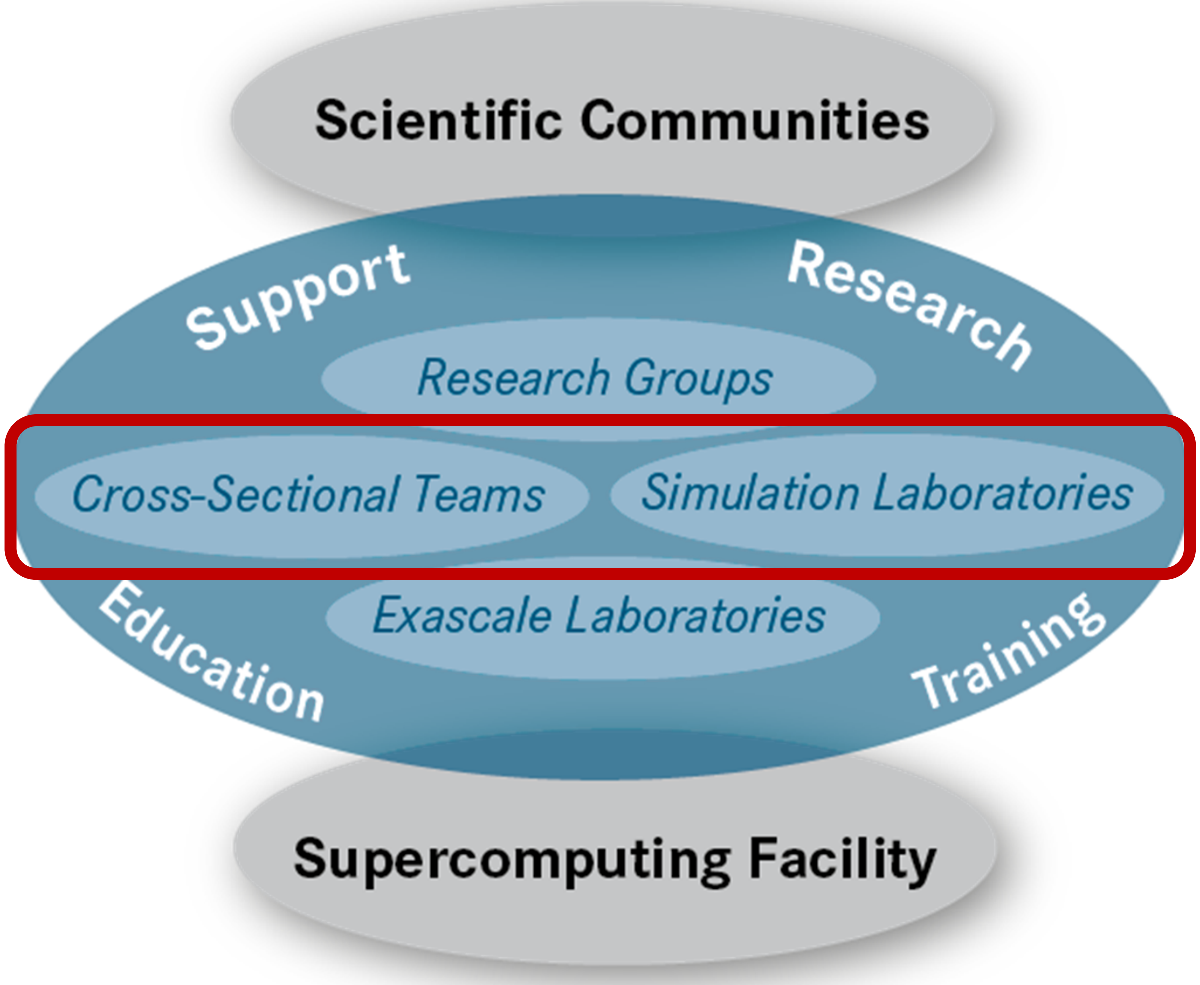
Towards a stand-alone booster system



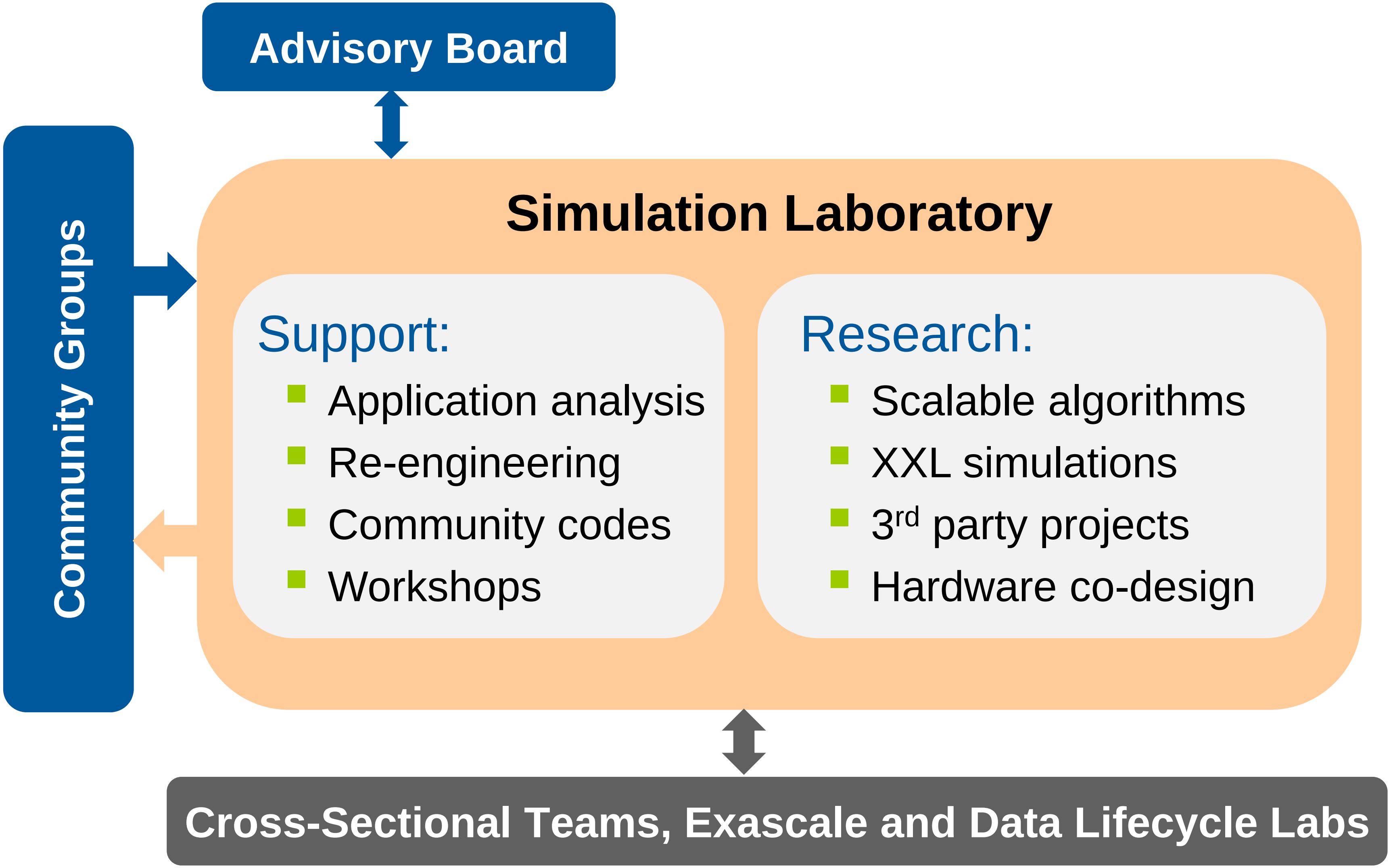
Legend:

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

R&D and Application Support at the Jülich Supercomputing Centre



The Simulation Laboratory as HPC Enabler



Advisory Board

Community Groups

Simulation Laboratory

Support:

- Application analysis
- Re-engineering
- Community codes
- Workshops

Research:

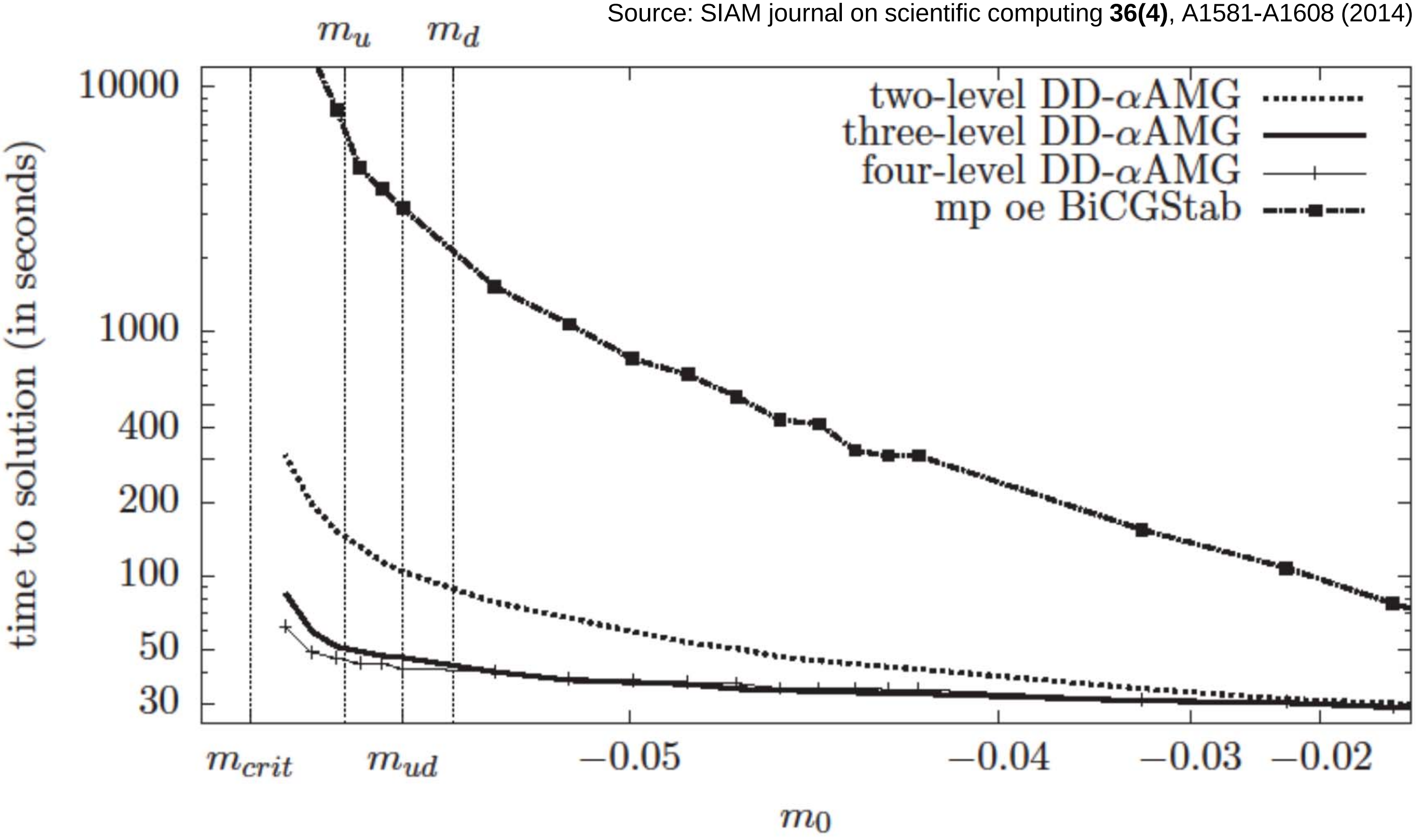
- Scalable algorithms
- XXL simulations
- 3rd party projects
- Hardware co-design

Cross-Sectional Teams, Exascale and Data Lifecycle Labs

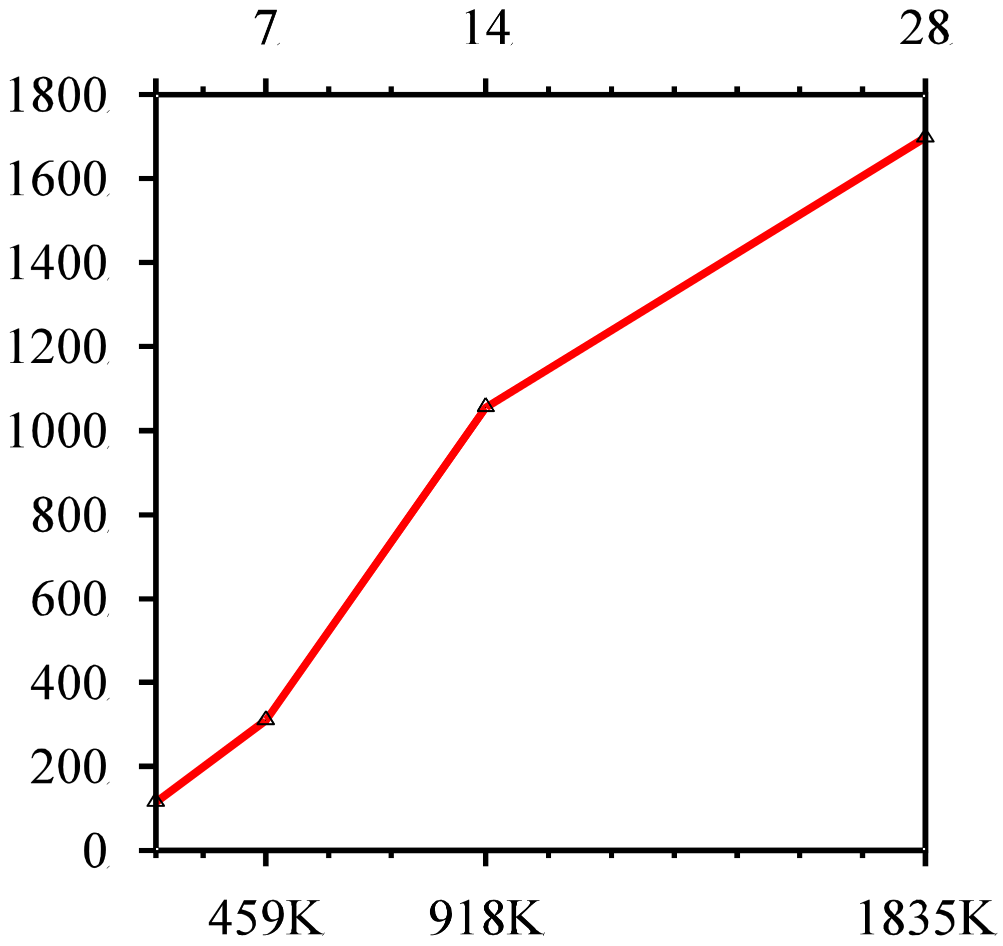
SimLab Nuclear and Particle Physics

Algorithm Research

Source: SIAM journal on scientific computing **36(4)**, A1581-A1608 (2014)



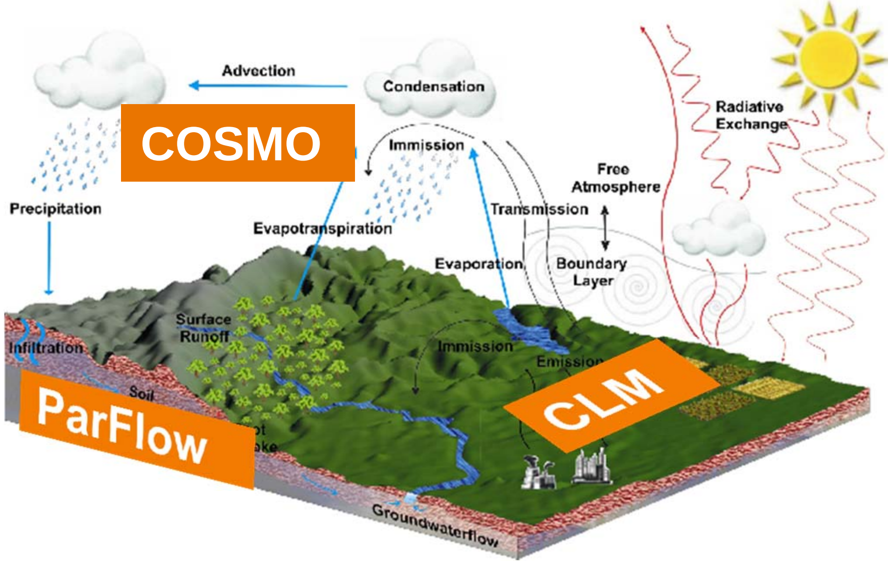
SimLab Nuclear and Particle Physics
Strong scaling analysis of LQCD simulation software



SimLab TerrSys

TerrSysMP:

- Fully integrated groundwater-vegetation-atmosphere simulation platform; earth system models at regional scale
- Water cycle processes and variability across scales
- Climate and land use impacts



- Scalasca performance analysis
- Refactoring of OASIS-MCT coupling interface to remove scaling bottleneck
- Scaling now to 32k cores: 64x increased problem size!

HPSC TerrSys

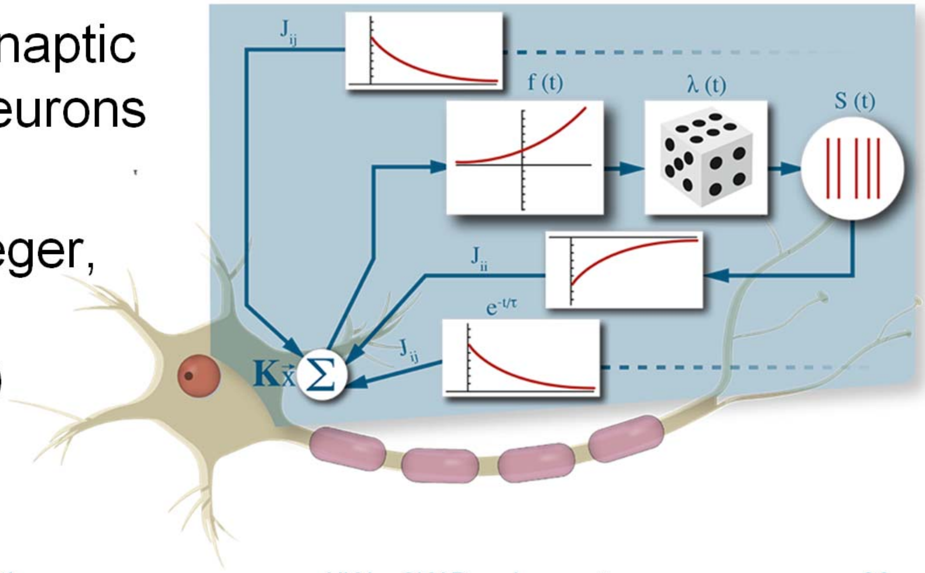
A B C/J geoverbund aachen bonn köln jülich

SimLab Neuroscience

- Alias: Bernstein Facility for Simulation and Database Technology
- Part of Helmholtz-funded activity 'Supercomputing and Modeling for the Human Brain' and JARA-HPC
- Supporting the European FET Flagship 'Human Brain Project'
- Bridge between Comp. Neuroscience community and HPC

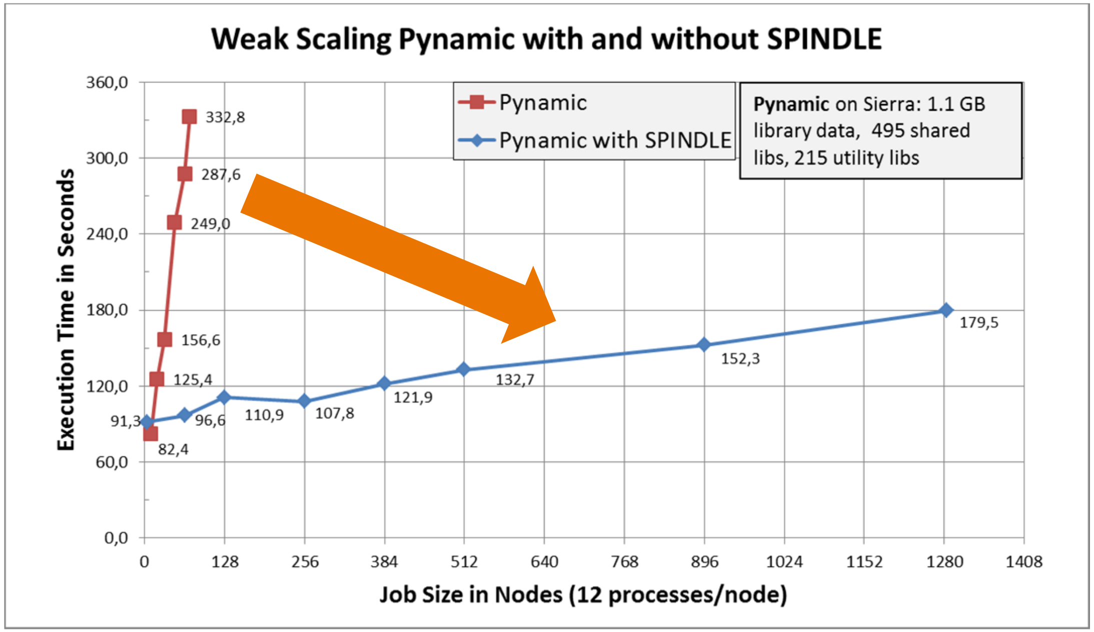
Recent Highlight

- Reconstruction of recurrent synaptic connectivity of thousands of neurons from simulated spiking activity
- Y. Zaytsev, A. Morrison, M. Deger, Journal of Computational Neuroscience **39** (1) 77 (2015)



CST Application Optimisation
Scalable library loading with SPINDLE

- Improving the library-loading performance of dynamically-linked HPC applications and Python files



Pynamic on Sierra: 1.1 GB library data, 495 shared libs, 215 utility libs

In cooperation with

Lawrence Livermore National Laboratory

- W. Frings et al., Best Paper award, 27th ICS, June 2013

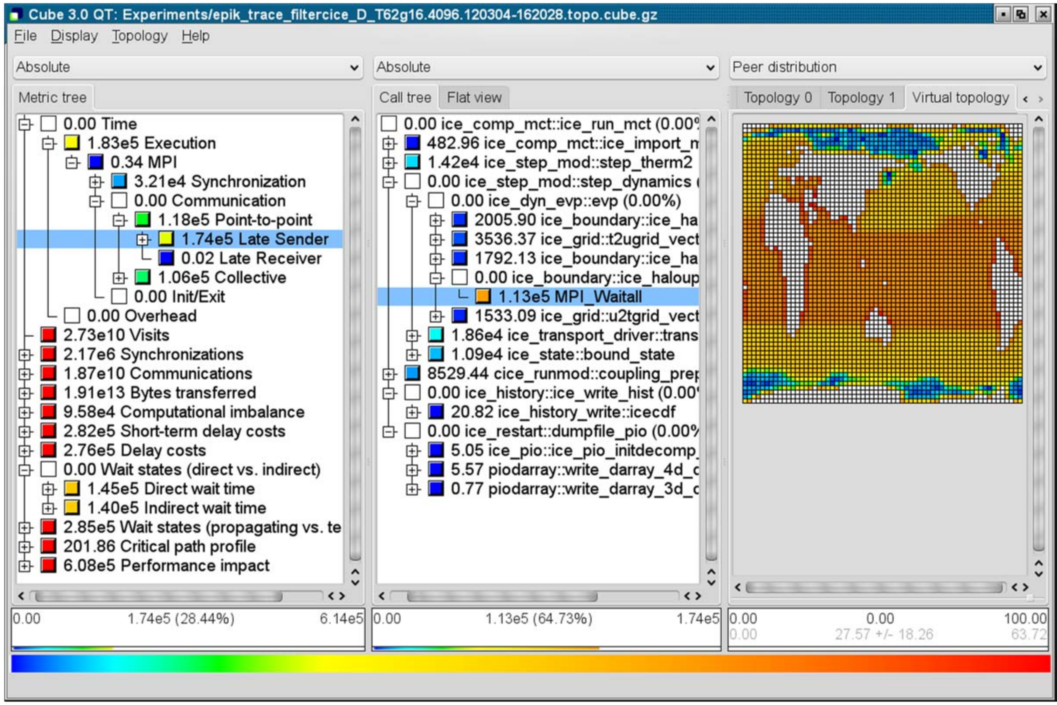
CST Parallel Performance: scalasca

Scalable **A**nalysis of
Large **Sc**ale **A**pplications   http://www.scalasca.org

Highly scalable parallel performance tool

Successful experiments with up to 1 million threads

Basis for user support, research and training



```
0.00 Time
    1.83e5 Execution
        0.34 MPI
            3.21e4 Synchronization
            0.00 Communication
                1.18e5 Point-to-point
                    1.74e5 Late Sender
                    0.02 Late Receiver
                1.06e5 Collective
            0.00 Init/Exit
    0.00 Overhead
2.73e10 Visits
2.17e6 Synchronizations
1.87e10 Communications
1.91e13 Bytes transferred
9.58e4 Computational imbalance
2.82e5 Short-term delay costs
2.76e5 Delay costs
0.00 Wait states (direct vs. indirect)
    1.45e5 Direct wait time
    1.40e5 Indirect wait time
2.85e5 Wait states (propagating vs. te
201.86 Critical path profile
6.08e5 Performance impact

0.00          1.74e5 (28.44%)          6.14e5
```

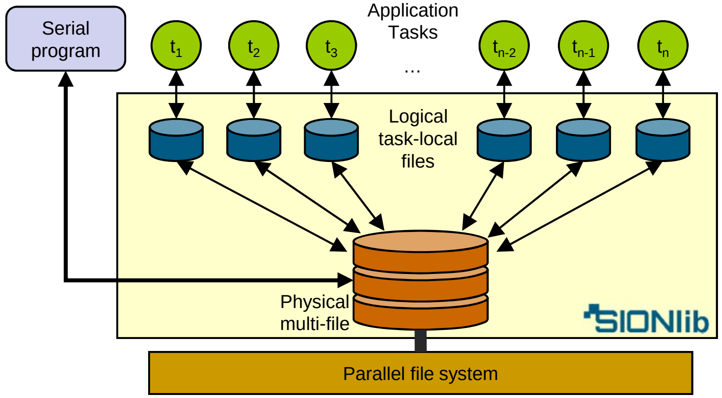
```
0.00 ice_comp_mct::ice_run_mct (0.00%
482.96 ice_comp_mct::ice_import_m
1.42e4 ice_step_mod::step_therm2
0.00 ice_step_mod::step_dynamics (
    0.00 ice_dyn_evp::evp (0.00%)
        2005.90 ice_boundary::ice_ha
        3536.37 ice_grid::t2ugrid_vect
        1792.13 ice_boundary::ice_ha
        0.00 ice_boundary::ice_haloup
            1.13e5 MPI_Waitall
        1533.09 ice_grid::u2tgrid_vect
    1.86e4 ice_transport_driver::trans
    1.09e4 ice_state::bound_state
8529.44 cice_runmod::coupling_prep
0.00 ice_history::ice_write_hist (0.00'
    20.82 ice_history_write::icecdf
0.00 ice_restart::dumpfile_pio (0.00%
    5.05 ice_pio::ice_pio_initdecomp_
    5.57 piodarray::write_darray_4d_c
    0.77 piodarray::write_darray_3d_c

0.00          1.13e5 (64.73%)          1.74e5
```

```
0.00          0.00                100.00
0.00          27.57 +/- 18.26     63.72
```

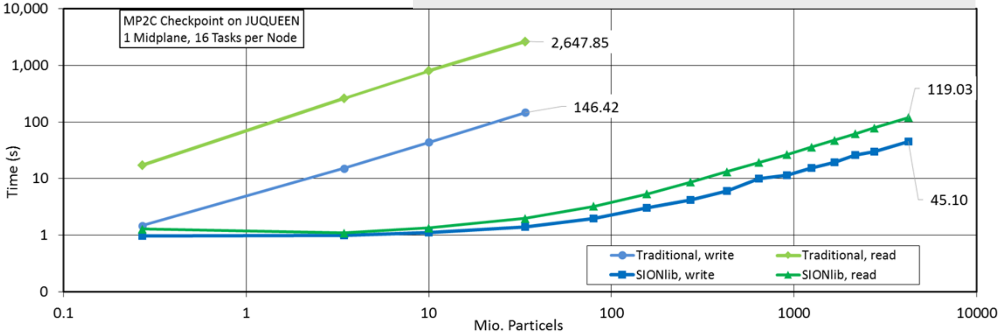
CST Application Optimisation

SIONlib: Parallel I/O to task local files at large scale

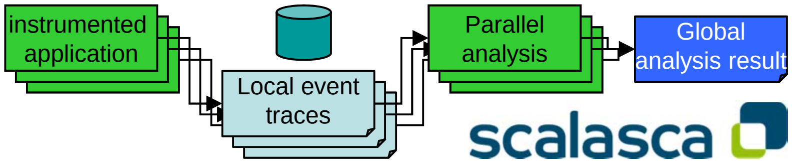


→ Checkpointing

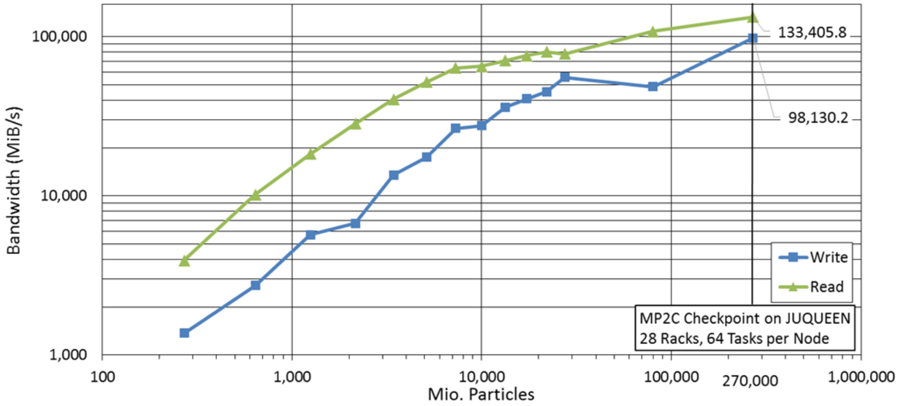
MP2C: Mesoscopic hydrodynamics + MD
Speedup and higher particle numbers through SIONlib integration



→ Tool Support



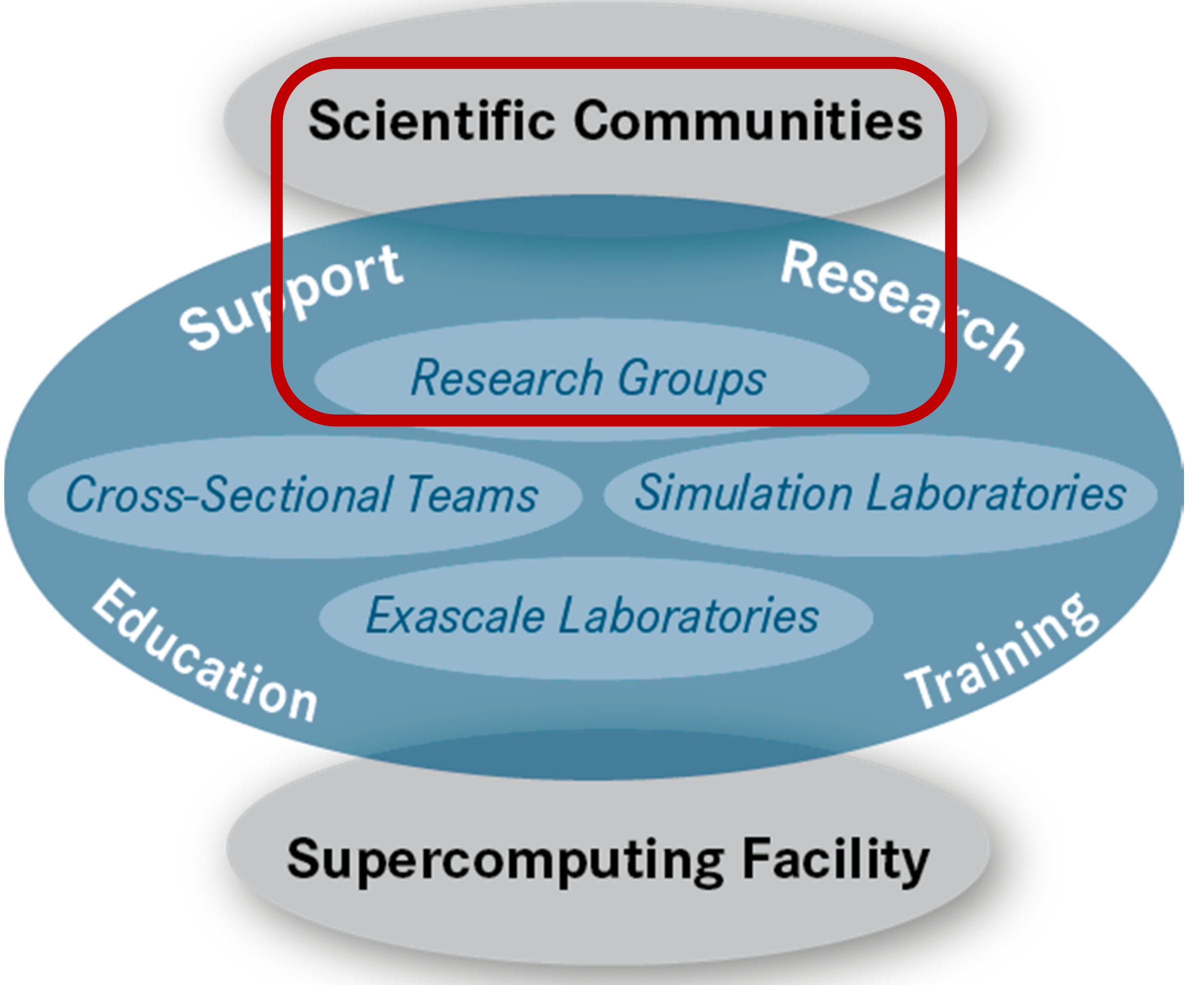
Score-P: Scalable Performance Measurement Infrastructure for Parallel Codes



→ Scalability

1.8 M tasks, ~100 GiB/s

R&D and Application Support at the Jülich Supercomputing Centre



Scientific Communities

Support

Research

Research Groups

Cross-Sectional Teams

Simulation Laboratories

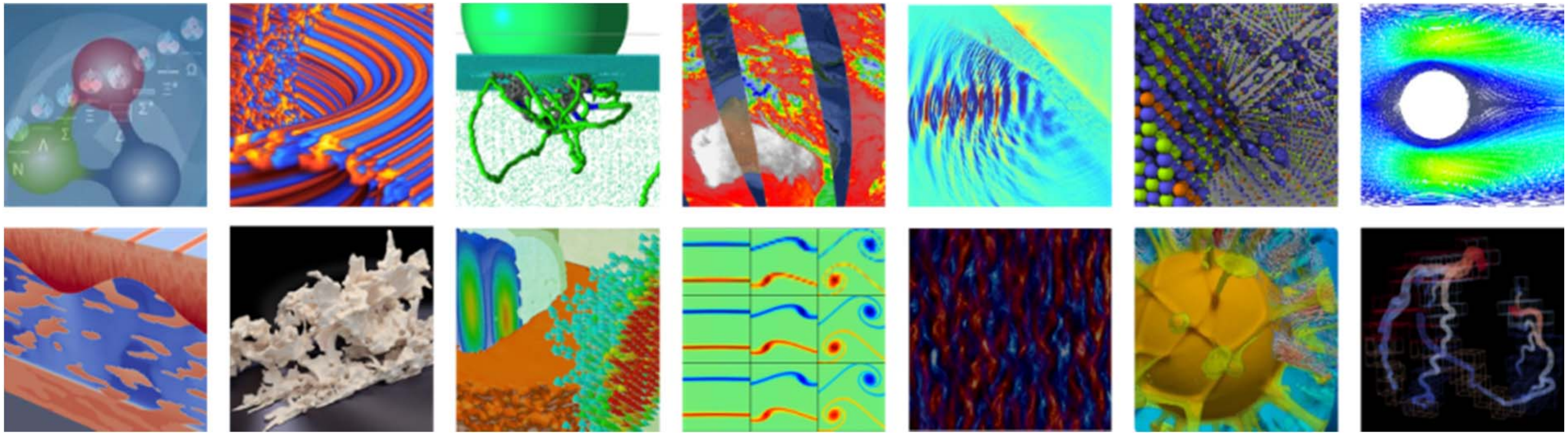
Exascale Laboratories

Education

Training

Supercomputing Facility

Promoting Exascale-Ready Applications



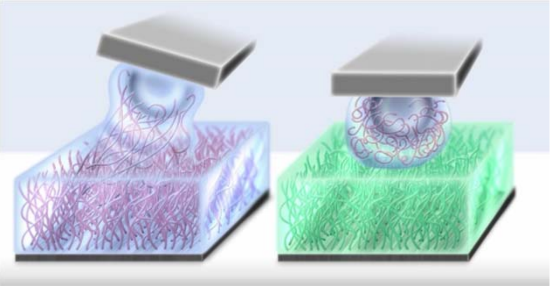
23 codes from a wide range of science fields, scaling across 458,752 cores and up to 1.8 million threads on JUQUEEN:

CoreNeuron, dynQCD, FE2TI, FEMPAR, Gysela, ICON, IMD, JURASSIC, JuSPIC, KKRnano, MP2C, muPhi, Musubi, NEST, OpenTBL, PEPC, PMG+PFASST, PP-Code, psOpen, SHOCK, Terra-Neo, waLBerla, ZFS

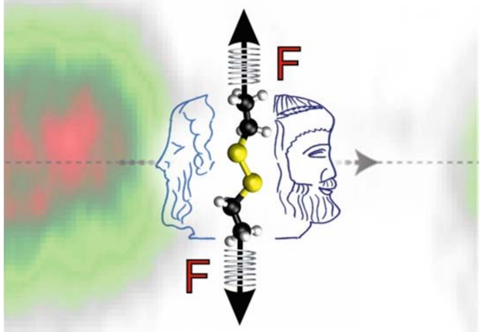
http://www.fz-juelich.de/ias/jsc/high-q-club

High Impact Publications

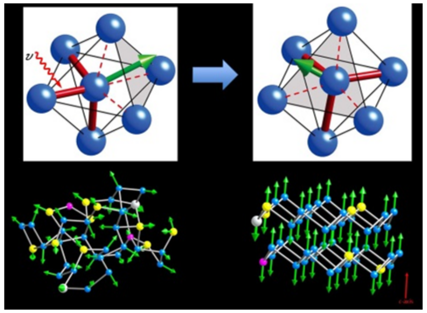
Users of the facility at JSC produce about 250 publications per year



S. de Beer, M. Müser
Nature Communications **5**
(2014) 3781

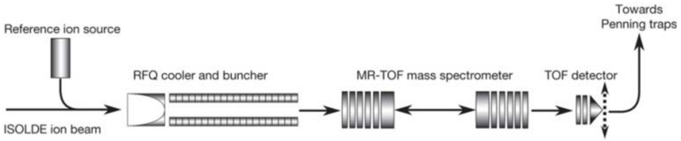


D. Marx et al.,
Nature Chemistry **5**
(2013) 685

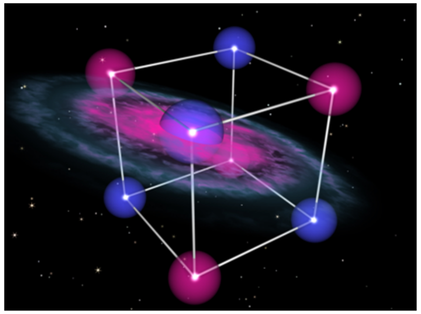


R.O. Jones et al.,
Nature Materials **10**
(2011) 129

Sz. Borsanyi et al.,
Science **347** (2015) 6229



A. Schwenk et al.,
Nature **498** (2013) 346



M. Lezaic et al.,
Nature Materials **9**
(2010) 649

End of Presentation