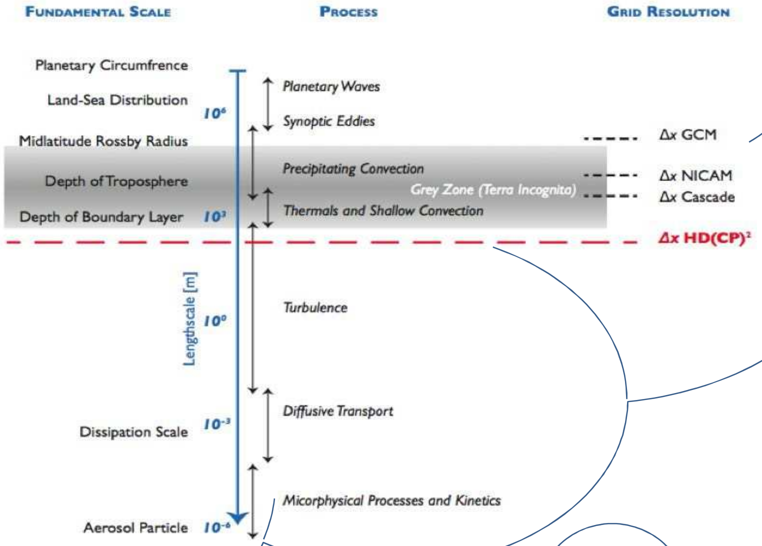
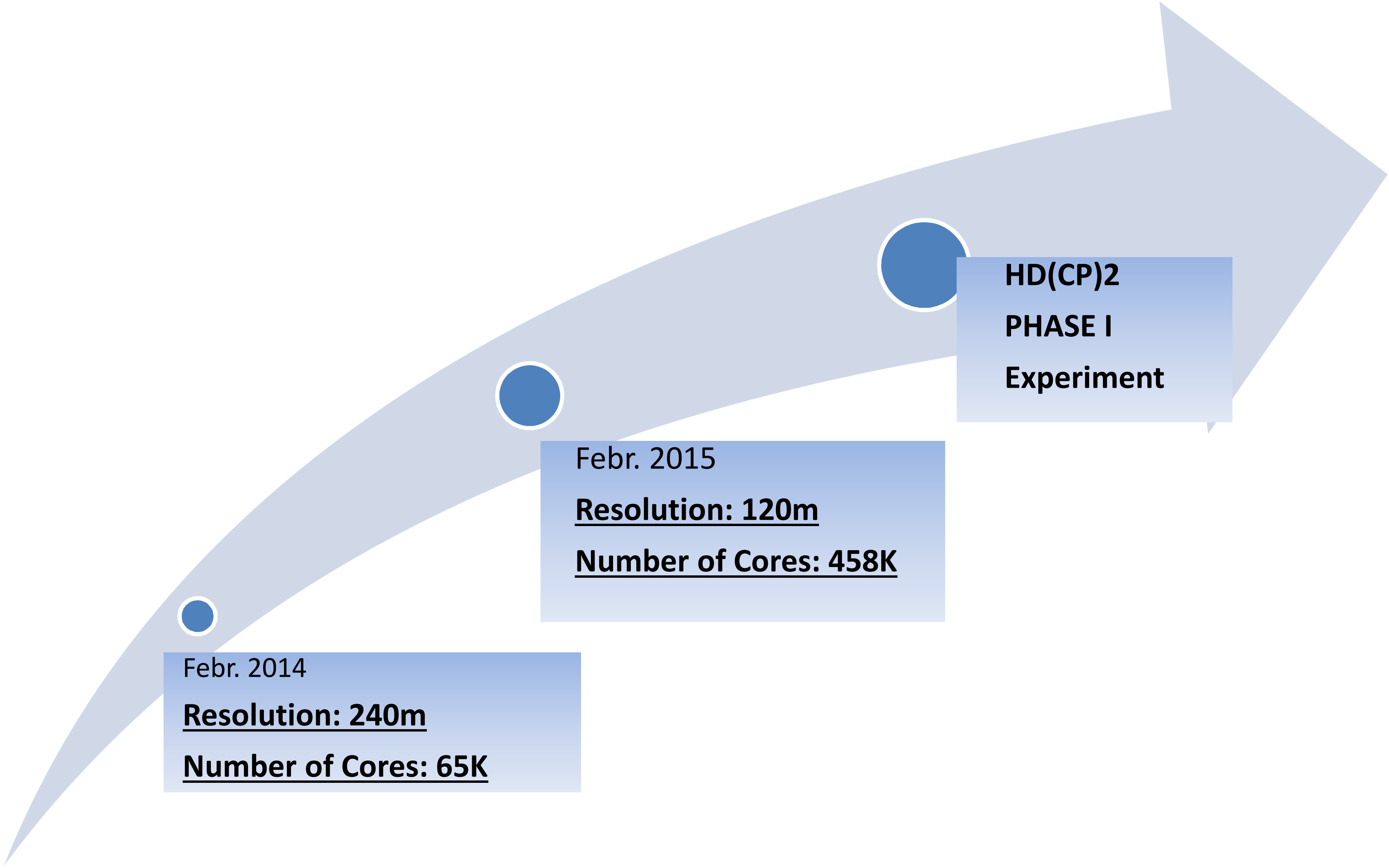
HPC Aspects of the ICON model and high-res simulations

<Panos Adamidis>
Deutsches Klimarechenzentrum (DKRZ)



Fundamental Scale
Process
Grid Resolution
Planetary Circumfrence
Land-Sea Distribution
Midlatitude Rossby Radius
Depth of Troposphere
Depth of Boundary Layer
Dissipation Scale
Aerosol Particle
10^6
10^3
10^0
10^{-3}
10^{-6}
Lengthscale [m]
Planetary Waves
Synoptic Eddies
Precipitating Convection
Grey Zone (Terra Incognita)
Thermals and Shallow Convection
Turbulence
Diffusive Transport
Micorphysical Processes and Kinetics
Δx GCM
Δx NICAM
Δx Cascade
Δx HD(CP)^2

ICON en route to Extreme Scale Computing



Febr. 2014

Resolution: 240m

Number of Cores: 65K

Febr. 2015

Resolution: 120m

Number of Cores: 458K

HD(CP)2

PHASE I

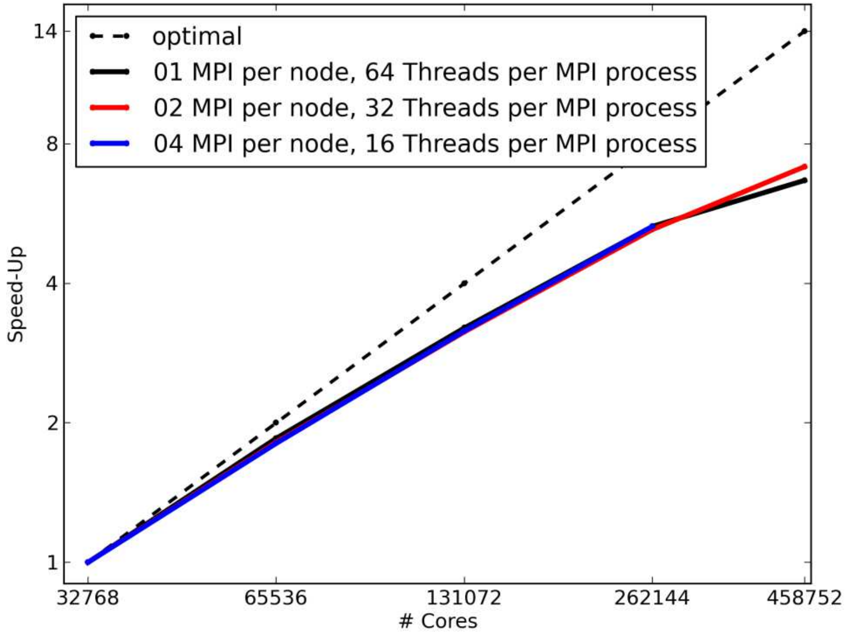
Experiment

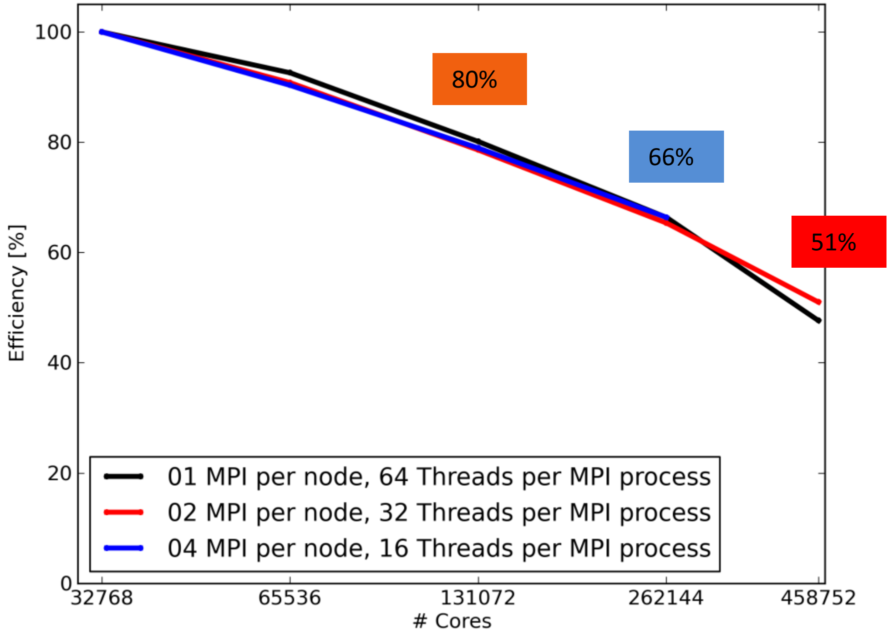
High-scale problems of ICON

➢ Parallelize and Distribute Everything

- Global arrays used in serial code portions to
    - compute decomposition (fixed by using distributed algorithm)
    - compute local halo information (fixed by rewriting algorithm)
    - generate local grid partition (fixed by using distributed data structures; based on shared memory)
    - store decomposition information (fixed by rewriting data structures)
    - read netcdf data; serial read + broadcast (fixed by using distributed read + scatter)
    - store gather communication pattern (fixed by using two-phase gather algorithm)
    - write output (needs to be fixed)

ICON on massive parallel system





ICON is Member of the High-Q Club

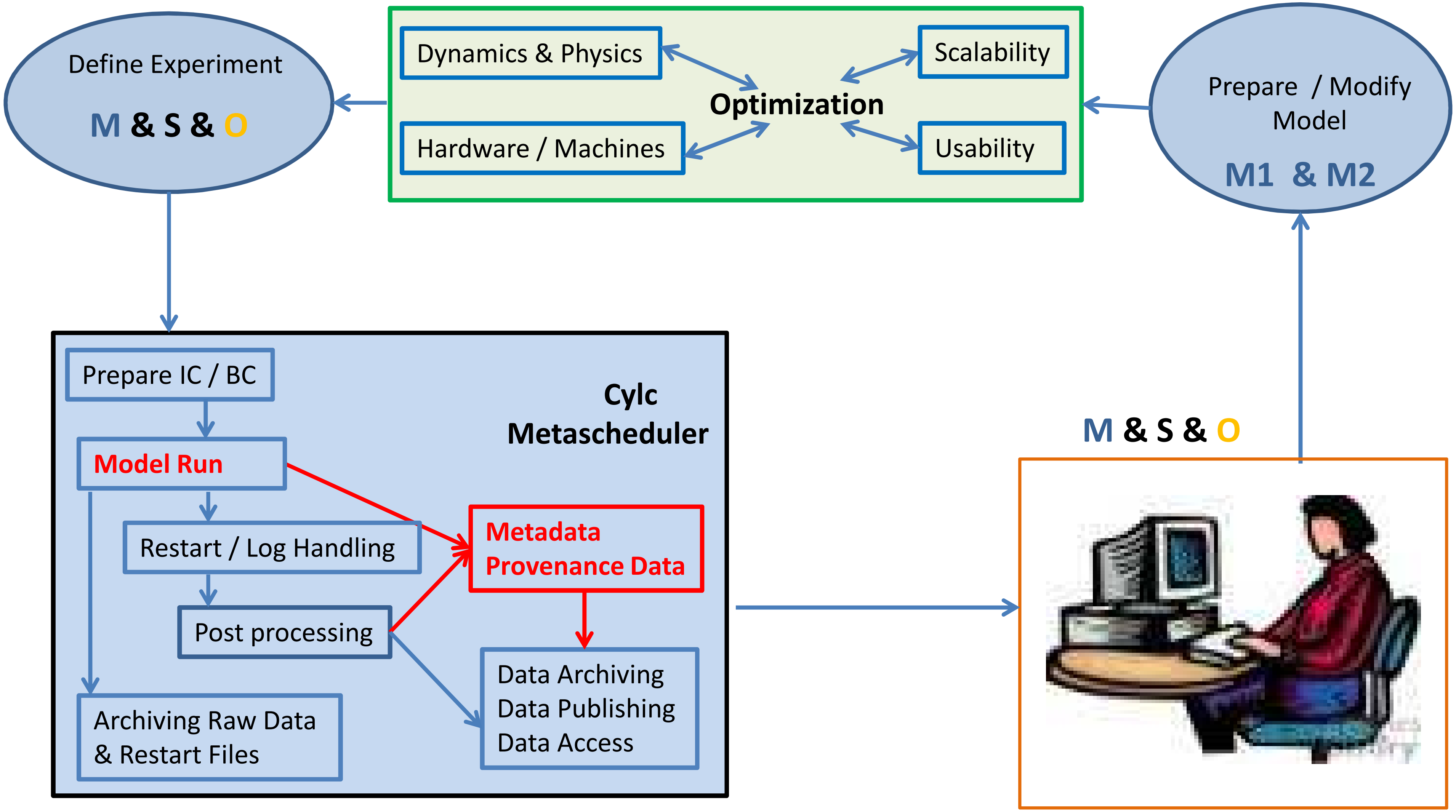
Member • Member • Member • Member • Member • Member • Member • Member •
HiQ
2015

- Highest Scaling Codes on JUQUEEN
- Codes that can utilize the entire BlueGene/Q system

YAXT – Yet Another eXchange Tool

- Library on top of MPI
- Simple to use:
  - Evaluates complete decomposition description
  - No explicit message passing visible
- Reducing latency problem:
  - Climate models & high scaling: many small messages
  - Latency: constant overhead
  - Aggregation reduces communication overhead

Time-to-Solution



Define Experiment

M & S & O

Dynamics & Physics

Hardware / Machines

Optimization

Scalability

Usability

Prepare / Modify Model

M1 & M2

Prepare IC / BC

Model Run

Restart / Log Handling

Post processing

Archiving Raw Data & Restart Files

Cylc Metascheduler

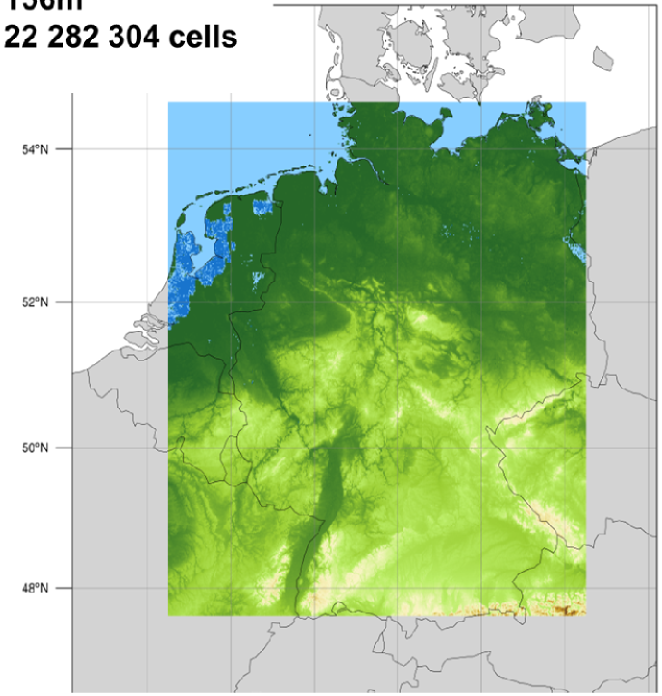
Metadata Provenance Data

Data Archiving
Data Publishing
Data Access

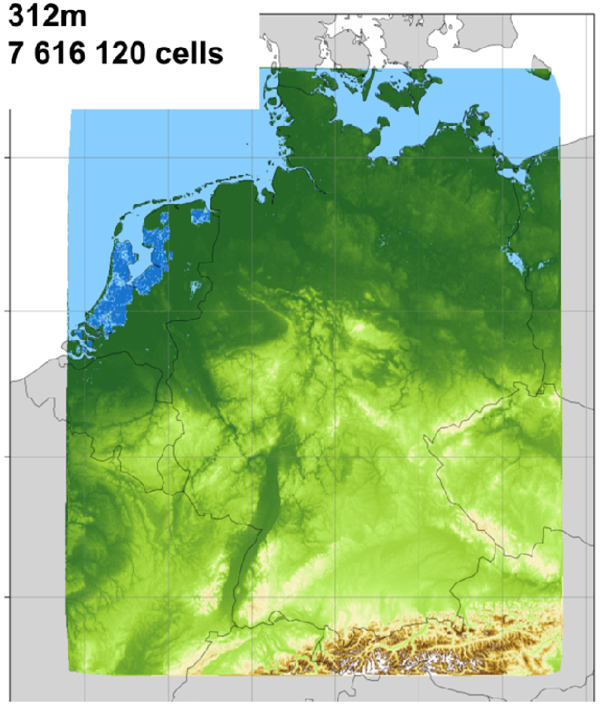
M & S & O

HD(CP)^2 Phase-I Final Experiment

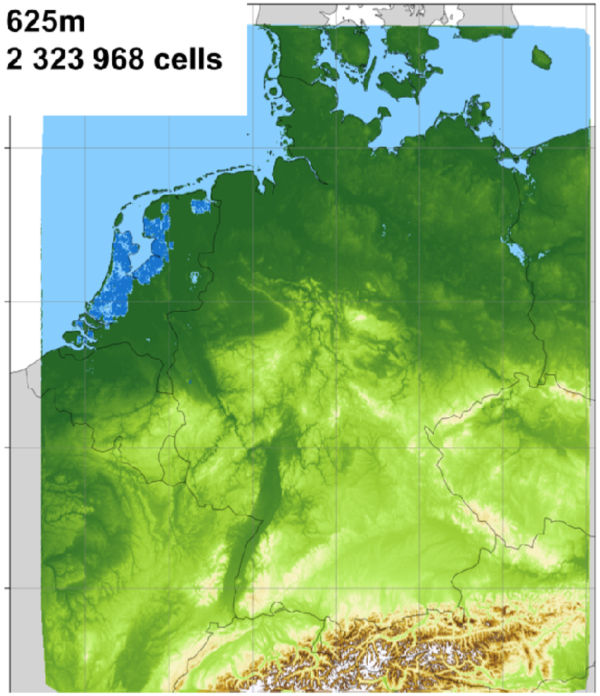
156m
22 282 304 cells



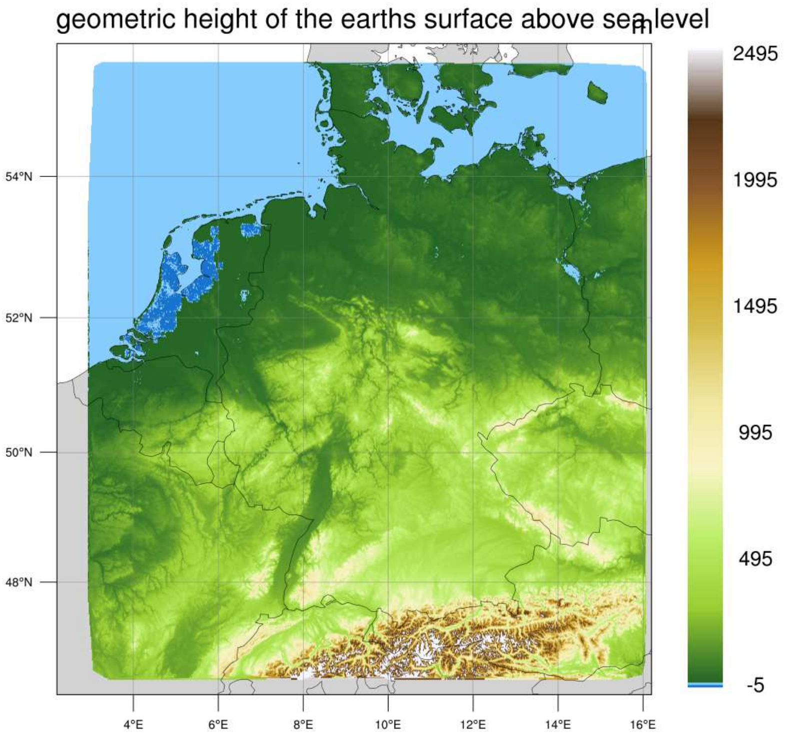
312m
7 616 120 cells

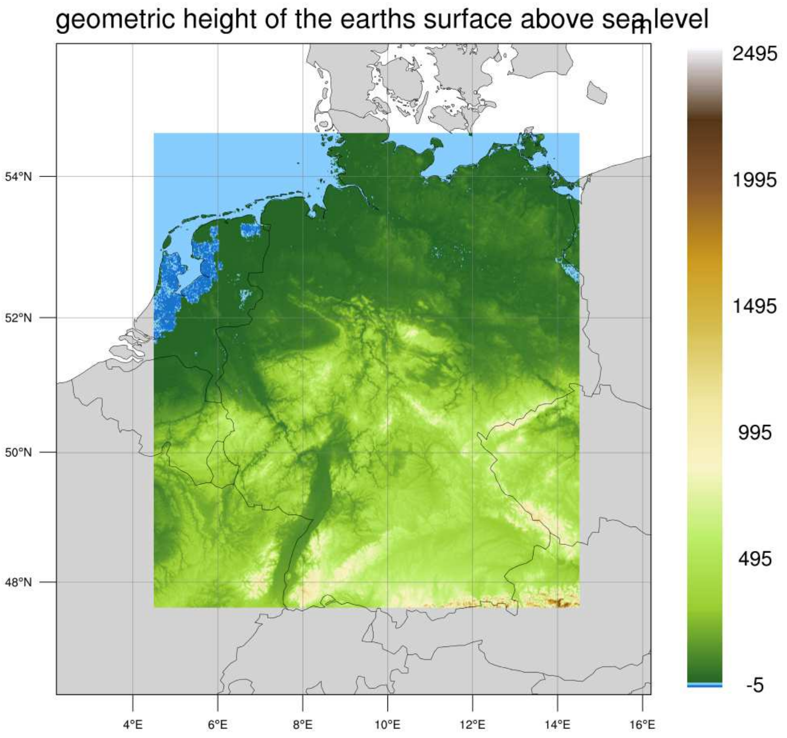


625m
2 323 968 cells



HD(CP)^2 Phase-I Final Experiment

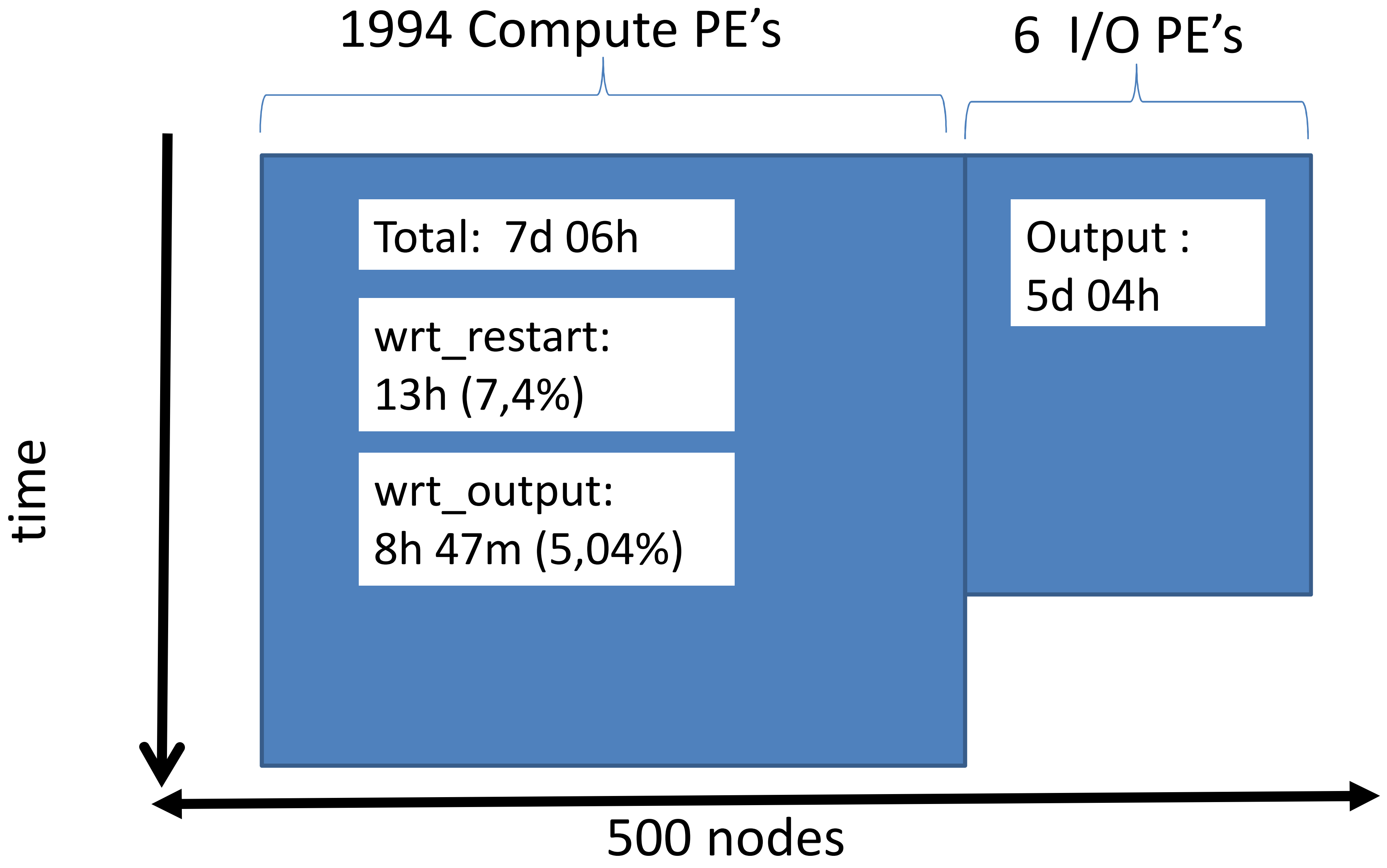

geometric height of the earths surface above sea level


geometric height of the earths surface above sea level

HD(CP)^2 Phase-I Final Experiment

- 3 Domains (625m, 312m, 156m)
- Output of 169 variables (2D/3D) at different intervals (9s, 10s,5min,15min,30min,1hour)
- 1 model day on 500 mistral nodes
  - with 4 MPI Processes x 6 OpenMP threads per node
  - Wallclock : 7days 6hours
  - Total size of Data: 48 TB

ICON-Parallel Asynchronous Output



1994 Compute PE's

6 I/O PE's

Total: 7d 06h

wrt_restart:
13h (7,4%)

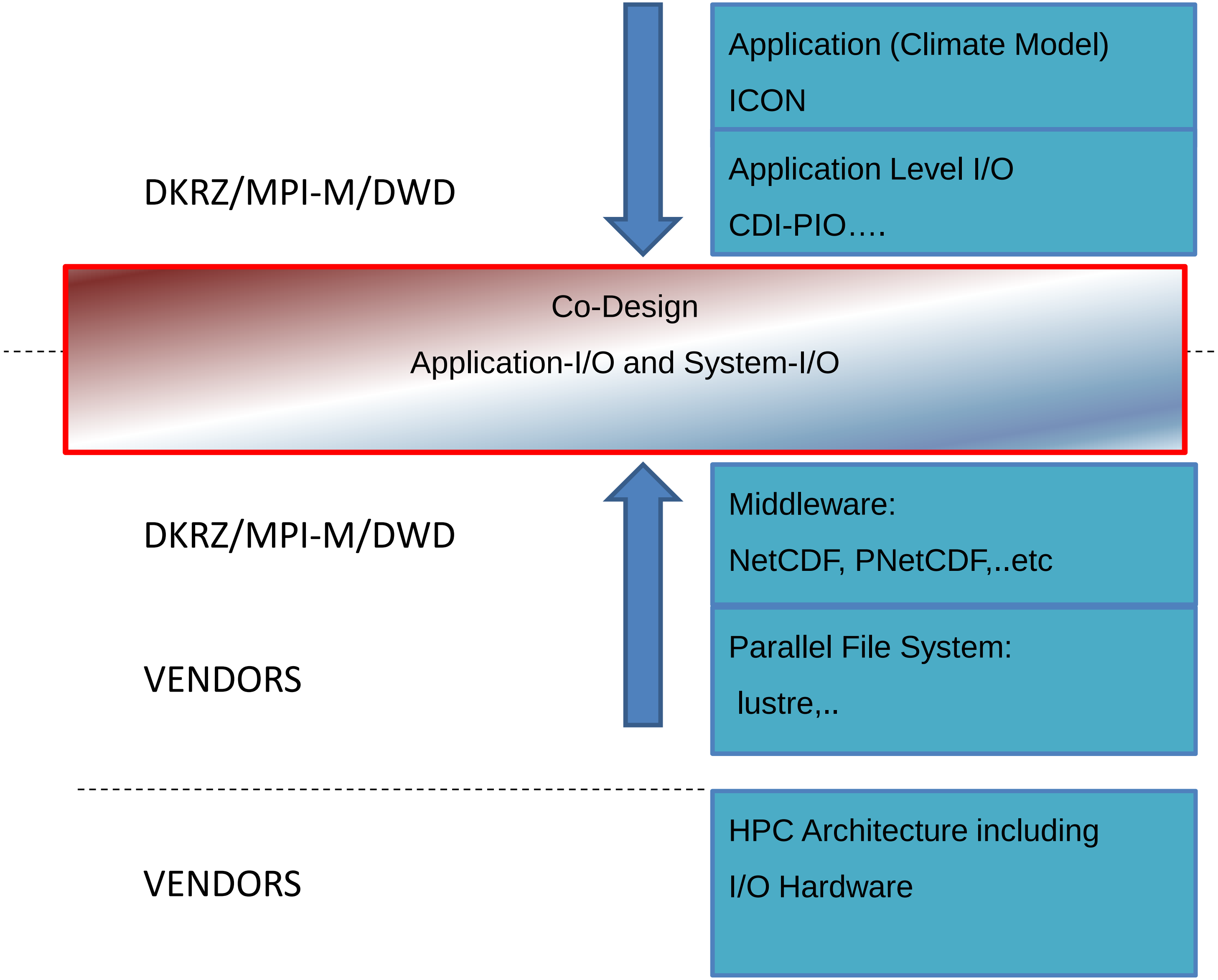
wrt_output:
8h 47m (5,04%)

Output :
5d 04h

time

500 nodes

Co-Design : I/O Now and in the Future



DKRZ/MPI-M/DWD

Application (Climate Model)

ICON

Application Level I/O

CDI-PIO….

Co-Design

Application-I/O and System-I/O

DKRZ/MPI-M/DWD

Middleware:

NetCDF, PNetCDF,..etc

VENDORS

Parallel File System:

lustre,..

VENDORS

HPC Architecture including

I/O Hardware

Requirements for ICON output

The ICON user base requires the following for any suitable Output solution:

1. GRIB and GRIB2 formats must be supported for NWP and climate
2. netCDF output formats are a necessity to also handle restart I/O and highly desirable for climate diagnostics
3. Output data and metadata must be handled in a completely parallelized/decomposed fashion such that no I/O server process violates the "no global size data can be handled by any process" requirement resulting from the scale at which HDCP2 is intended to work
4. The ICON decomposition must be addressable
5. Also flexible control of output should follow the selection pattern already implemented in mo_name_list_output
6. The library must be available in a form that is usable from ICON
7. High throughput should be possible on all relevant platforms

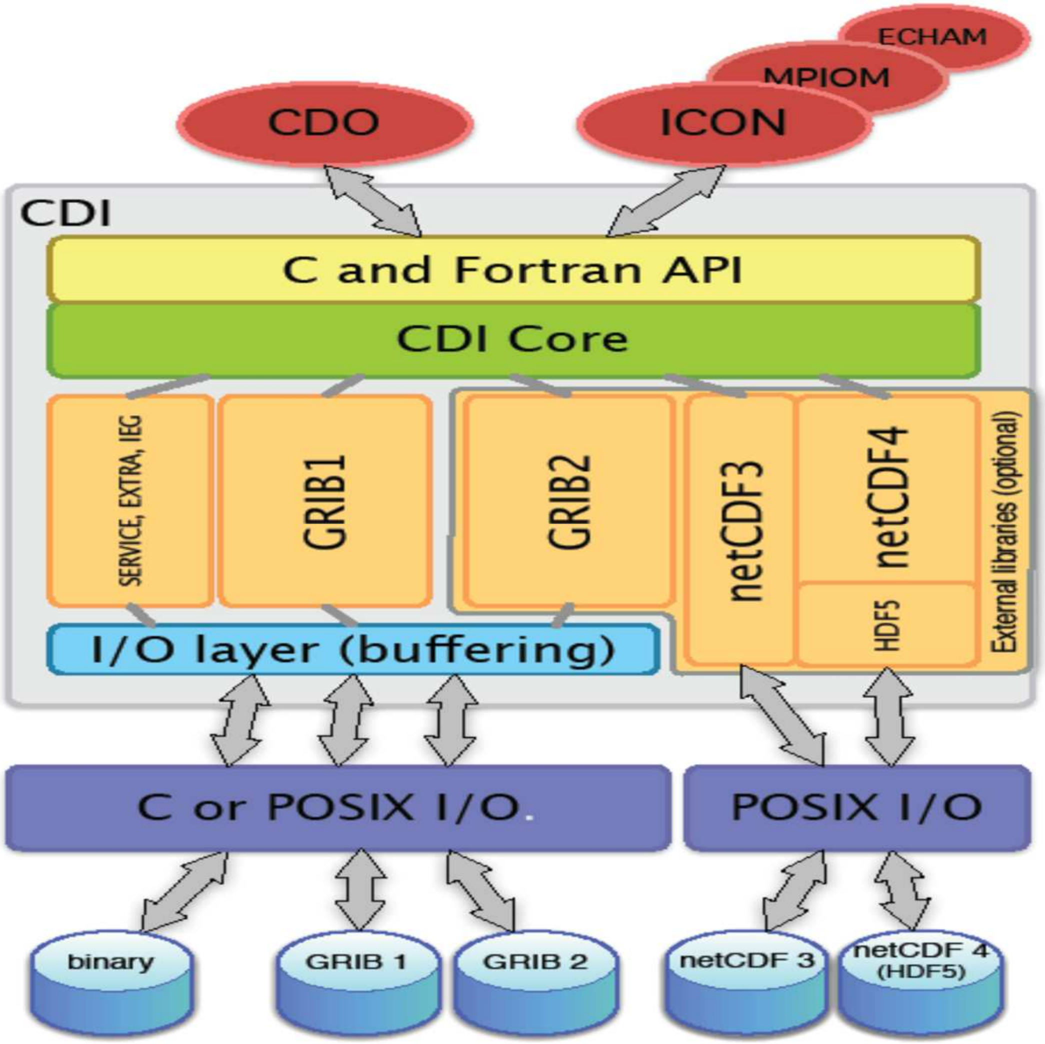
Status of different I/O libraries

The following parallel I/O solutions are available but none yet fully addresses all requirements:

- NCAR PIO does not meet requirement 1. and only meets 4. when further code and data in ICON is added
- XIOS does not meet any of 1. and 5.
- UM Parallel output does meet none of 6. and 2.
- CFIO does not meet 1. and 4. with the same problems as NCAR PIO
- ADIOS neither fulfills 1. nor 2. unless a way from its BP data format to a well-established format is devised and additionally creates voluminous intermediate data that needs to be post-processed extensively, how well 4. is addressed remains unclear
- CDI-PIO is currently not fully able to handle 3. and needs further tuning for 7.

DKRZ

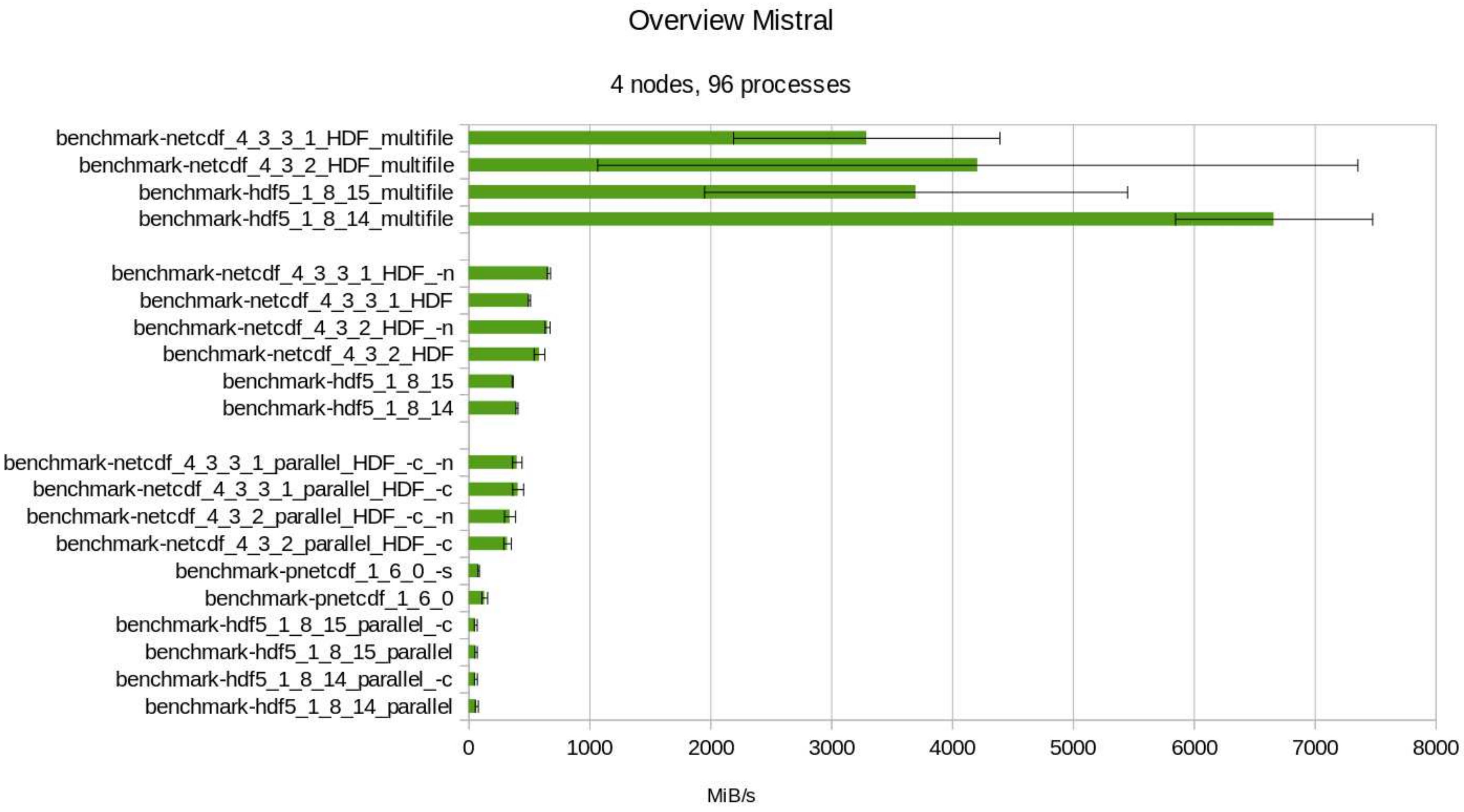
CDI-PIO


ECHAM
MPIOM
CDO
ICON
CDI
C and Fortran API
CDI Core
SERVICE, EXTRA, IEG
GRIB1
GRIB2
netCDF3
netCDF4
HDF5
External libraries (optional)
I/O layer (buffering)
C or POSIX I/O.
POSIX I/O
binary
GRIB 1
GRIB 2
netCDF 3
netCDF 4
(HDF5)

CDI-PIO in MPI-ESM1

- ECHAM - T127L95 (192x384x95)
  MPIOM - TP04L40
- On mistral 1 model year and output every 6h
  - without CDI-PIO on 108 nodes takes 8000 sec
    => 864000 node hours
  - with CDI-PIO on 109 nodes 4800sec
    => 523200 node hours
- Has been tested with Intel MPI and BullxMPI

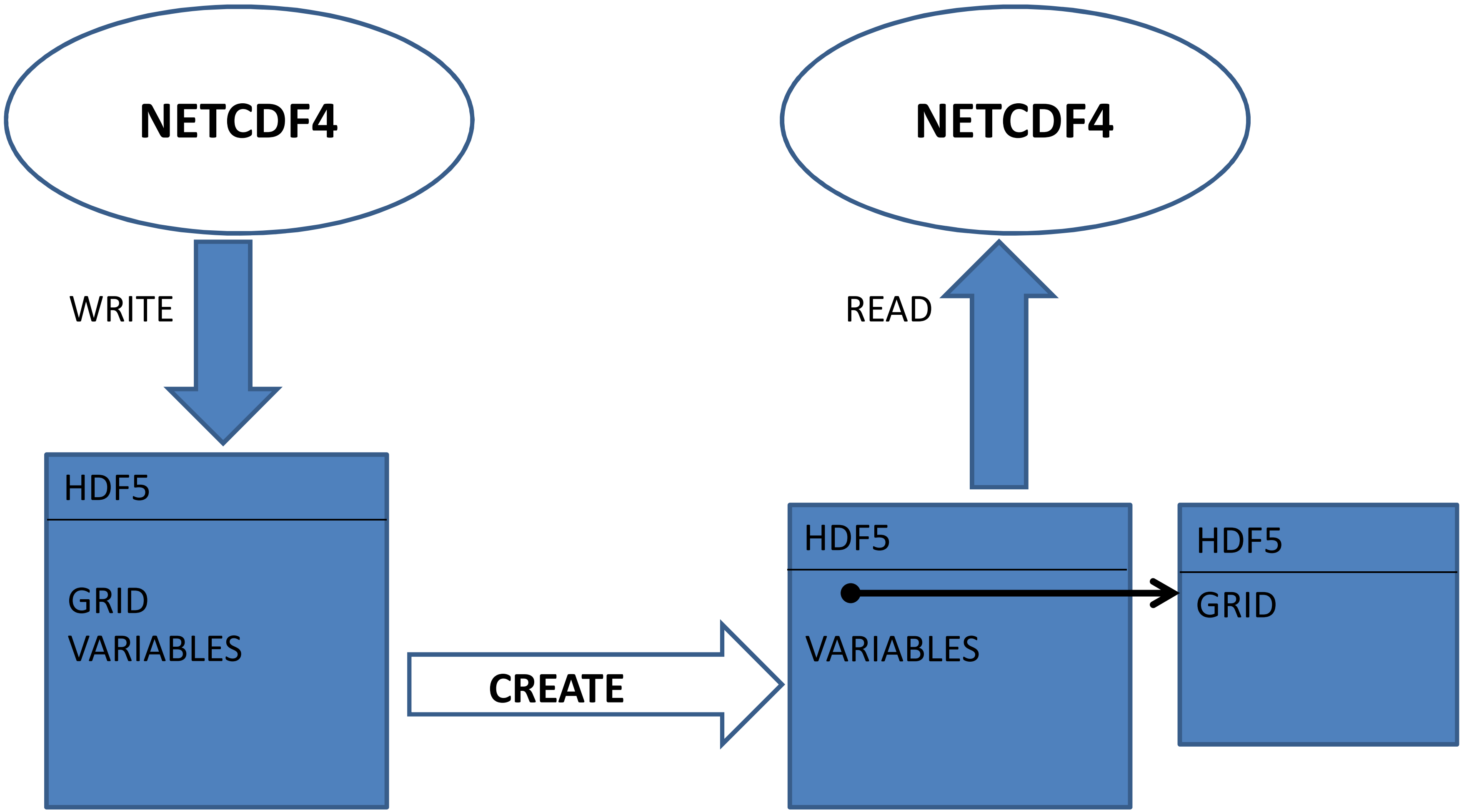
Output Benchmarks



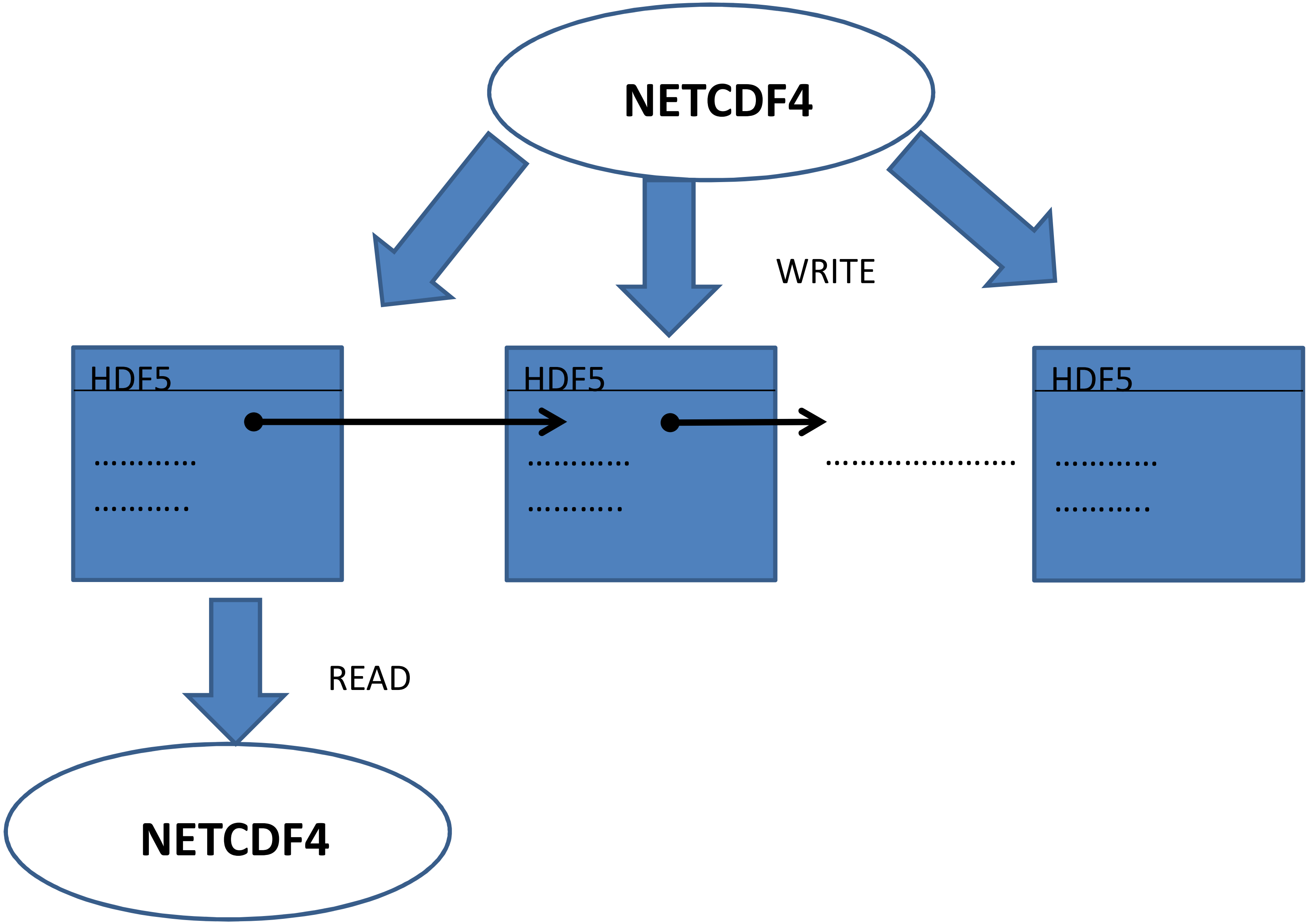
Exploring “multifile” Approach

- Benchmarks show that multifile approach promises best performance
  - “multifile” method : ~ 6500 MiB/sec
  - parallel ntecdf/HDF method: ~ 600 MiB/s
- The community though requires 1 single file
- Exploring the way of having one “logical file” which under the hood is spread among several files
- From the user’s point of view still 1 file is being accessed for any post processing or other work

Stub File Testing Workflow



File Stubing in the future ?



Outlook

- Implementation of a new communication scheme in ICON based on YAXT
- Performance measurements of ICON with CDI-PIO
- Analysis of multifile approach with more complicated workflows
  - Test the separation of the grid with HDCP2 output files