DKRZ Status (HW & SW)

Joachim Biercamp

Panos Adamidis, Jörg Behrens, Irina Fast, Moritz Hanke, Thomas Jahns, Niklas Röber

Deutsches Klimarechenzentrum
Hamburg

ORGANISATION

HARDWARE

SOFTWARE

TOOLS

CO-DESIGN

ORGANISATION

HARDWARE

SOFTWARE

TOOLS

CO-DESIGN

The German Climate Computing Center (DKRZ)

Founded in 1987 as a national institution

Operated as a non-profit limited company with four shareholders

- Max Planck Society for Research (55%)
- The City of Hamburg represented by the University of Hamburg (27%)
- Alfred Wegener Research Institute in Bremerhafen (9%)
- Helmholtz Center for Research in Geesthacht (9%)

Mission

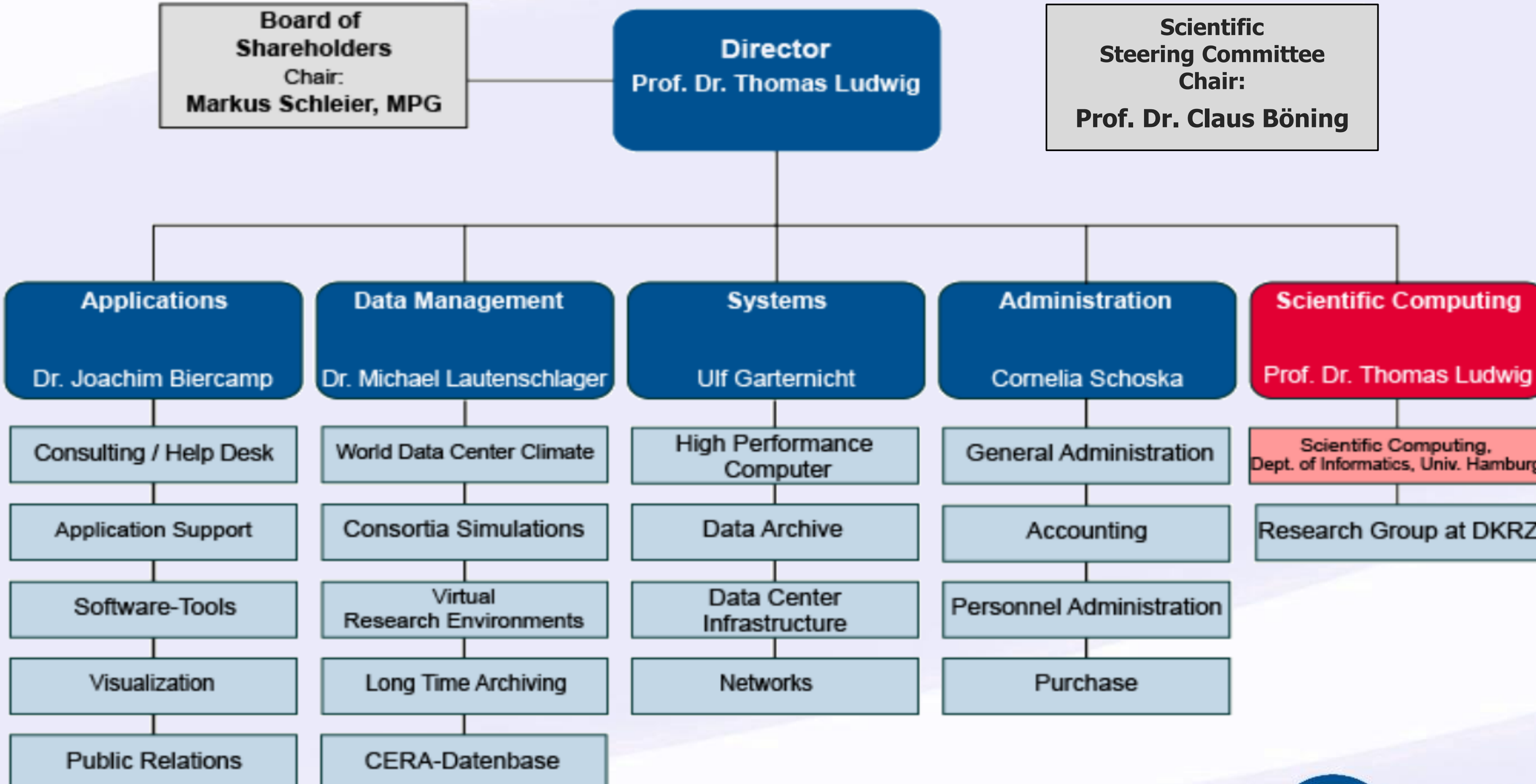
DKRZ – Partner for Climate Research

Maximum Compute Performance.

Sophisticated Data Management.

Competent Service.

DKRZ´s Structure



DKRZ

Money

Budget

- > € 8M per year regular budget
  - ~ € 2.5M per year for electrical power
  - ~ € 0.5M for tapes
- Additional third party funding

HW Funding and Procurements

- Every 5-7 years
- 2014-2019: € 41M for computer, storage, etc.

ORGANISATION

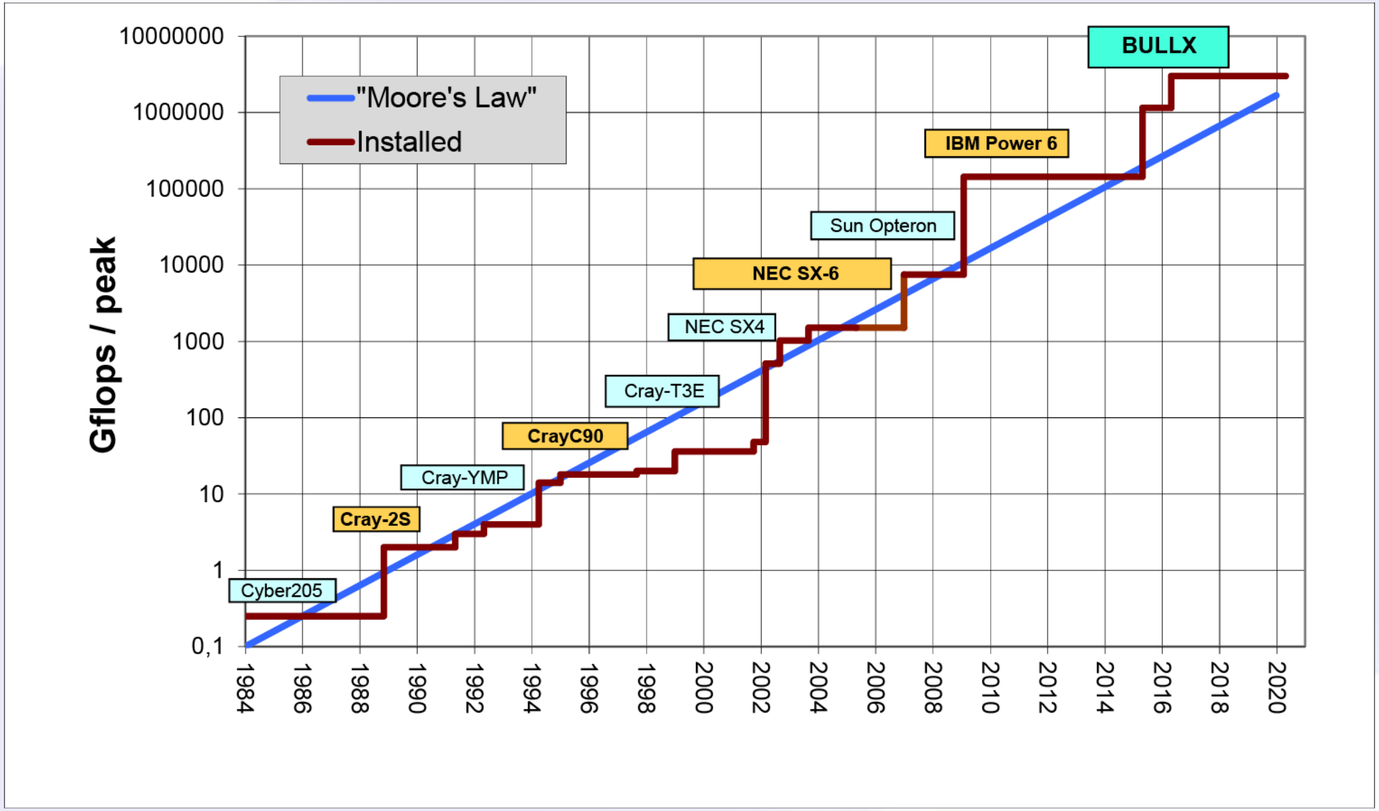
HARDWARE

SOFTWARE

TOOLS

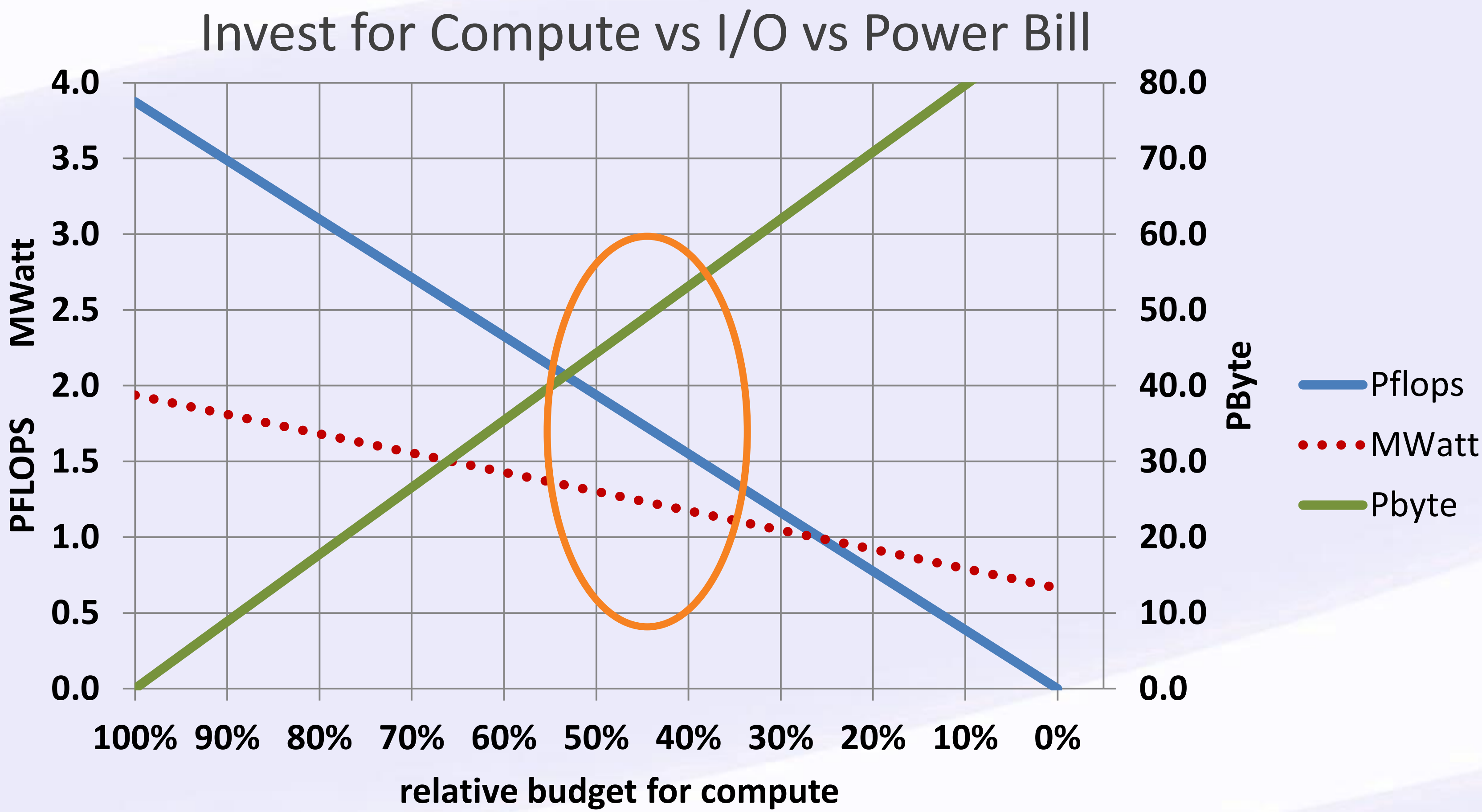
CO-DESIGN

Evolution of compute performance



DKRZ

Pre-Procurement Considerations (1)



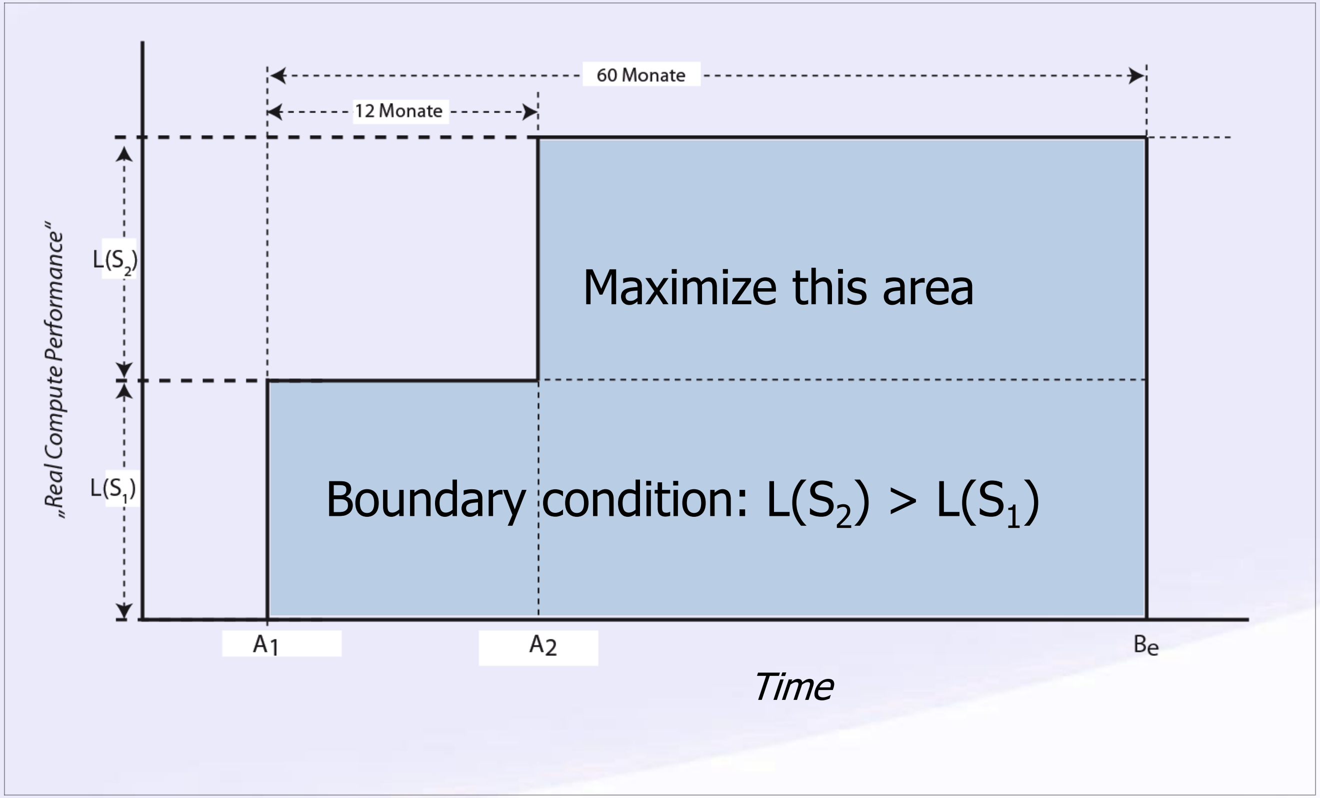
DKRZ

Pre-Procurement Considerations (2)

Principle selection criteria for new system

- Maximize application performance
  - measured via application benchmarks
- Stay inside given envelope for electrical power
  - Max 1350 MW in mean everyday operations
  - includes cooling overhead
- Deliver 45 PByte of usable (net) disk space
  - Specify price for optional extension of disk space

How to define application performance



How to define application performance

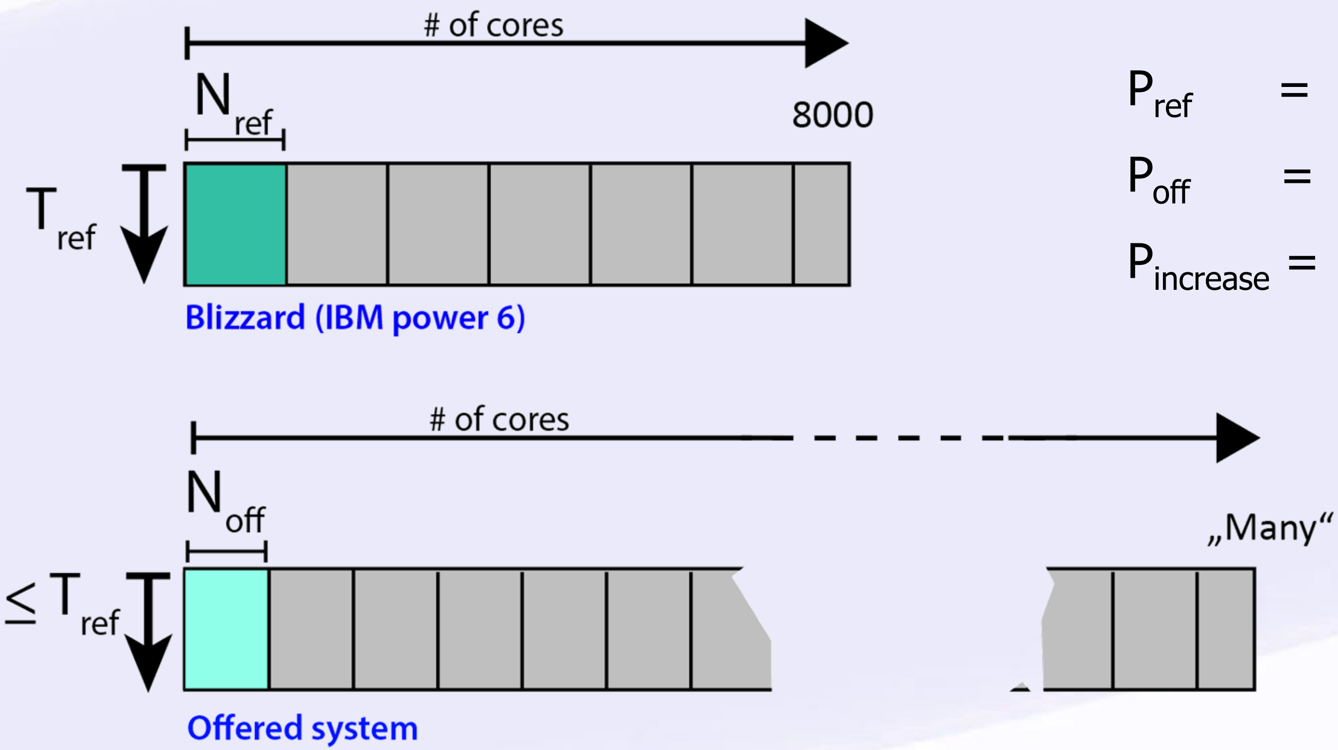
A suite of real models selected by user group.

- Configuration (=resolution) as expected to be used in 2015-20
- (but no realistic I/O)

For each: maximal allowed time-to-solution

- The number of cores used for to beat this time defined a troughput for this individual BM on the offered system
- A weighted mean of these throughputs is score of the offer

How to define application performance



$P_{ref} = 8000 : N_{ref}$

$P_{off} = M_{any} : N_{off}$

$P_{increase} = P_{off} : N_{ref}$

DKRZ

How to define application performance

"Real" Application Benchmarks

- **ICON global, 20km (N_{ref}: 7872)**
- **ICON local area 416m (N_{ref}: 4096)**
- **CCLM (COSMO_RAPS_5.1_CLM) 12 km (N_{ref}: 1024)**
- **FESOM ocean unstructured grid (N_{ref}: 1024)**
- **EMAC T42L90, 250 km (256)**
- MPI-ESM (coupled ESM, T63L95/TP04L40, CMIP5 version) (**N_{ref}**: 192)
- METRAS (openMP code, meso-scale Atmosphere) (**N_{ref}**: 32)
- EH6-CDI-PIO (Test for IO server)

Benchmarking electrical power

Mix of the individual benchmarks to simulates mean everyday load on the system

- Fills (nearly) the whole system. All jobs run concurrently
- Settings and performance (e.g. turbo/non turbo, #cores, SYPD) have to be identical than those used to deliver the performance for individual BMs

This troughput benchmark is used

1. To measure average electrical power
2. To guarantee that no tricks can be played for individual measurements

The new HPC system
(„Mistral“)

BullX B720 DualSocket DLC blades

- SLURM
- Intel MPI and or BULLX MPI
- 64 Gbyte/node (10% of the nodes 128 GB/node)
- Lustre file system (Xyratec/Seagate)

Phase 1; April 2015;

- Performance increase ("capacity") vs bizzard: ca 6 x
- Haswell, 12 core, 2,5 GHz
- Ca 36000 core
- Ca 1.5 TFlops peak
- 20 PByte net disk capacity

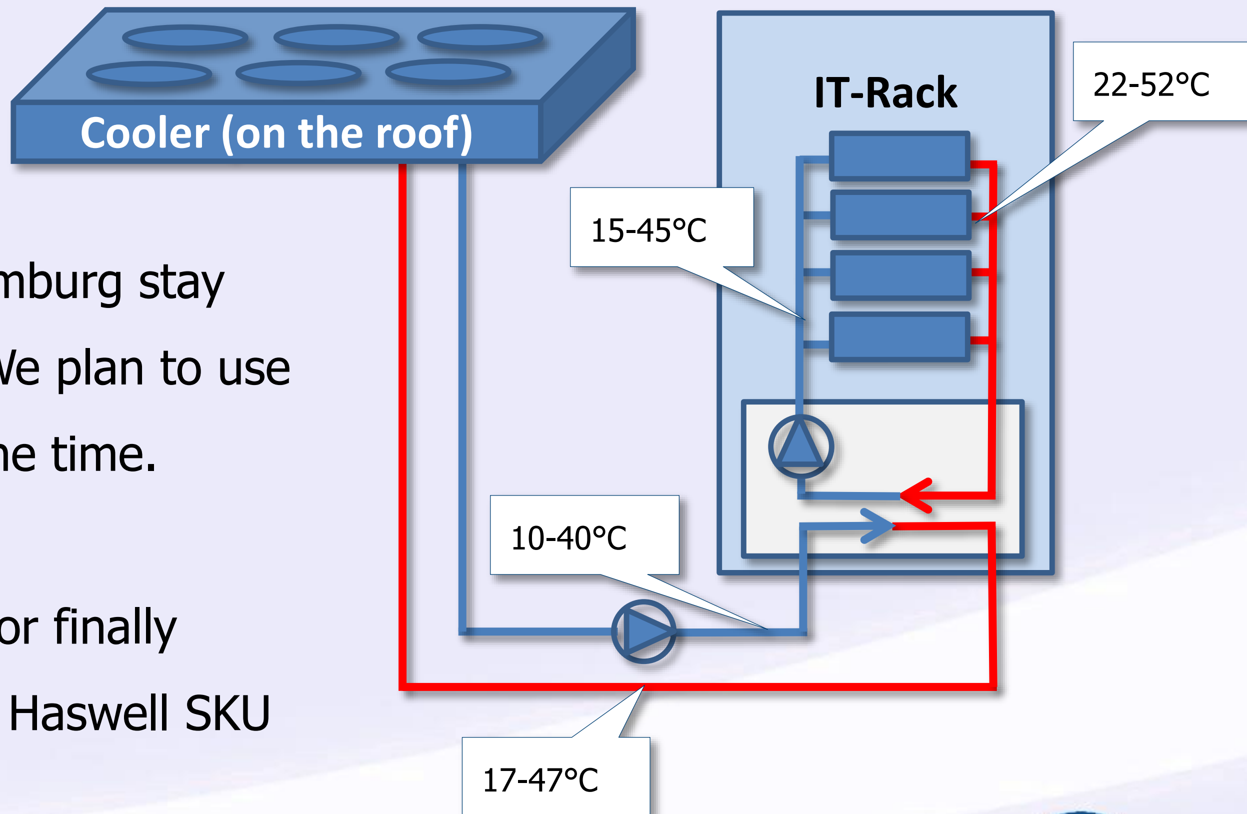
Phase 2; April 2016; In total:

- Performance increase vs blizzard: > 20 x
- Add new nodes, probably Broadwell
- Ca 80000 cores (total P1 + P2)
- > 3.2 Tflops peak
- 50 PByte net disk capacity (x8 vs blizzard)

DKRZ

Infrastructure

Direct Liquid Cooling (DLC) to keep the PUE low



Temperatures in Hamburg stay below 35° usually. We plan to use free cooling all of the time.

That is one reason for finally deciding on 2.5 Ghz Haswell SKU

The HSM System: We stay with HPSS (IBM)

- IBM X-Series x86-Server (2 Core-Server, 6 Diskmover, 5 Tapemover)
- 5 PB disk-cache (NetApp E-Series DS5300)
- 8 Oracle StorageTek SL8500 tape-libraries (75.000 Slots, 70 LTO-drives)

	Phase 1 (2015)	Final (2016)	
Capacity	200 PB (LTO 6)	500 PB (LTO 7)	
Agg. Bandwidth sust. read/write	6 GB/s	15 GB/s	
Agg. Bandwidth peak. read/write	12 GB/s	18 GB/s	
Annual storage volume	75 PB	75 PB	Ca x6 vs current system

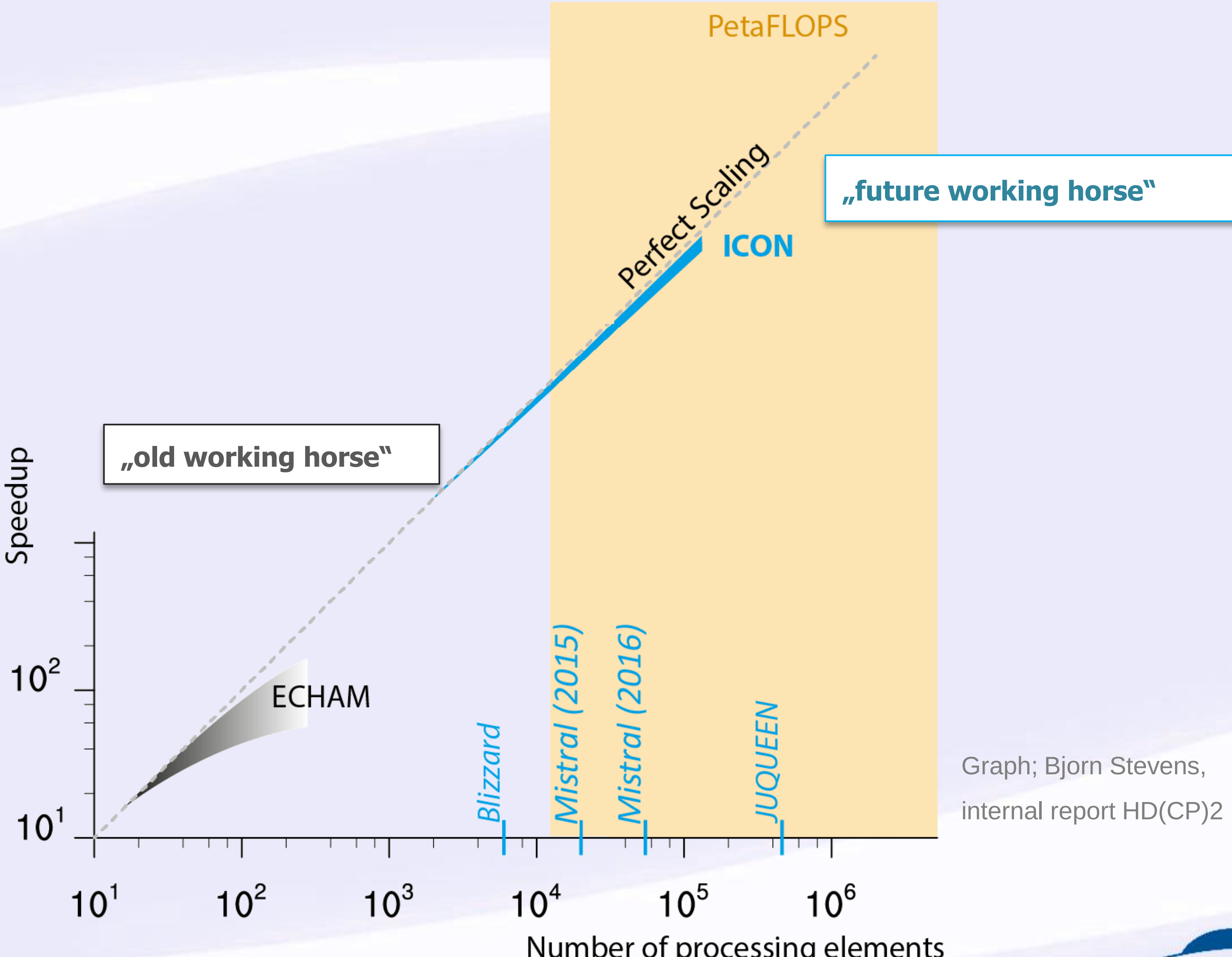
ORGANISATION

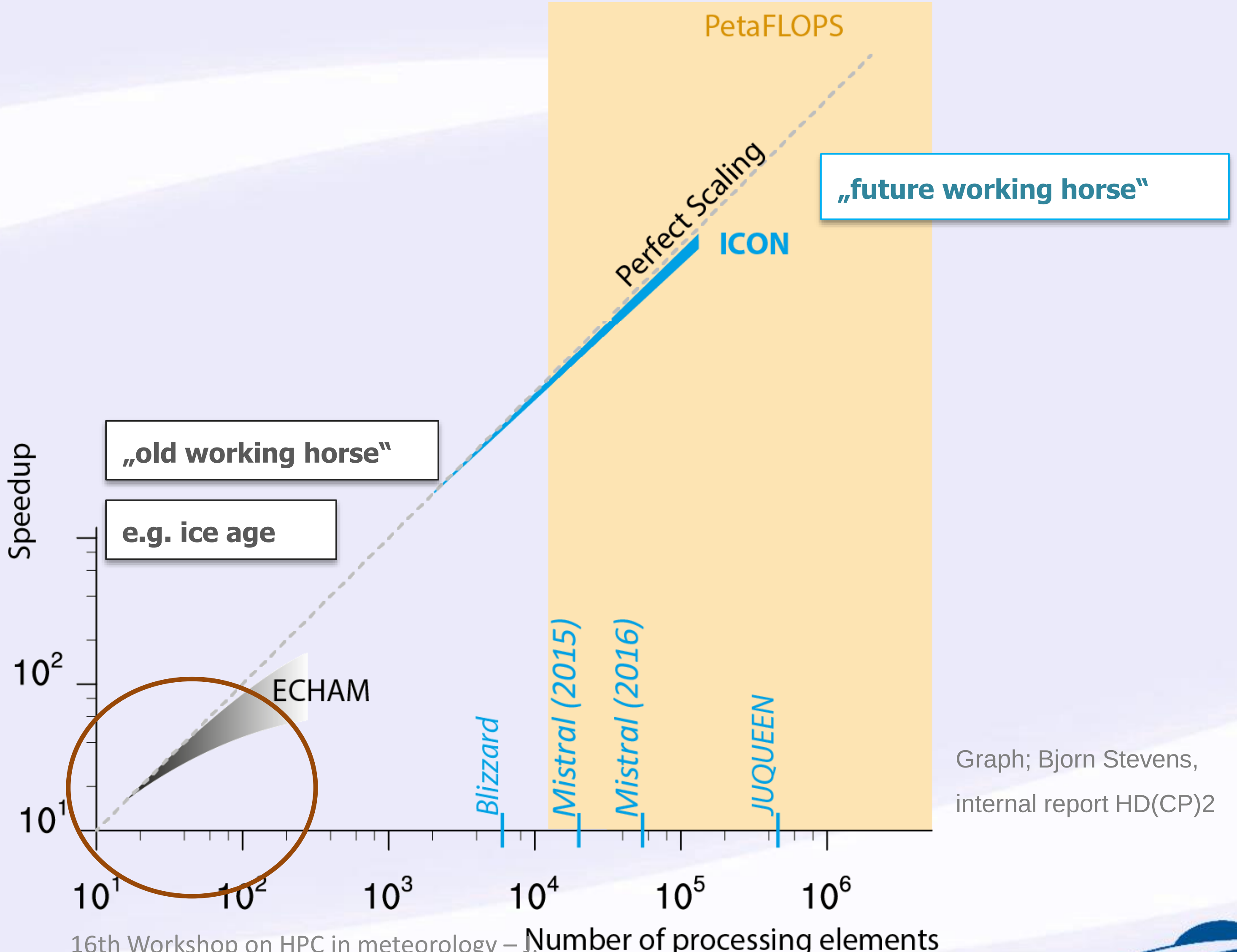
HARDWARE

SOFTWARE

TOOLS

CO-DESIGN

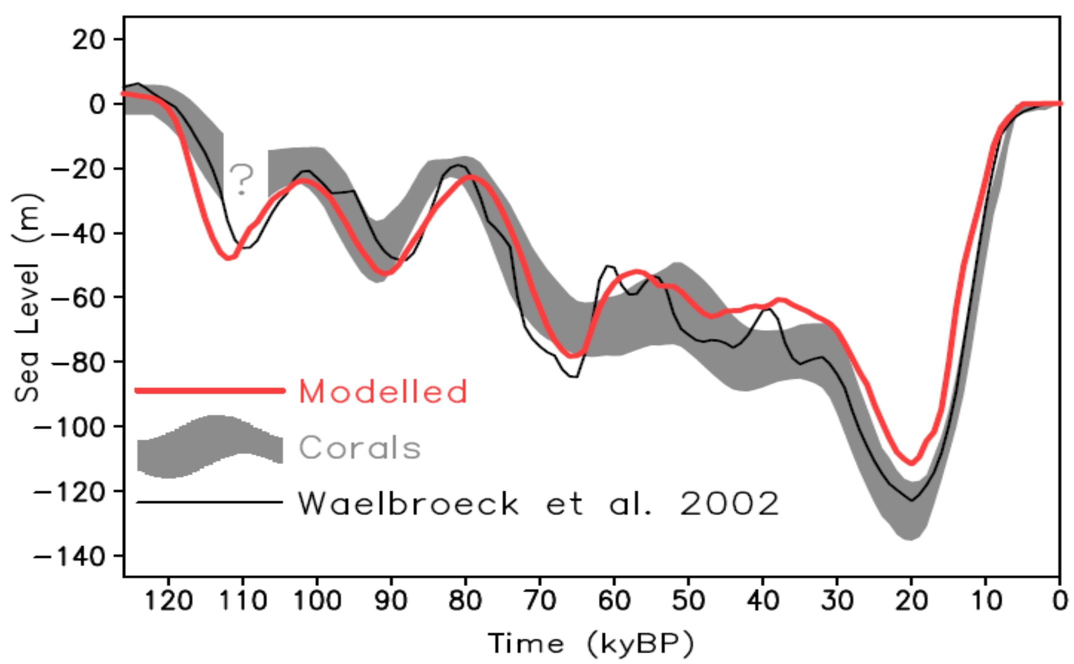

PetaFLOPS
„future working horse“
Perfect Scaling
ICON
„old working horse“
Speedup
ECHAM
Blizzard
Mistral (2015)
Mistral (2016)
JUQUEEN
Number of processing elements
Graph; Bjorn Stevens, internal report HD(CP)2



Graph; Bjorn Stevens, internal report HD(CP)2

National Climate Initiative „135 ka“

Project Goal:
Simulate the full range of climate variability over ice age cycle



Sea-level evolution during the last glacial cycle relative to present-day values. (Ganopolski et al., 2010, using CLIMBER)

DKRZ

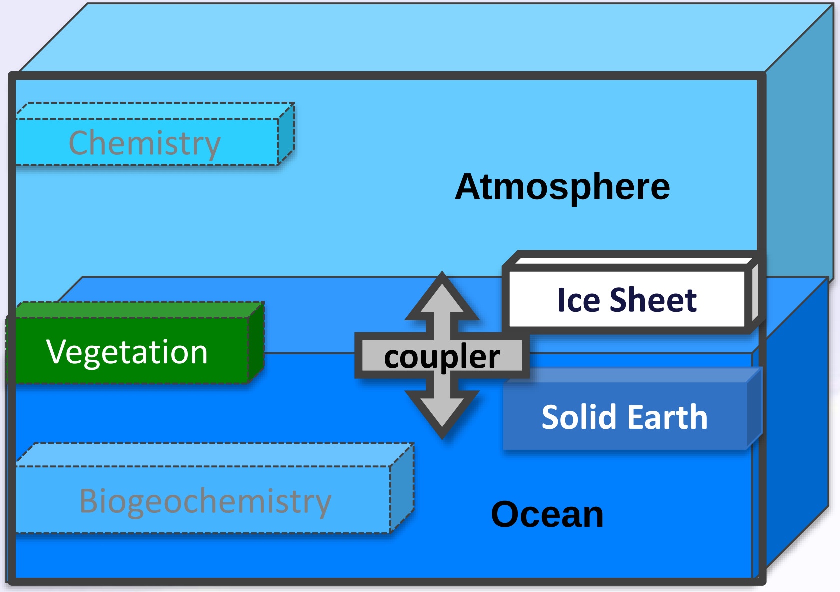
National Climate Initiative „135 ka"

Possible design of a coupled AOISE model (AOGCM-ISM-SE) that could be used for the transient simulations

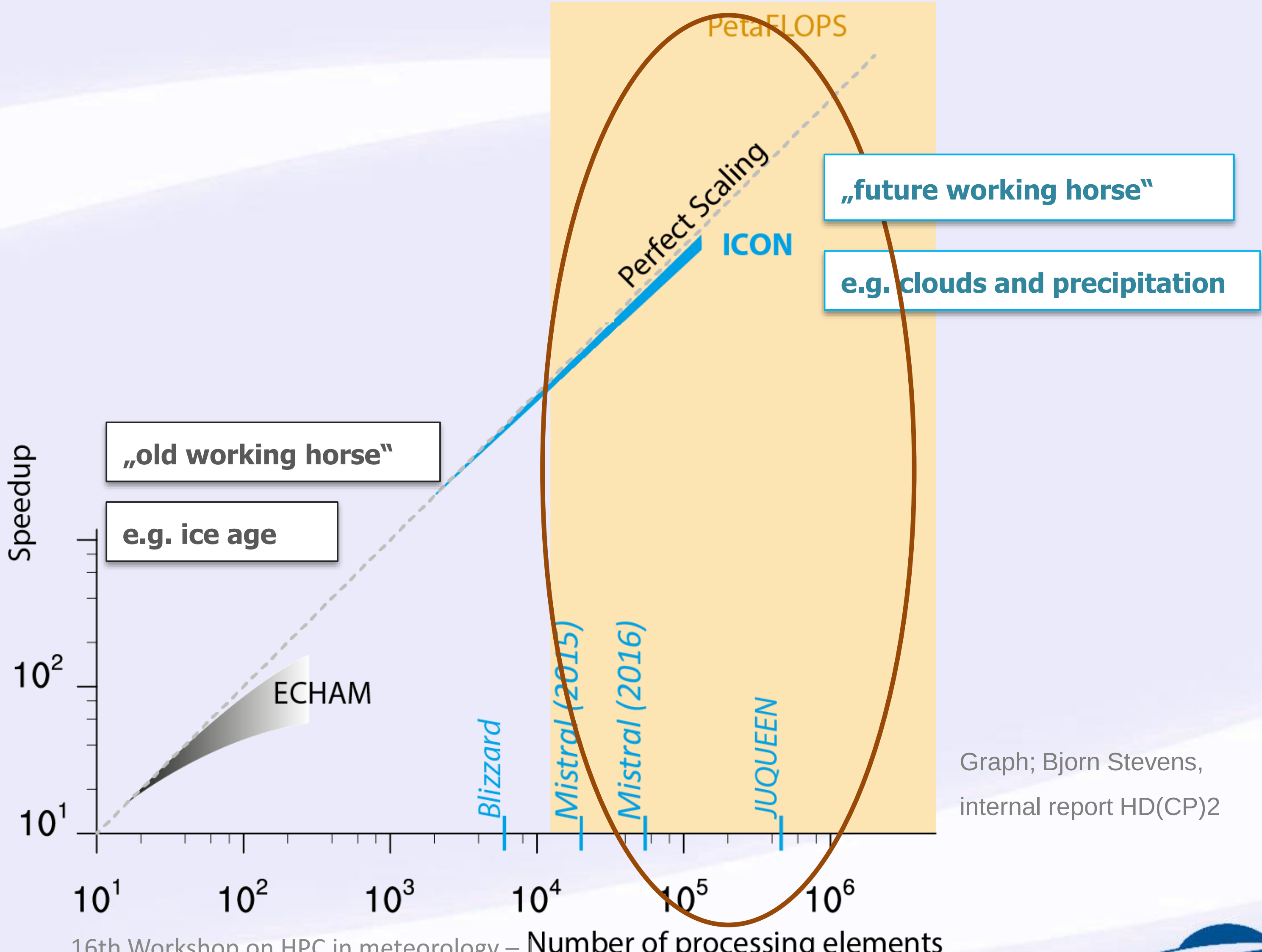
AOGCM: MPI-ESM is a candidate

Physical feedbacks (atmosphere-ocean-ice sheet-solid earth) for long-term climate change

Modular AOISE System Model as basis for the overall program



DKRZ



Graph; Bjorn Stevens, internal report HD(CP)2

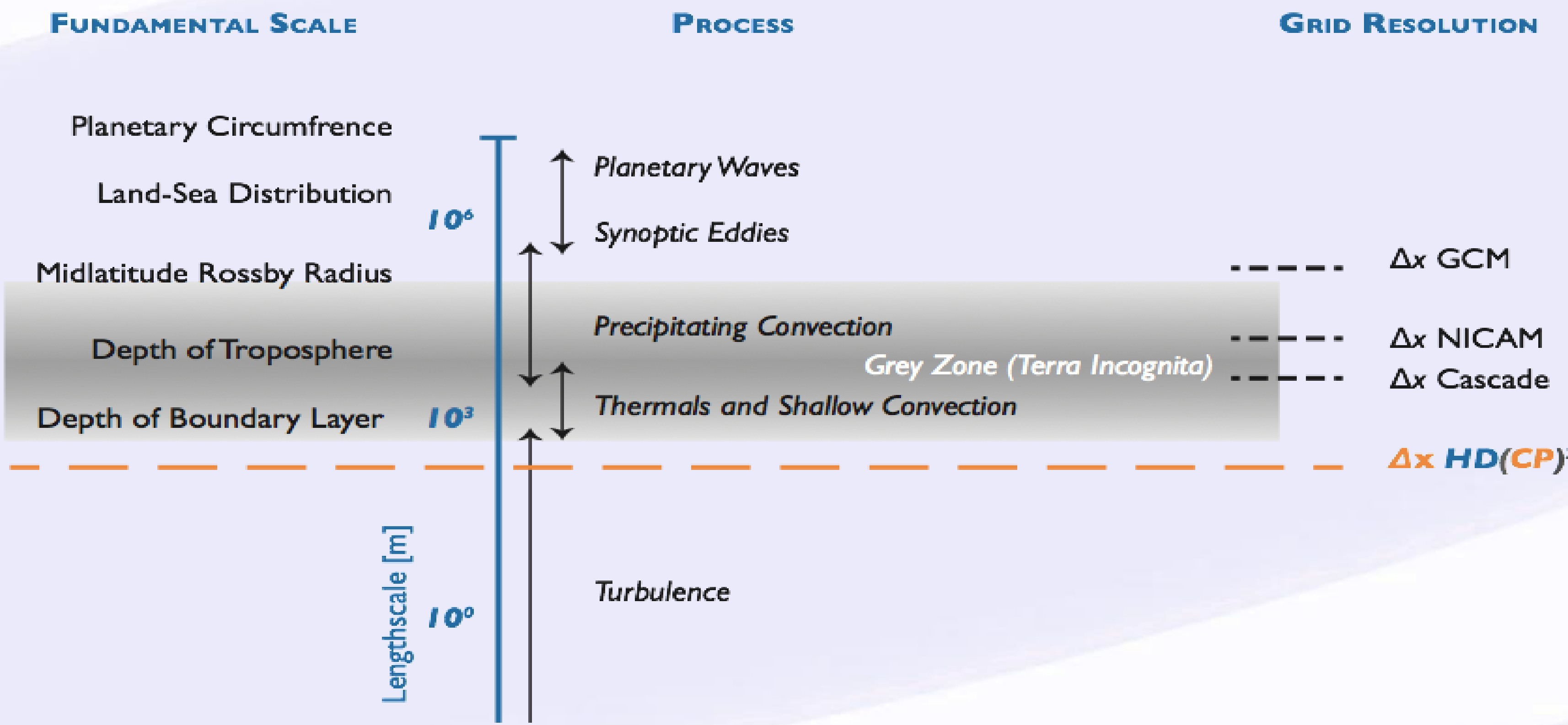
DKRZ

$HD(CP)^2$

High Definition Clouds and Precipitation for Climate Prediction

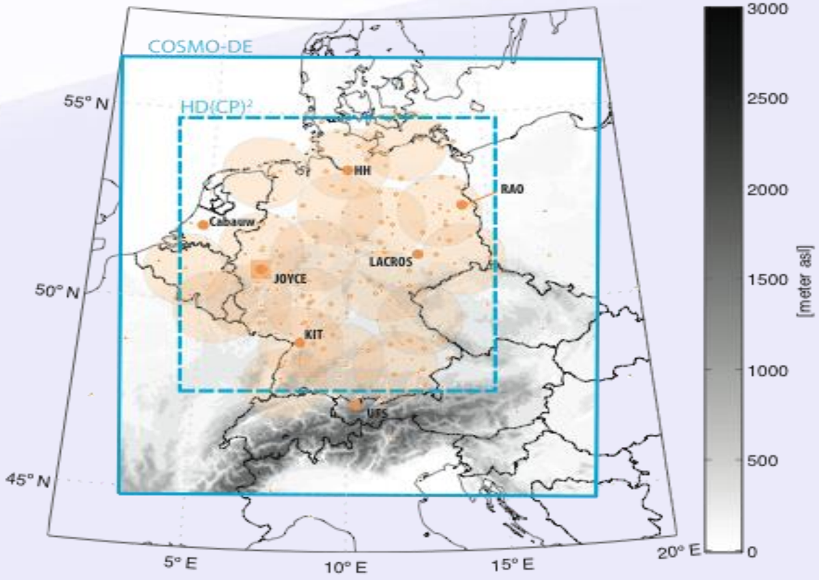
Bjorn Stevens, Joachim Biercamp, Ulrike Burkhardt, Susanne Crewell, Sarah Jones, Andreas Macke, Axel Seifert, Clemens Simmer and Johannes Quaas

The Grey Zone



ICON and a Mature Observational Network

HD(CP)²



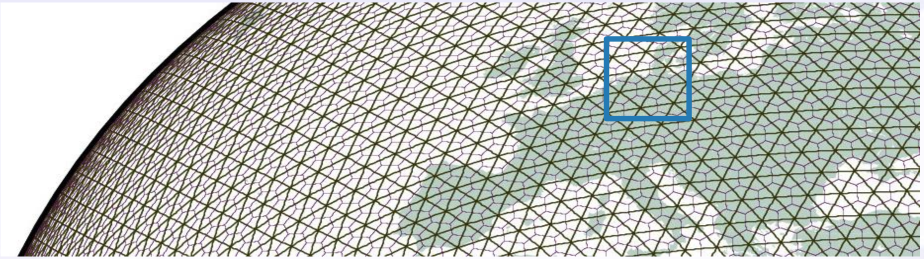
Unprecedented observational network

Across Germany, and Europe more broadly with supersites at CABAUW, RAO (Lindenberg) and other locations that are comparable to the best instrumented sites anywhere in the world.

Project Goal:
coordination and standardization.

Modell-Development based on ICON

Project Goal:
ICON local area over Germany with 100 meter resolution



DKRZ

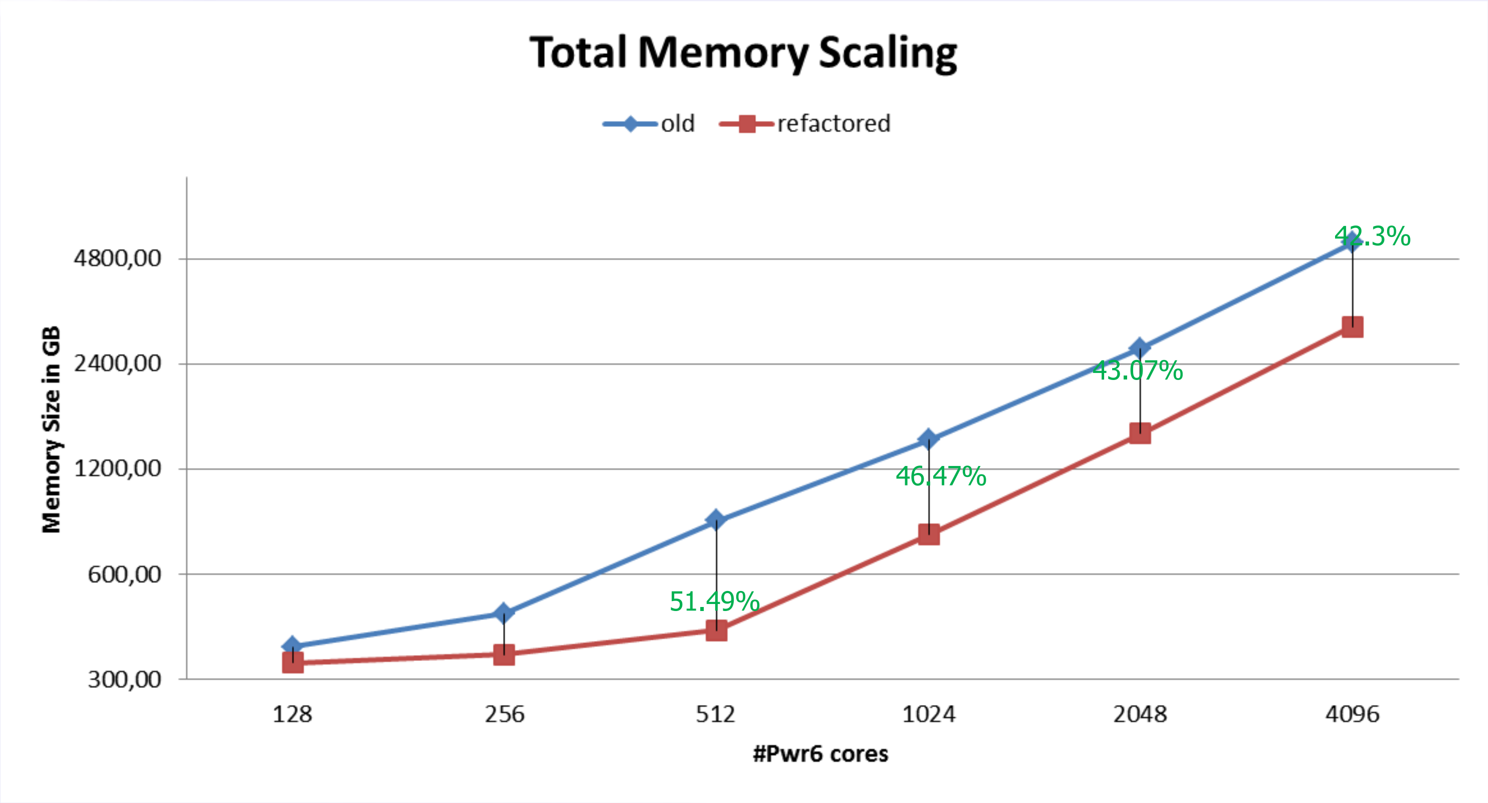
Contact: Moritz Hanke, Thomas Jahns

HD(CP)² led to refacturing of ICON to deal with scalability walls and maintainability

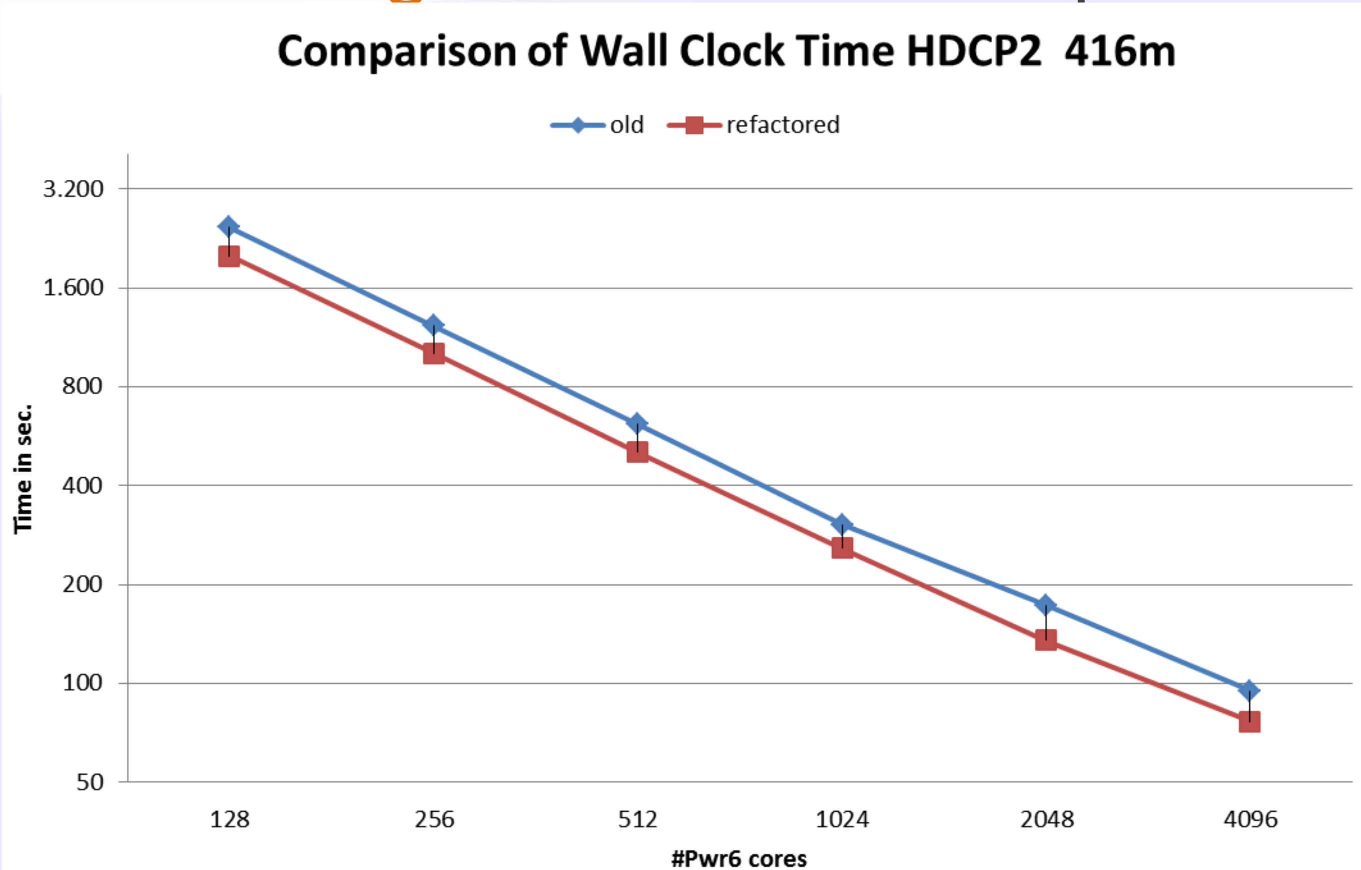
- Memory scaling:
  - Arrays sized to the global number of cells/vertices/edges on every process
    → memory consumption is proportional to number of processes
  - Used in serial code portions to
    - compute decomposition (work in process by implementing distributed algorithm)
    - compute local halo information (fixed rewriting algorithm)
    - store decomposition information (fixed rewriting data structures)
    - read netcdf data; serial read + broadcast (fixed using distributed read + scatter)
    - store gather communication pattern (fixed using two-phase gather algorithm)
    - write output (needs to be fixed)

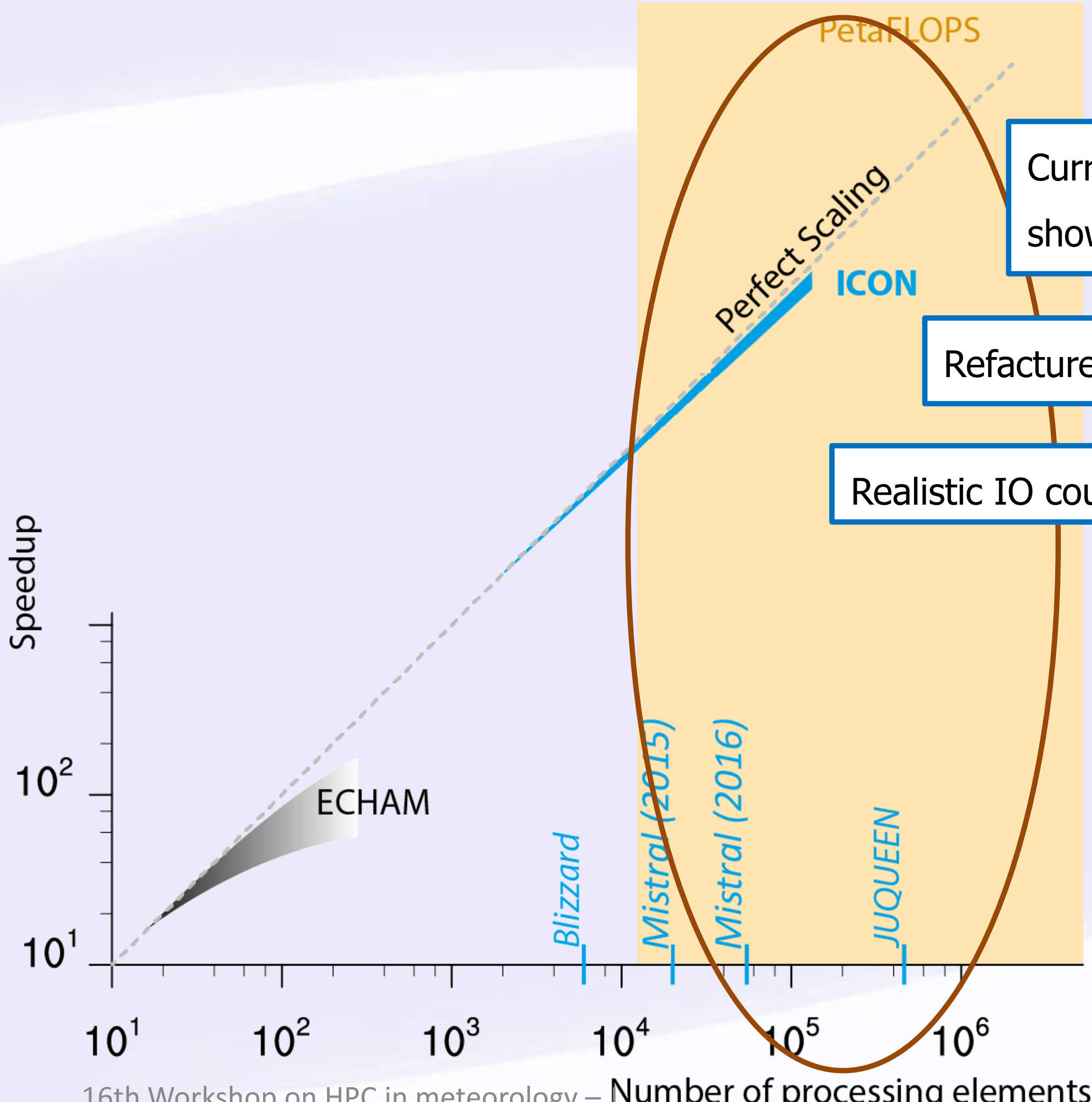
- Code quality:
  - code duplication (removed wherever we found it)
  - single responsibility principle often violated (fix attempted whenever possible...not often)

Refactoring of Domain Decomposition



Refactoring of Domain Decomposition





Current HD(CP)^2 version of ICON shown to scale up to 130,000 cores

Refactured version should scale further

Realistic IO could change the picture

Graph; Bjorn Stevens, internal report HD(CP)2

ORGANISATION

HARDWARE

SOFTWARE

TOOLS

CO-DESIGN

Parallel Output with CDI-PIO (1)

Contact: Thomas Jahns. Irina Fast

Parallel Output with CDI-PIO (1)

Contact: Thomas Jahns. Irina Fast

CDI (Climate Data Interface) is generic I/O interface library abstracting away differences of several file formats relevant in weather and climate research (GRIB1/2, netCDF etc.).

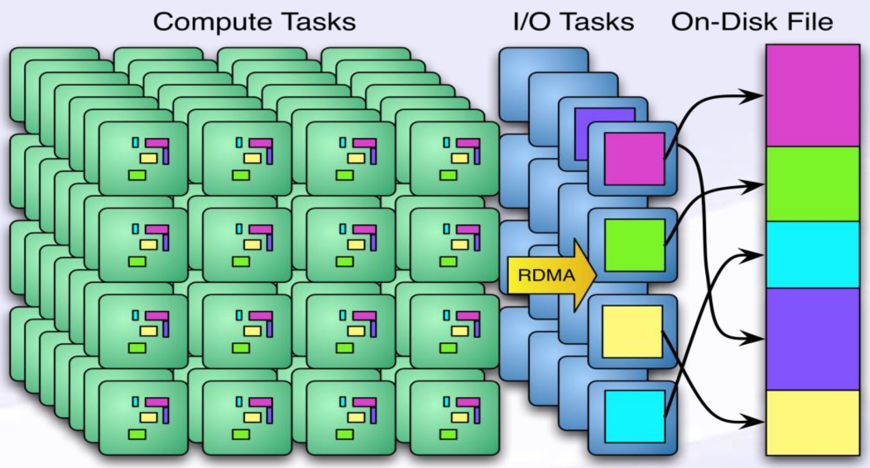
CDI-PIO is the MPI-parallelization and I/O client/server infrastructure of CDI.

Features

- High-throughput/low-latency output of model data supporting different file formats
- Cross-platform support for all relevant HPC systems
- Minimized disturbance of on-going computations via asynchronous RDMA transfer

Current status

- GRIB and netCDF4 working
- Implementation in ECHAM6 and MPI-ESM
- Tests on different hardware architectures and with different MPI implementations
    - IBM Power6 Cluster (IBM PE)
    - Intel Xeon Cluster (MVAPICH2, OpenMPI)

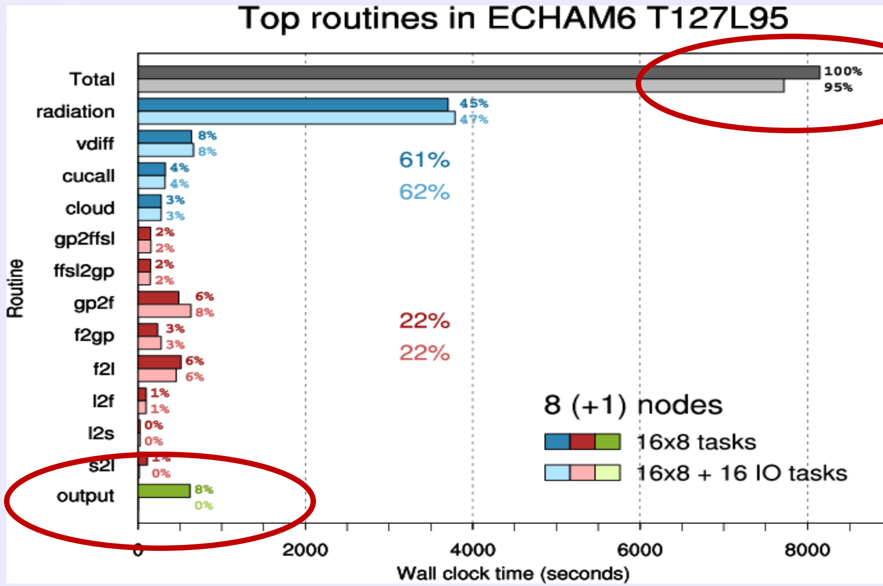


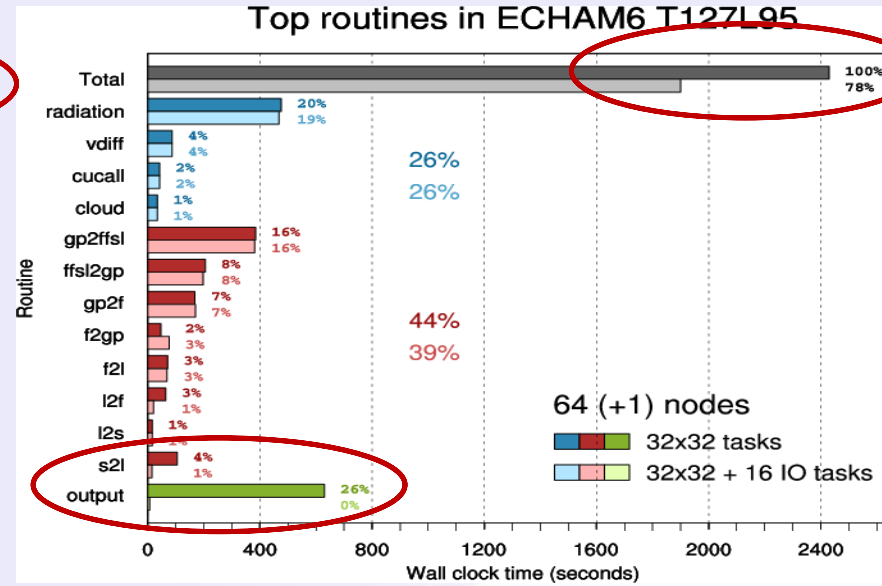
Plans

- better mapping of compute to I/O tasks
- flexible decomposition on compute tasks

Parallel Output with CDI-PIO (2)

Runtime profiles of ECHAM6-HR (T127L95)
on "Thunder": Intel Xeon Linux Cluster, MPI: MVAPICH2





Legend (dark colors: serial output, light colors: output via I/O servers)

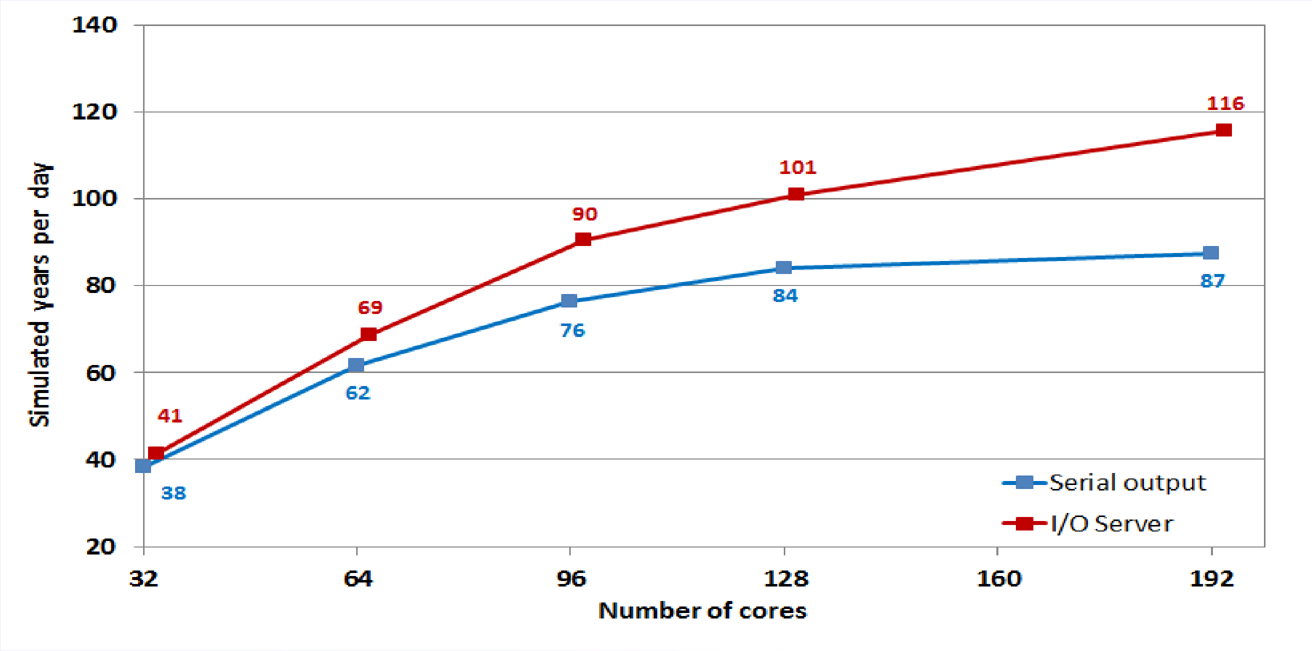
Grey:	*total*	*Red:*	*transpositions*
Blue:	*physics*	*Green:*	*output*

DKRZ

Parallel Output with CDI-PIO (3)

Contact: Thomas Jahns. Irina Fast

Throughput rates for ECHAM6-CRT (T31L31)
on "Blizzard": IBM Power6 Cluster, MPI: IBM PE

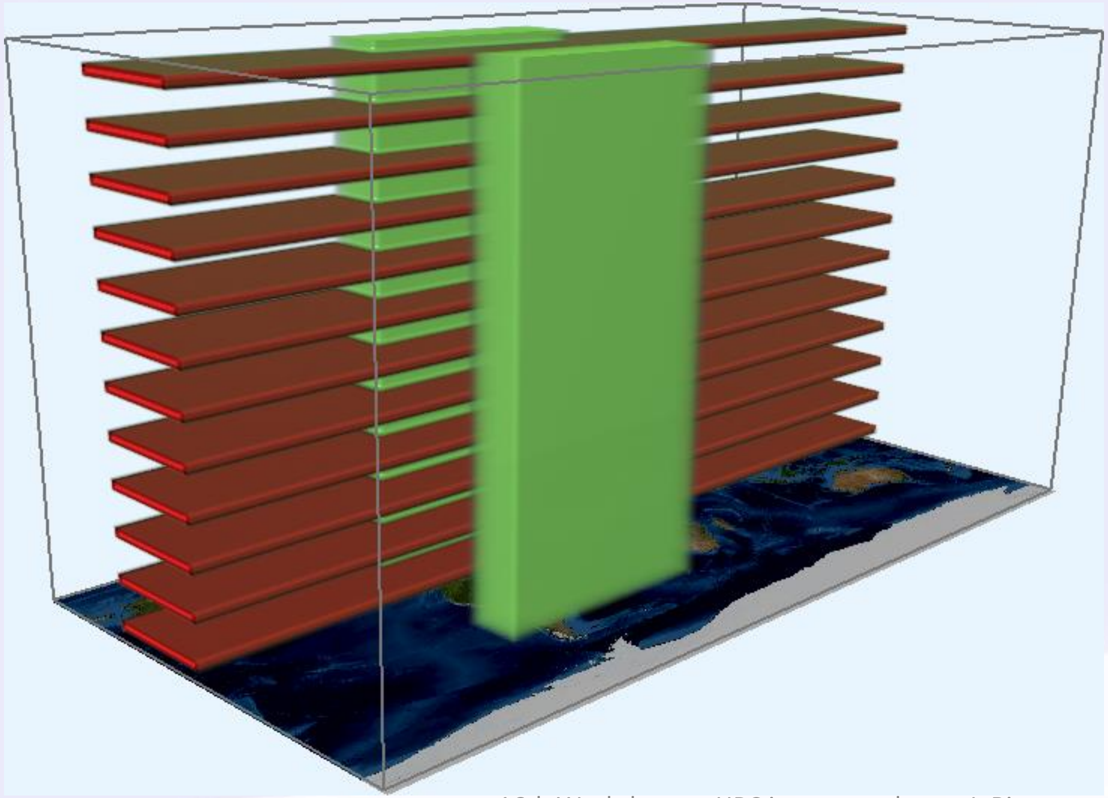


DKRZ

Communication using YAXT (1)

Contact: Jörg Behrens

ECHAM Transposition: From gridpoint (GP) decomposition to Flux Form Semi-Lagrangian (FFSL) decomposition



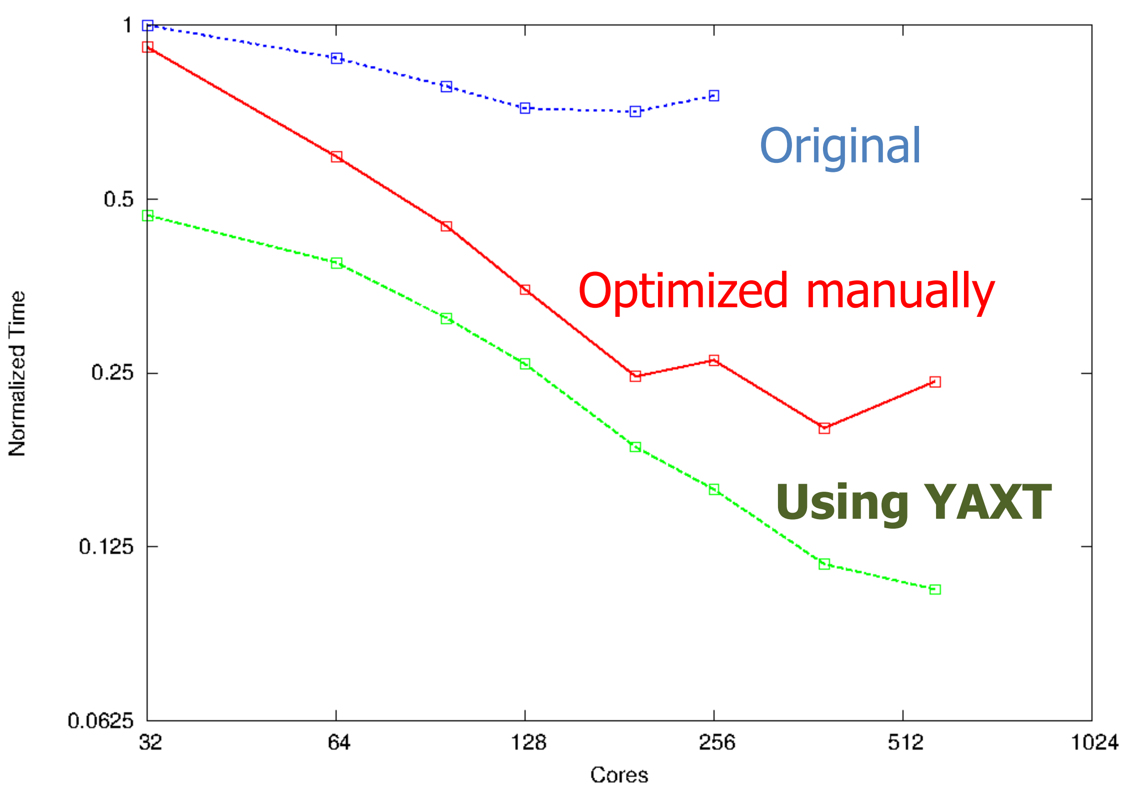
Data distribution of one selected MPI-process:

before (green) and after (red) the gp2ffsl transposition

DKRZ

Communication using YAXT (2)

Contact: Jörg Behrens



ECHAM Transposition gp->ffsl reinvented

T63L47, synchronized measurement

DKRZ

Communication using YAXT (3)

Contact: Jörg Behrens

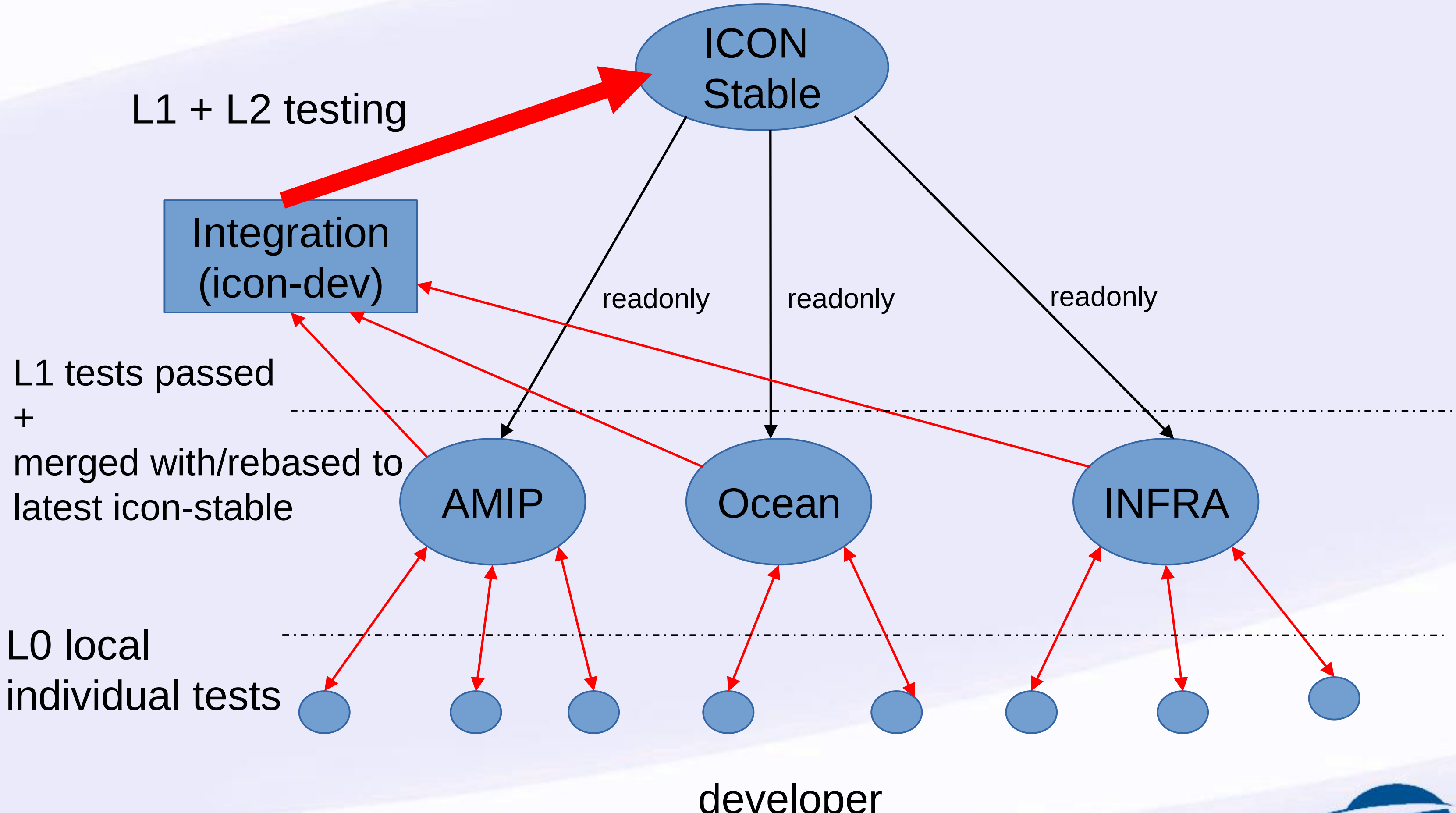
ECHAM Transposition: gp->ffsl

- Reimplementation of gp2ffsl using MPI directly:
  - Effort: weeks
  - Improved scaling
  - Bugs removed after extensive testing

- Reimplementation of gp2ffsl using YAXT:
  - Effort: days
  - Improved scaling even more
  - Bugs removed early (YAXT internal checks

Contact: Panos Adamidis

ICON hierarchical testing



Contact: Panos Adamidis

ICON Hierarchical Testing (2)

Level-1 Tests

- Portability
  - Compilers for which the code must compile :
    - Cray - DWD
    - IBM - DKRZ
    - Intel(Bull) – DKRZ
    - GNU – MPI
    - NAG – MPI
    - PGI - CSCS

- Technical Properties
  - compare result to reference revision
  - MPI processes
  - OpenMP threads
  - OpenACC (once in use)

- Experiments to be used for testing the technical properties
  - AMIP
  - OMIP
  - ICON-LES limited area
  - Joblonowski Wiiliamson with nesting etc.
  - NWP forecast

WORKFLOWS, e.g CYLC

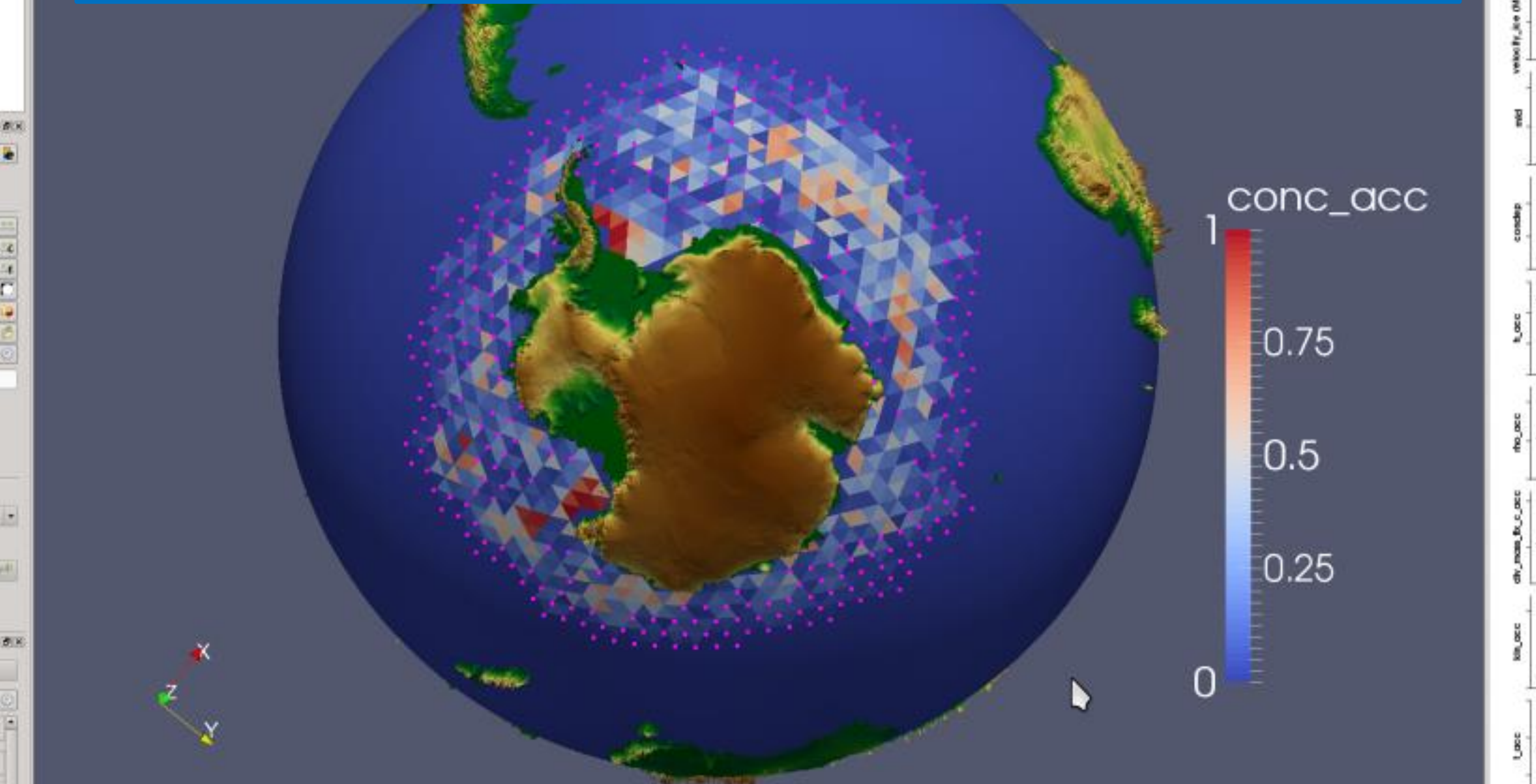
Contact: Kerstin Fieg

- Will be used at DKRZ in collaboration with MPI – M as new runtime environment to design and control experiment workflows
  (e.g. for projects MIKLIP & HD(CP)2)

- Cylc ("*silk*") is a **suite engine** and **meta-scheduler** that specializes in suites of cycling tasks for weather and climate forecasting and related
- http://cylc.github.io/cylc/
- Using Python (=> platform independent)
- enables running of multiple cycles and / or processes in parallel
- Developed by H.J. Oliver (NIWA, NZ) under GNU Licence

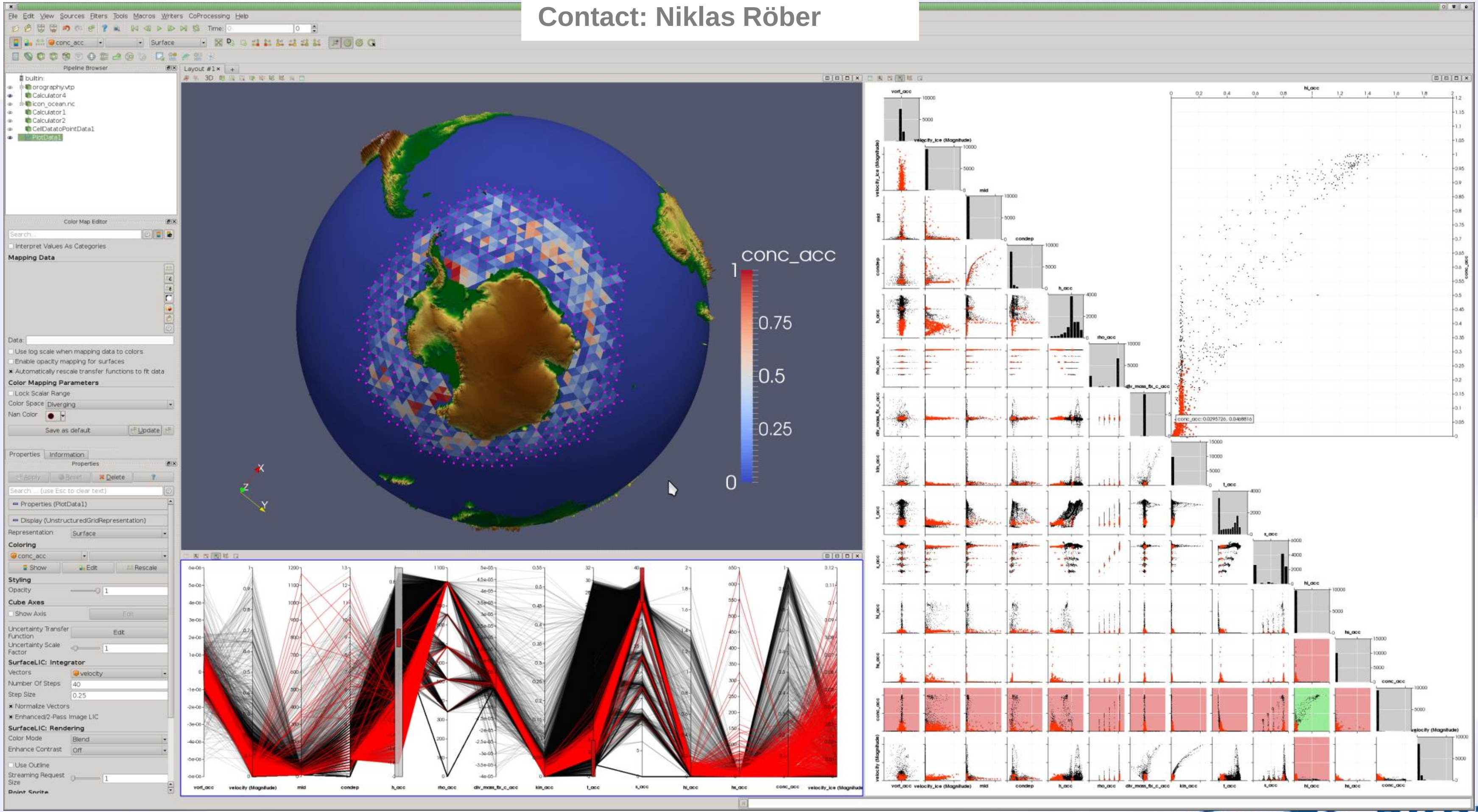
Visualisation of data on the ICON Grid

Contact: Niklas Röber



Visualisation of data on the ICON Grid

Contact: Niklas Röber



ORGANISATION

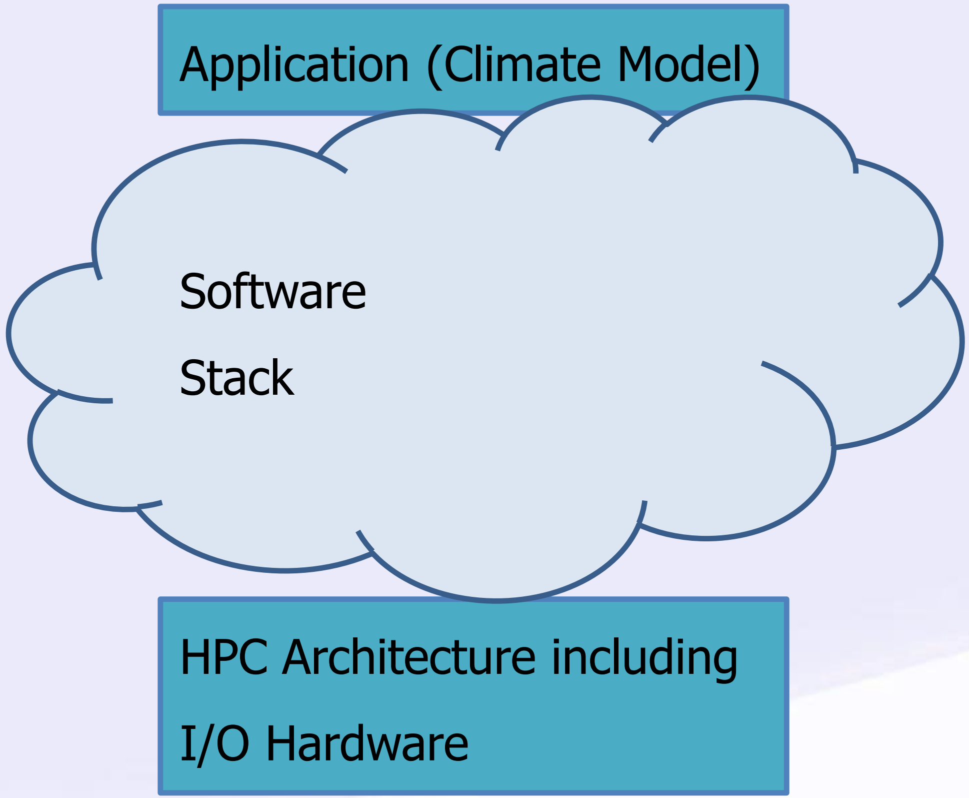
HARDWARE

SOFTWARE

TOOLS

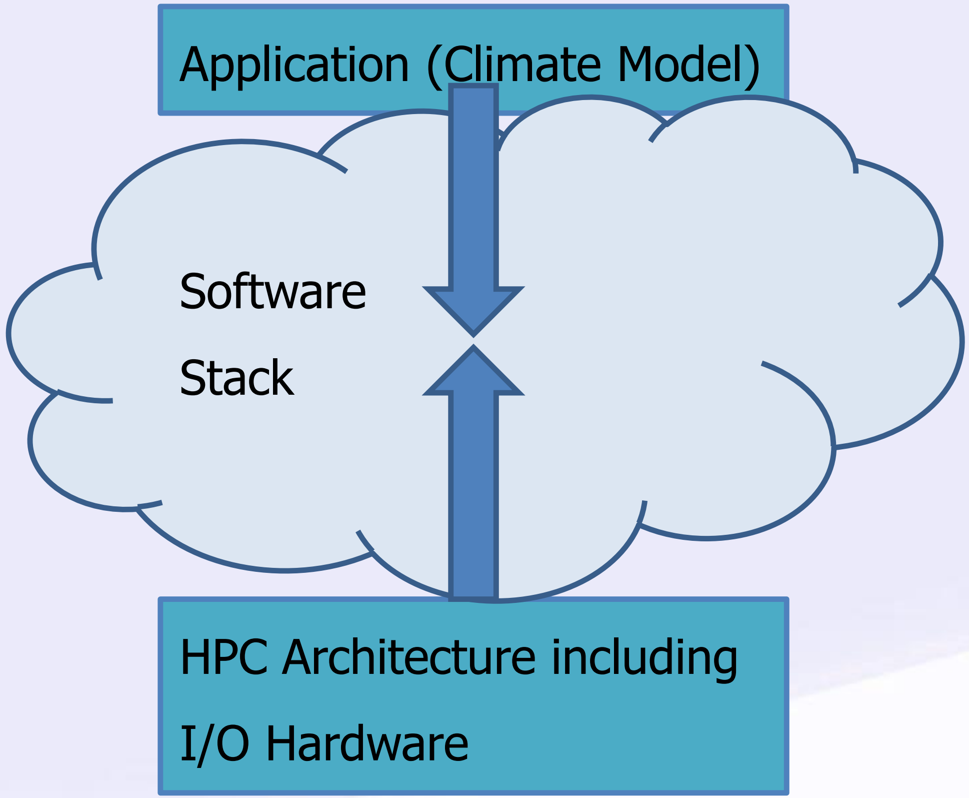
CO-DESIGN (Here: Cooperation with BULL)

Computational Challenges for EXASCALE

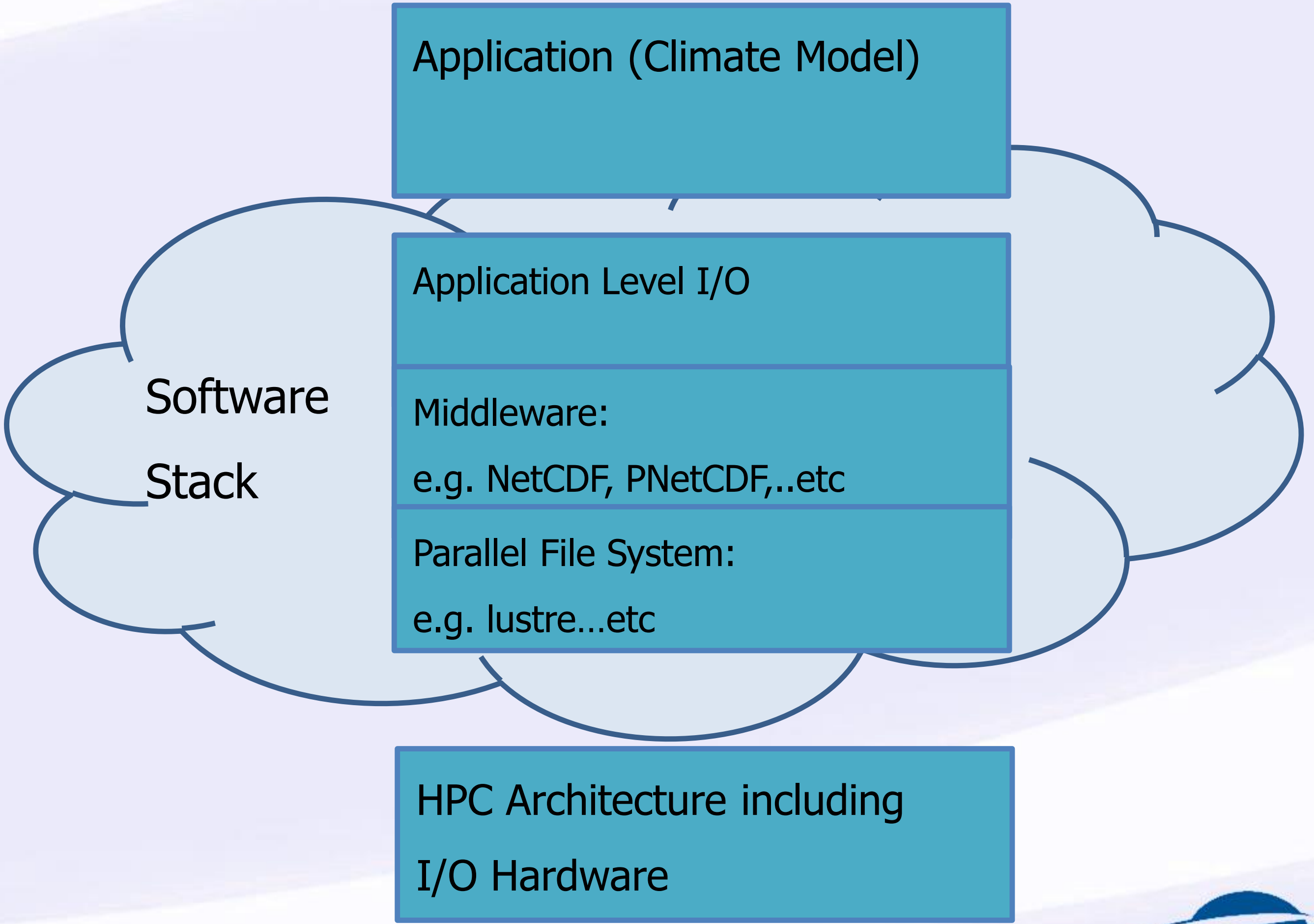


DKRZ

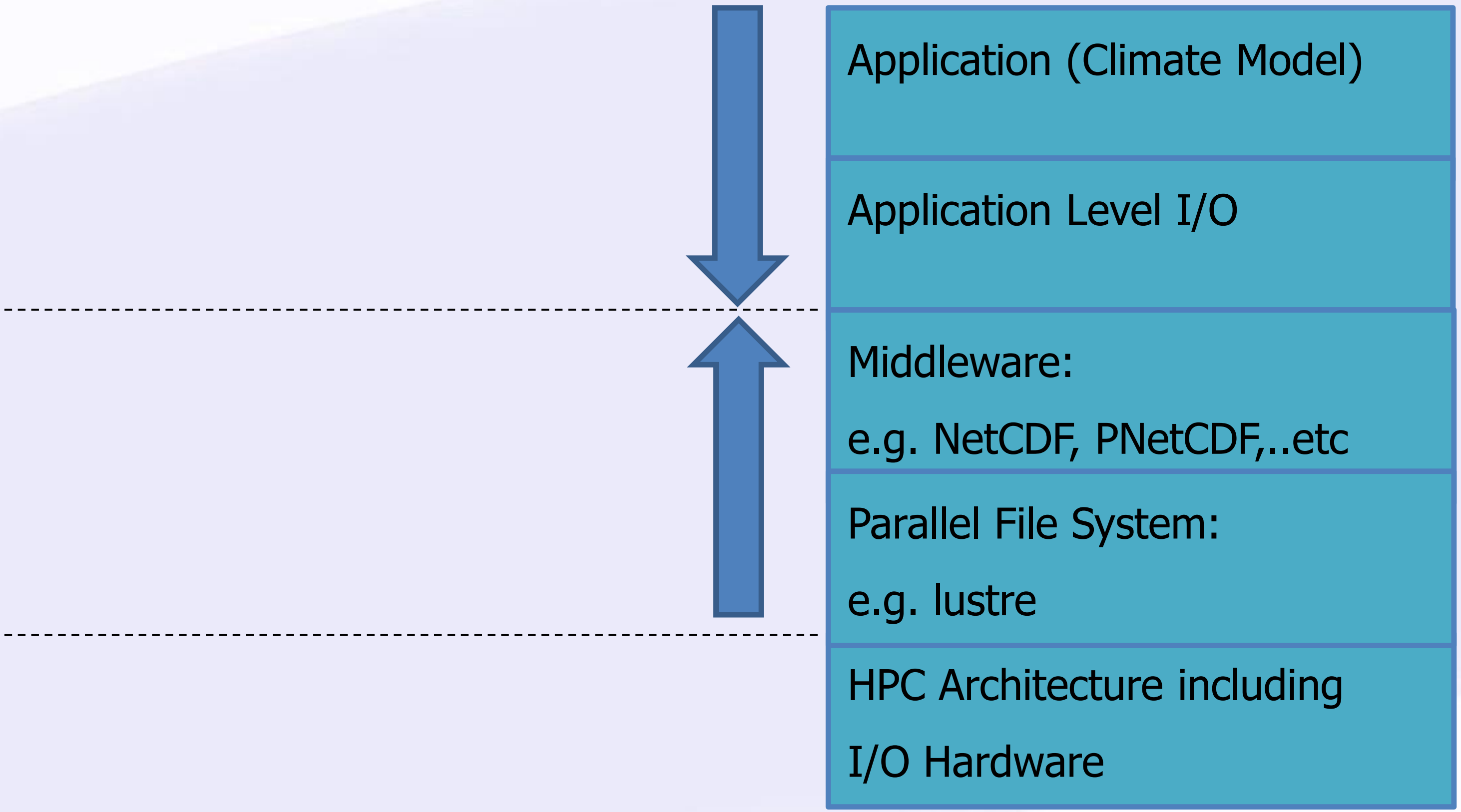
Computational Challenges for EXASCALE



Computational Challenges for EXASCALE



Set-up of BULL DKRZ Co-operation



Set-up of BULL DKRZ Co-operation

The application guy (DKRZ):

Focus: Application Tuning

including I/O

The computer science guy (DKRZ/UNI HH):

Focus: Parallel File System

& Middleware Tuning

The guys from the harware vendor (Bull):

Focus: EXASCALE Systems

Application (Climate Model)
Application Level I/O
Middleware: e.g. NetCDF, PNetCDF,..etc
Parallel File System: e.g. lustre
HPC Architecture including I/O Hardware

Outcome of Co-Design

New algorithms and data structures optimized in terms of computation, communication and I/O

Analysis and better tuning for climate models

Application (Climate Model)

Application Level I/O

Middleware:
e.g. NetCDF, PNetCDF,..etc

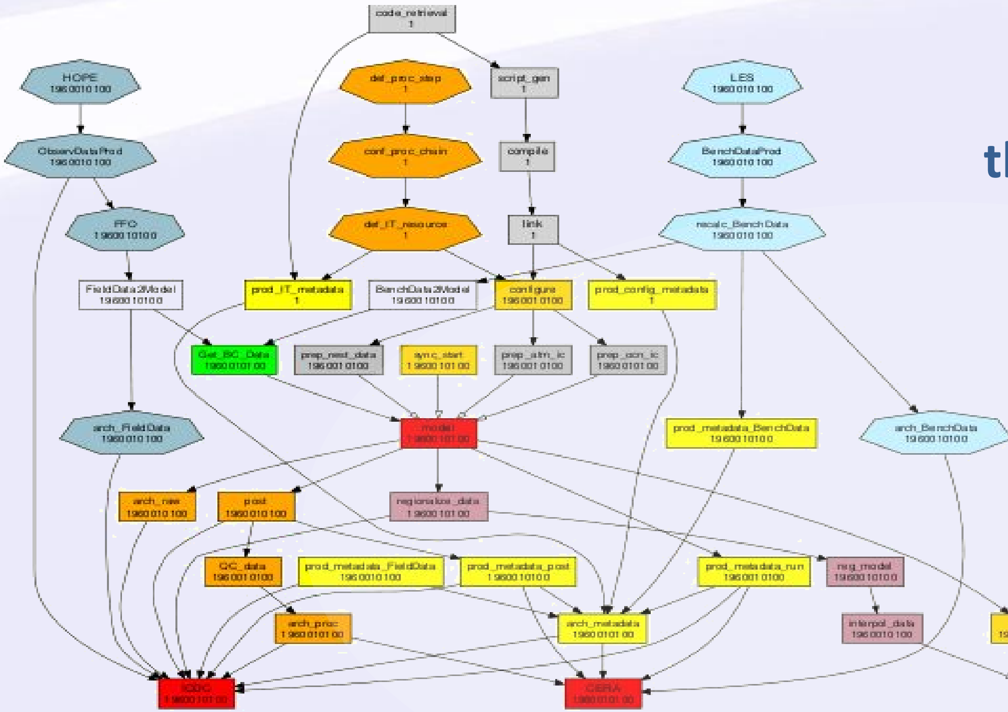
Parallel File System:
e.g. lustre

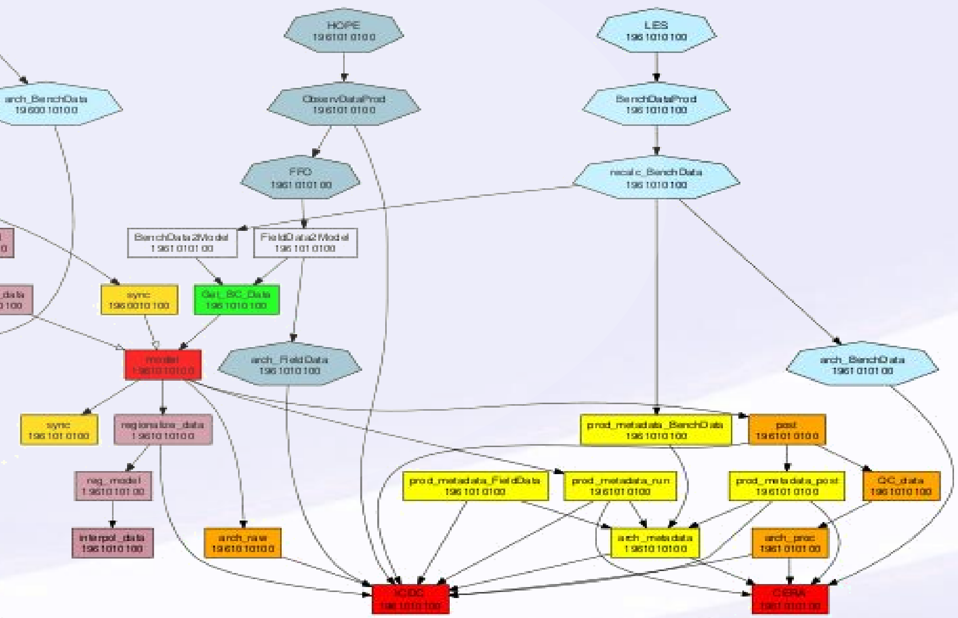
HPC Architecture including
I/O Hardware

DKRZ

Conclusion:

DKRZ tries hard to support the end-to-end workflow of climate modeling





Thank you for staying