Deutscher Wetterdienst
Wetter und Klima aus einer Hand

DWD Applications

on Cray XC30

and Beyond

Elisabeth Krenzien (Technical Infrastructure)
Ulrich Schättler (Numerical Modeling, COSMO)
Florian Prill (Numerical Modeling, ICON)
Harald Anlauf (Data Assimilation)

Contents

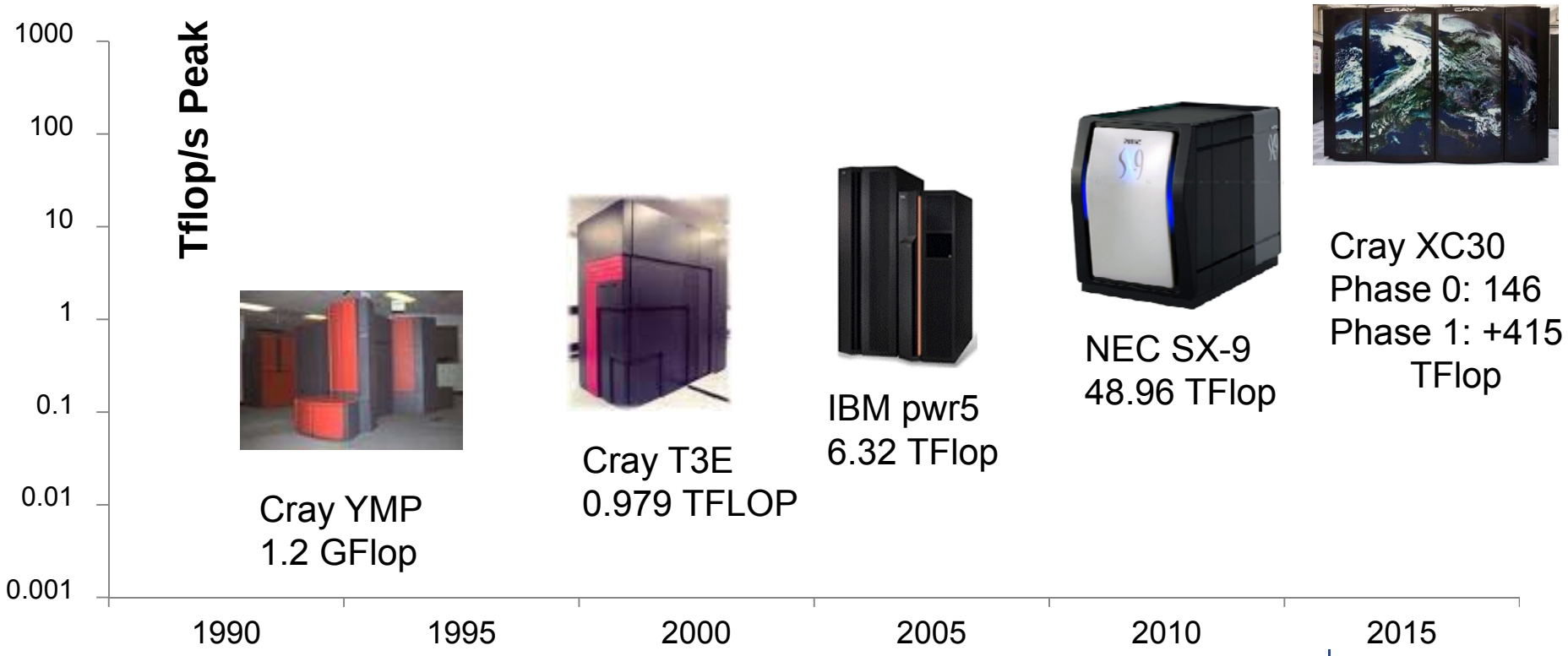
Deutscher Wetterdienst
Wetter und Klima aus einer Hand

- Result from last procurement: DWD‘s Cray XC30
- Scalability of the COSMO-Model
- Scalability of ICON
- (Ensemble) Data Assimilation
- And Beyond

The new Cray XC30 at DWD

Dec. 2012: Decision on new System

Deutscher Wetterdienst
Wetter und Klima aus einer Hand
DWD



GM (200 km)
EM (50 km)
DM (14km)
GME (60 – 40 – 30 – 20 km)
COSMO-EU (7 km)
COSMO-DE (2.8 km
COSMO-DE-EPS
ICON
(13 km)
2.2 km)
2.2 km)
EDA

Configuration Overview

Deutscher Wetterdienst
Wetter und Klima aus einer Hand
DWD

Components per cluster	Phase 0	Phase 1
Cray Cascade LC cabinets / chassis / blades	2 / 6 / 96	5 / 13 / 208
Compute blades (IvyBridge / Haswell)	91 / -	91 / 108
Compute nodes (IvyBridge / Haswell)	364 / -	364 / 432
Cores (IvyBridge / Haswell)	7280 / -	7280 / 10368
Memory per node (GB) (IvyBridge / Haswell)	64 / -	64 / 128
Memory total (TB) (IvyBridge / Haswell)	22.8 / -	22.8 / 54.0
Performance peak (TF) (IvyBridge / Haswell)	146 / -	146 / 415
Power consumption (max.) kW	148	~ 325
Home filesystem (Panasas) (Prod. / R&D)	98 / 165 (TiB)	98 / 165 (TiB)
HPC filesystem (Sonexion) (Prod. / R&D)	327 / 655 (TiB)	982 / 1965 (TiB)
Starting	Dec. 2013	Dec. 2014

Migration Work

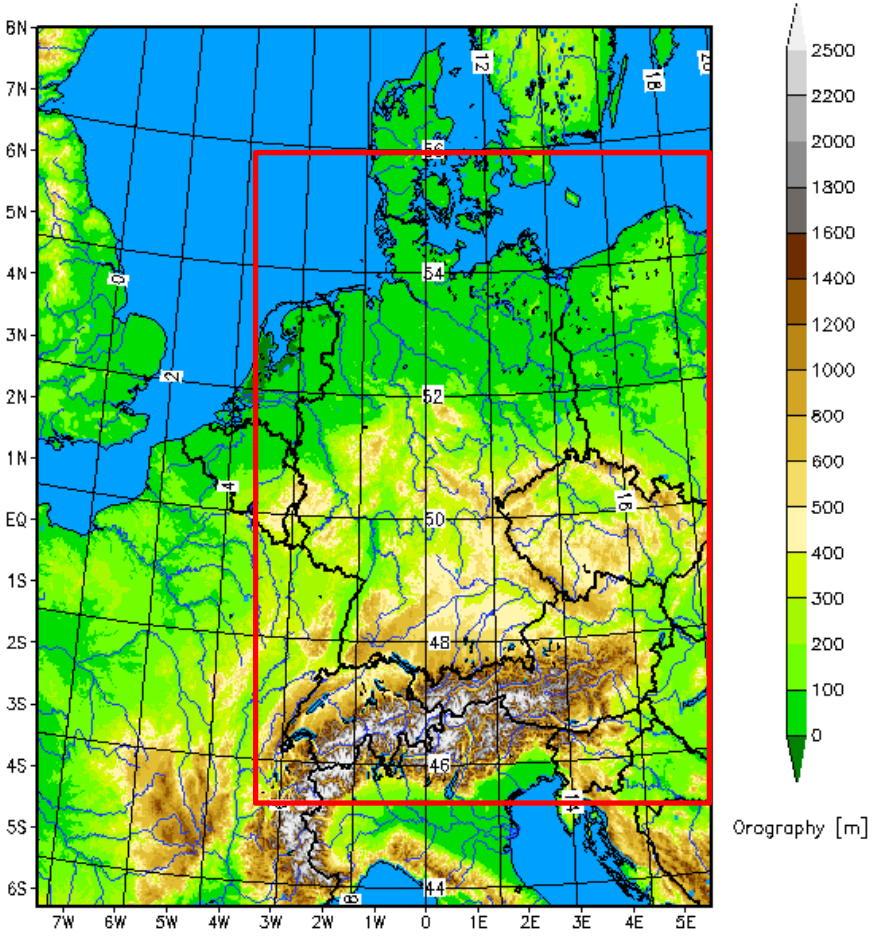
- Challenges of the porting:
  - Architecture: vector computer ⇒ scalar MPP.
  - Parallelism: 480 ⇒ ca. 18.000 cores.
- Nevertheless, the models could be migrated to the new XC30 system successfully in about 3 months.
- But with some problems and (painful) experiences:
  - compiler (Fortran 90/95 implementation; optimization; vectorization; memory management), RMA, huge pages.
  - batch system: suspend-resume, memory management.
  - parallel & heterogeneous file system: Panasas / Lustre: stability problems.
- Challenges of Phase 1:
  - heterogeneous CPUs: Ivy-Bridge (10 core) and Haswell (12 core).
  - implement the operational chain, and the numerical experimentation system.

Scalability of the COSMO-Model

The new COSMO-DE65

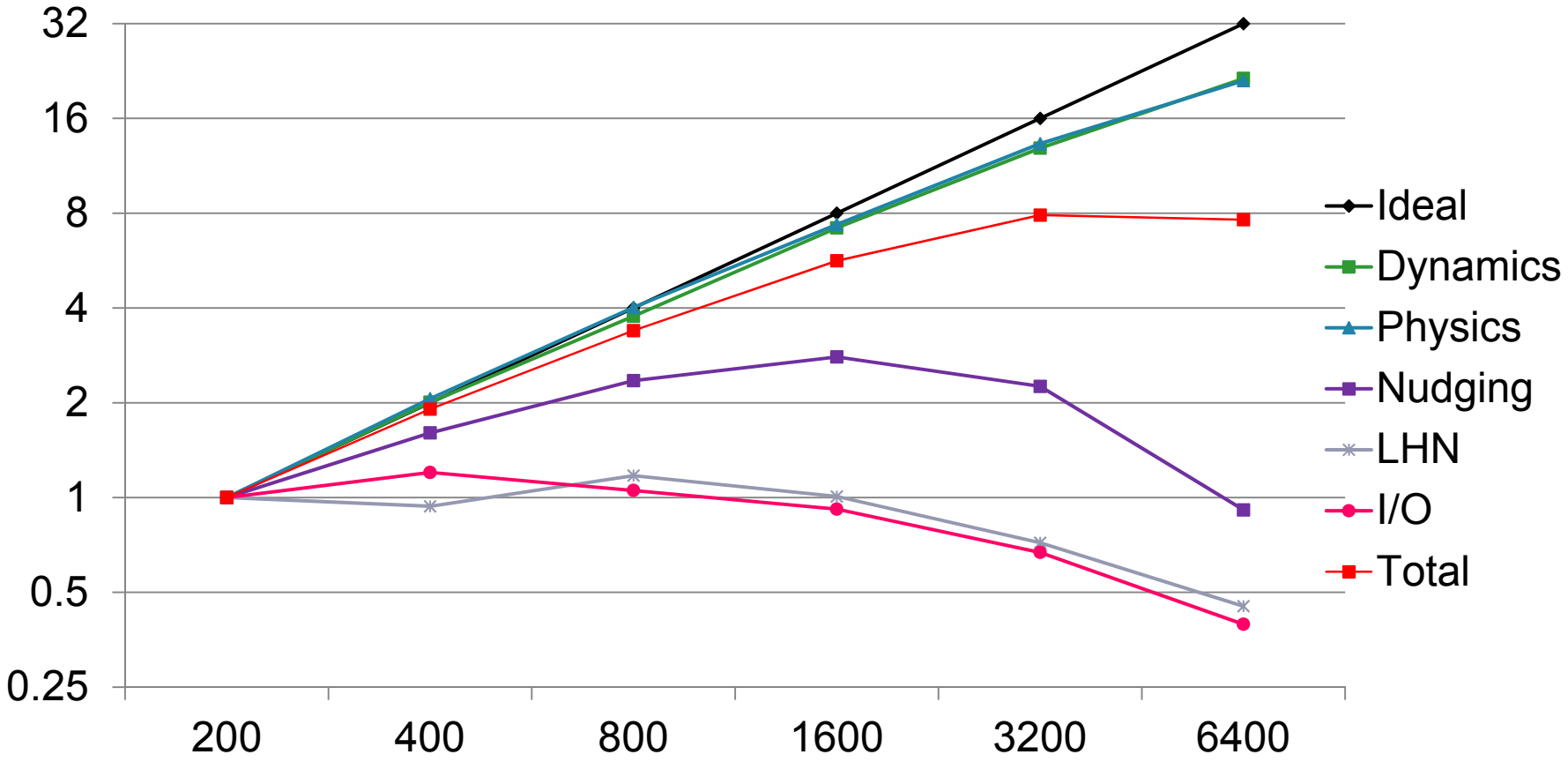
Deutscher Wetterdienst
Wetter und Klima aus einer Hand

- In 2015, DWD will upgrade the COSMO-DE to a larger domain and 65 vertical levels: $651 \times 716 \times 65$ grid points.
- Some specialities about COSMO-DE forecast:
  - 12 hour forecast should run
    - in ≤ 1200 s in ensemble mode,
    - in ≤ 400 s in deterministic mode.
  - „nudgecast“ run: nudging and latent heat nudging in the first 3h.
  - SynSat pictures every 15 minutes.
  - amount of output data per hour: 1.6 GByte: asynchronous output is used with 4 or 5 output cores.
- How many cores are necessary?



Scalability of COSMO Components (incl. Comm.)

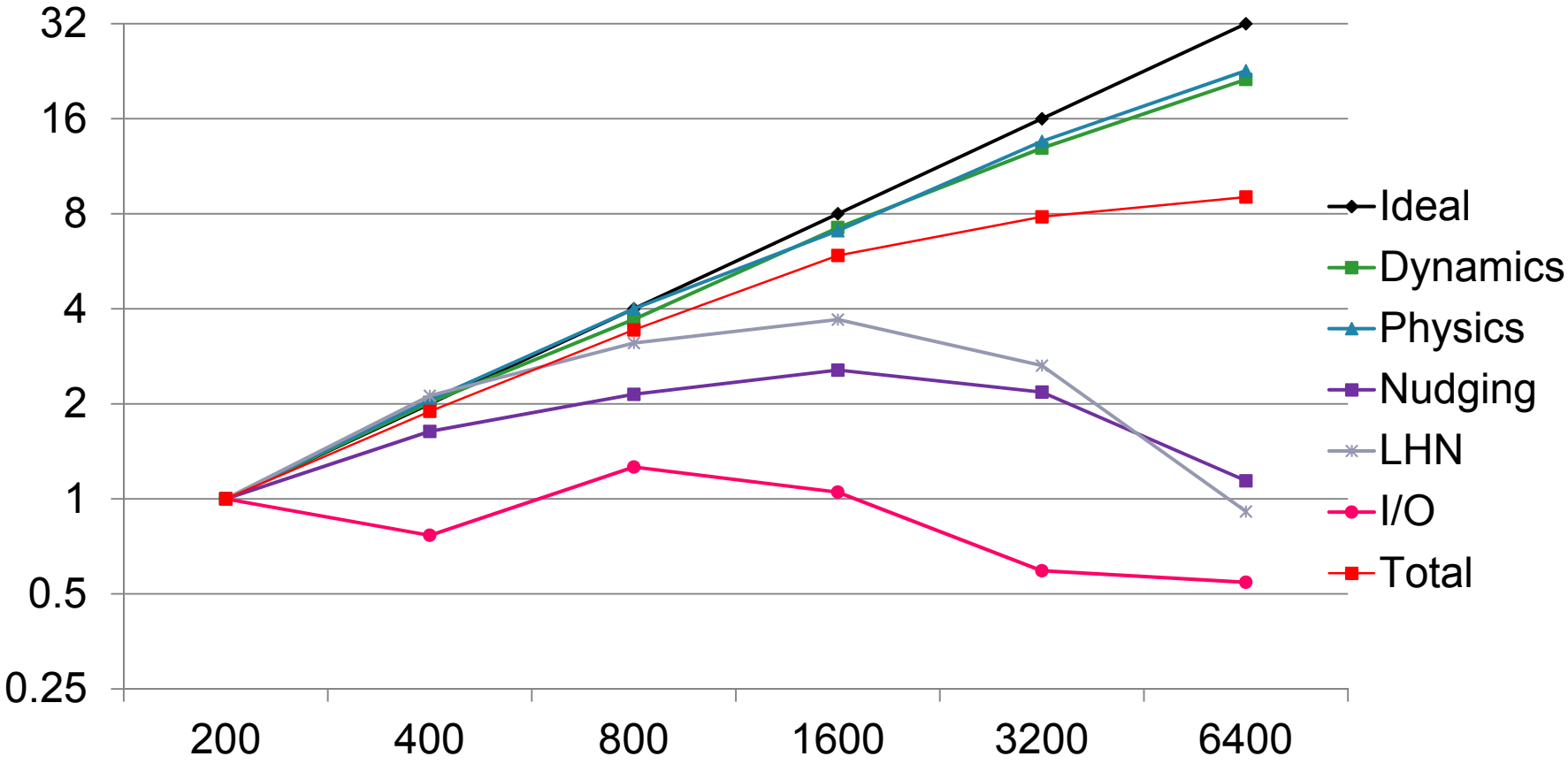
Deutscher Wetterdienst
Wetter und Klima aus einer Hand



Scalability of COSMO Components (incl. Comm.)

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

Could optimize a global communication in Latent Heat Nudging



Timings for COSMO-DE65 (new)

Deutscher Wetterdienst
Wetter und Klima aus einer Hand
DWD

Timings were done with COSMO-Model 5.1 (to be released in Nov. 2014)

# cores	196+4	396+4	795+5	1596+4	3196+4	6396+4
Dynamics	1848.06	908.56	474.62	240.43	132.24	77.56
Dyn. Comm.	256.83	133.92	93.56	50.69	30.82	21.23
Physics	326.42	157.02	80.12	43.44	23.20	13.68
Phy. Comm.	16.55	9.20	5.73	2.24	1.03	0.46
Nudging	26.72	15.24	10.75	7.61	6.56	9.43
LHN	16.51	8.62	5.20	4.41	6.86	20.73
Nud.+ LHN Comm.	13.22	7.74	7.36	7.55	10.82	23.98
Add. Comp.	754.63	416.42	223.49	111.80	56.37	34.75
Input	36.30	66.27	31.99	38.22	74.47	52.87
Output	33.92	25.38	23.69	28.66	44.20	76.21
Total	3355.67	1772.82	978.67	568.88	428.91	371.35

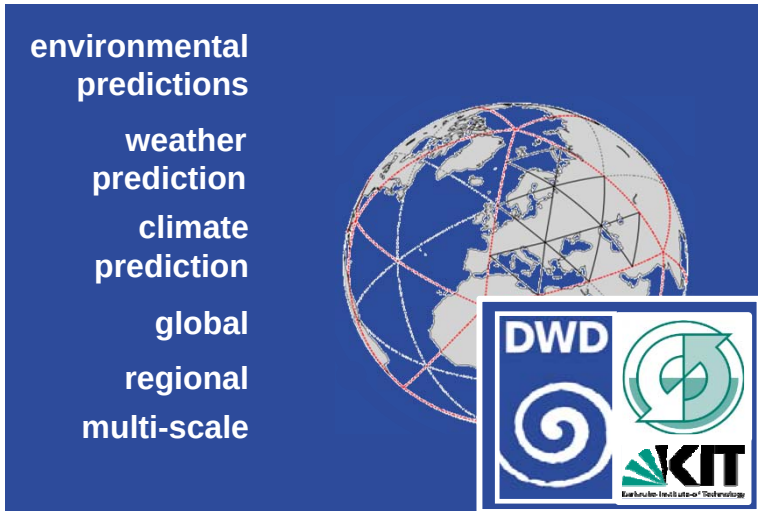
Comments

- Scalability of COSMO for COSMO-DE65 domain size is reasonably well up to 1600 cores. Dynamics and Physics also scale beyond up to 6400 cores.
- Meeting the operational requirements:
  - Deterministic mode: using nearly the full Phase 0 machine, it is (now) possible to run COSMO-DE65 in less than 400 seconds.
  - Ensemble mode: The ensembles are running without synthetic satellite images. Using 440 cores, a 12 hour forecast can be done in about 1320 seconds (1200 seconds were planned). But this is still within the tolerance limit.
- This is mainly a problem of some expensive components!
  - New fast-waves solver is more expensive than old one (40-50% of dynamics time; 20-25% of total time).
  - Additional Computations: is almost only in RTTOV10:
    - factor of about 10-15 compared to RTTOV7.
- Tests were done on a machine „crowded as usual“ and really reflect the operational setups (no tricks, no cheating, no beautifying).

Scalability of ICON

ICON: ICOsahedral Nonhydrostatic Model

Deutscher Wetterdienst
Wetter und Klima aus einer Hand



Joint development project of DWD and Max-Planck-Institute for Meteorology

- about 40 active developers from meteorology and computer science
- ~ 600,000 lines of Fortran code

Lately joined by KIT to implement ICON-ART (environmental prediction)

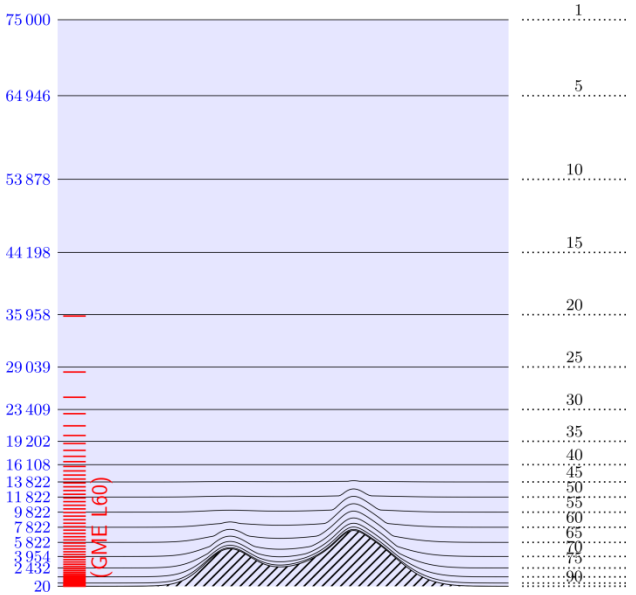
Future Milestones 2014-2020

- replacement of GME/COSMO-EU by ICON global model and Europe grid
- ensemble data assimilation for ICON
- ICON-EPS global ensemble system

ICON to GME comparison

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

GME	ICON
hydrostatic, Arakawa A	non-hydrostatic, Arakawa C
pressure-based vertical grid, TOA 35 km	hybrid z-based, TOA ca. 75 km
flat-MPI parallelization	hybrid MPI-OpenMP
1.5 Mio. grid points/level	2.9 Mio. grid points (planned operational setup)
	~ 3.5x more output on the model grid!



currently operational

HPC challenges with the new model

Deutscher Wetterdienst
Wetter und Klima aus einer Hand
DWD

5 km resolution tests at ECMWF Cray XC30 reveal possible future bottlenecks:

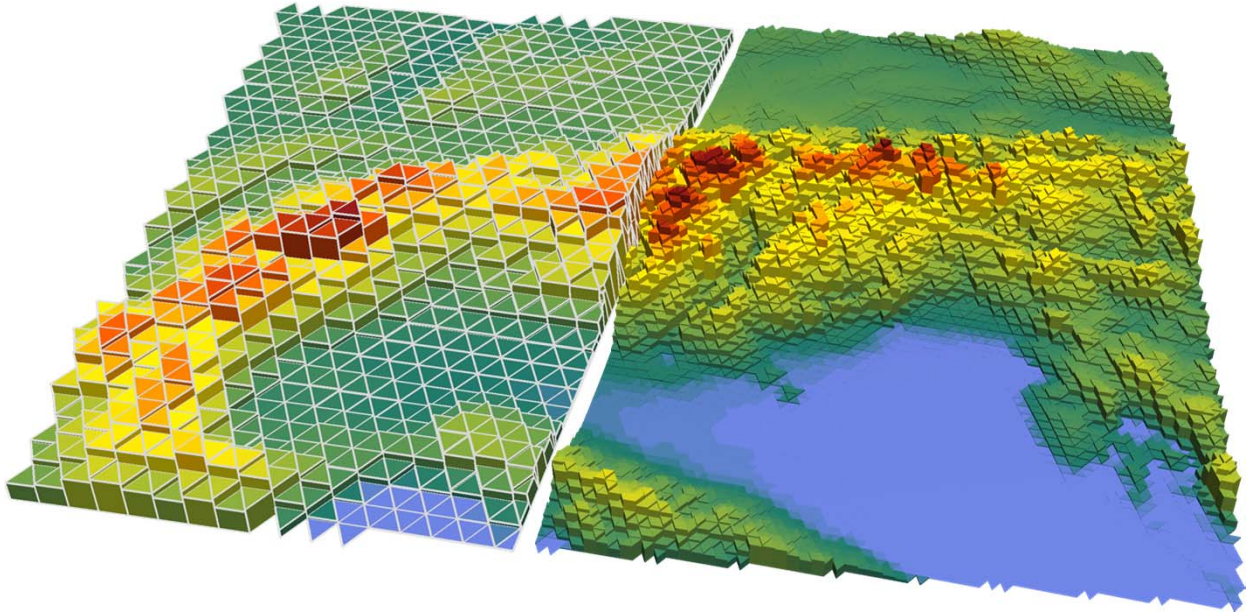
- memory scaling:
  avoid global arrays
- I/O scaling:
  NetCDF4 input files with ~125 GB data
- One-sided MPI:
  Cray-specific problems with RMA (DMAPP)
- load balancing:
  e.g. day-night load imbalance for radiation scheme

Static Domain Decomposition

Example: 1100 PEs

HPC challenges with the new model

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

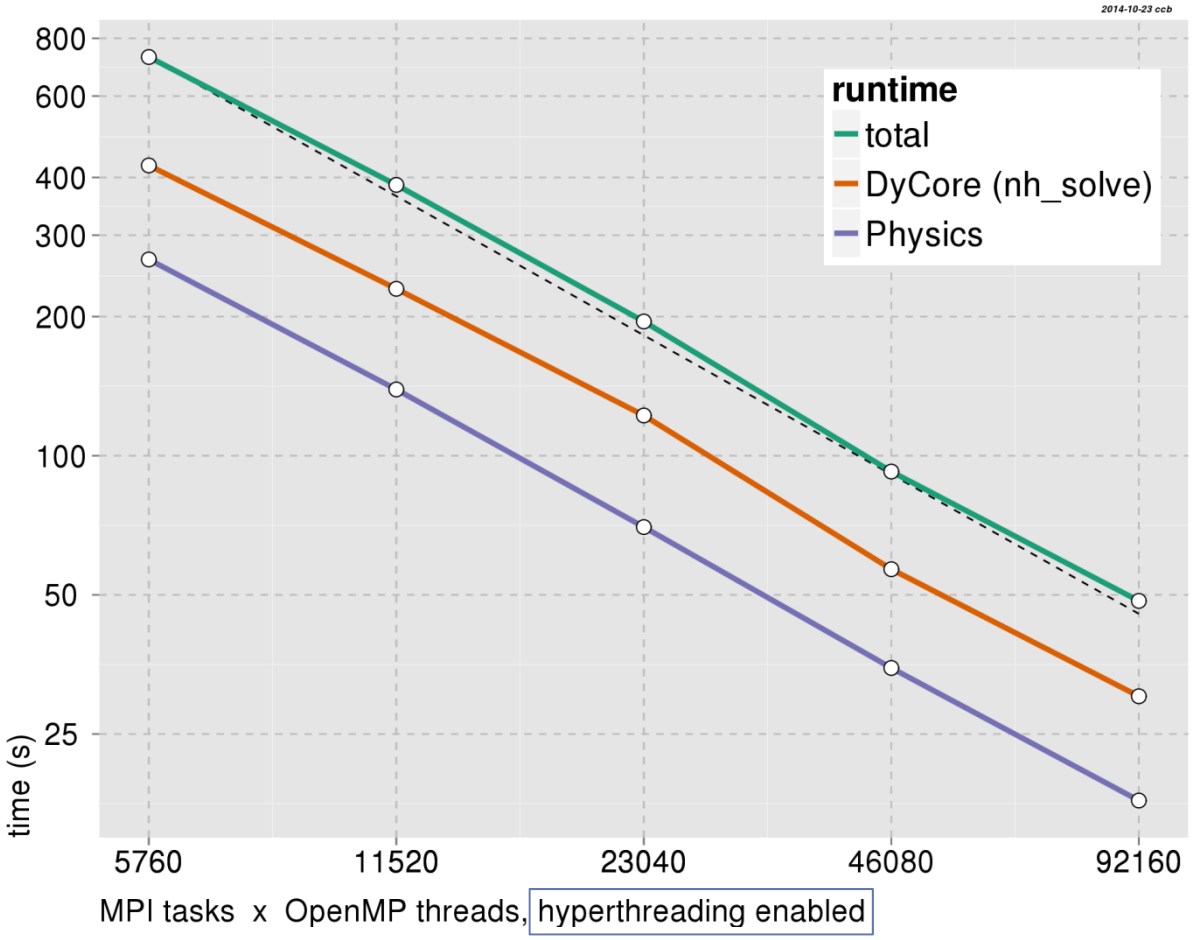


„R03B07“ 13 km resolution
soon-to-be operational

„R02B09“ scaling test (Oct `14)
5 km global resolution

ICON Parallel Scaling on ECMWF´s XC30 („ccb“)

Deutscher Wetterdienst
Wetter und Klima aus einer Hand



Real-Data Test Setup

(date: 01.06.2014)

5 km global resolution

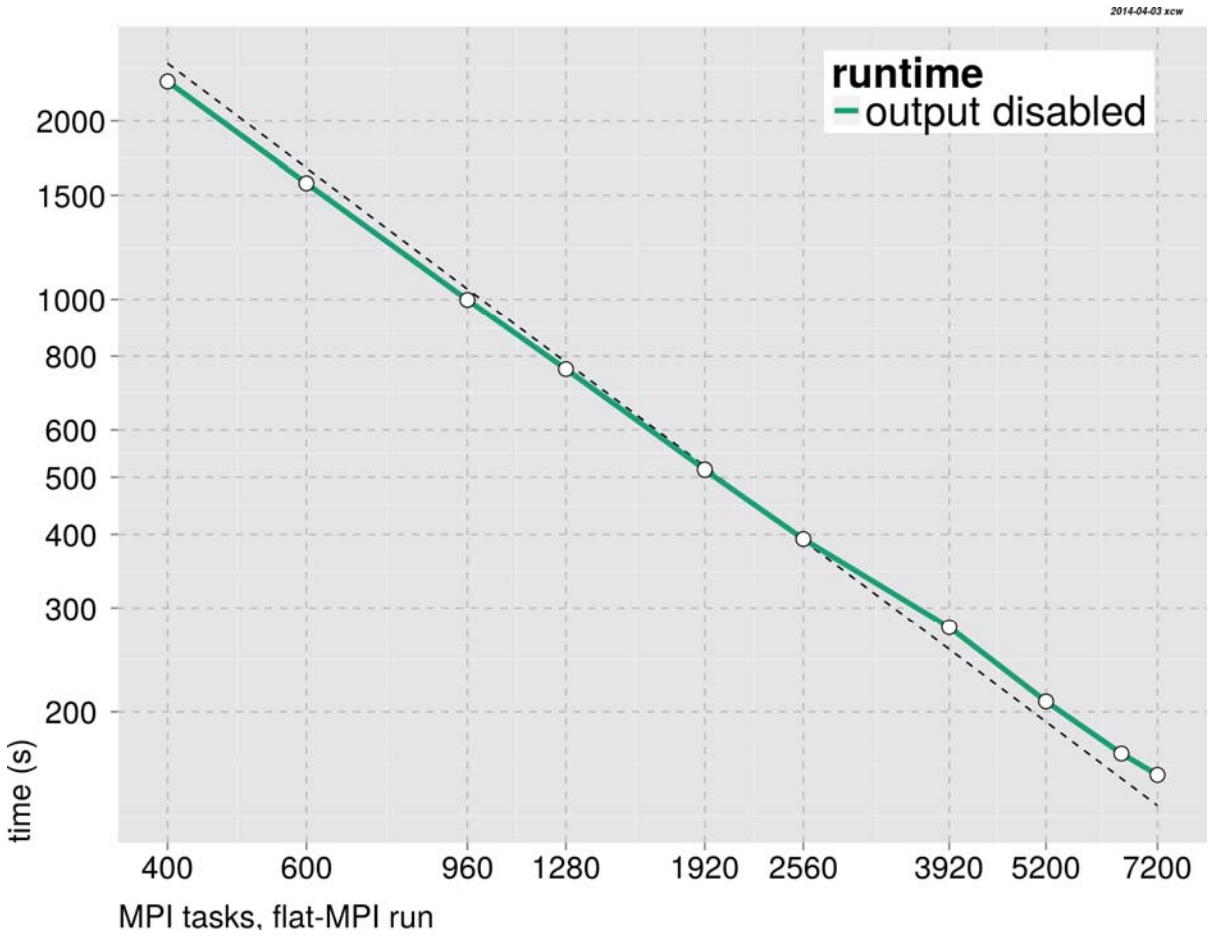
20,971,520 grid cells

hybrid run, 4 threads/task

1000 steps forecast,
w/out reduced radiation grid,
no output

ICON Parallel Scaling on DWD´s XC30

Deutscher Wetterdienst
Wetter und Klima aus einer Hand



Real-Data Test Setup

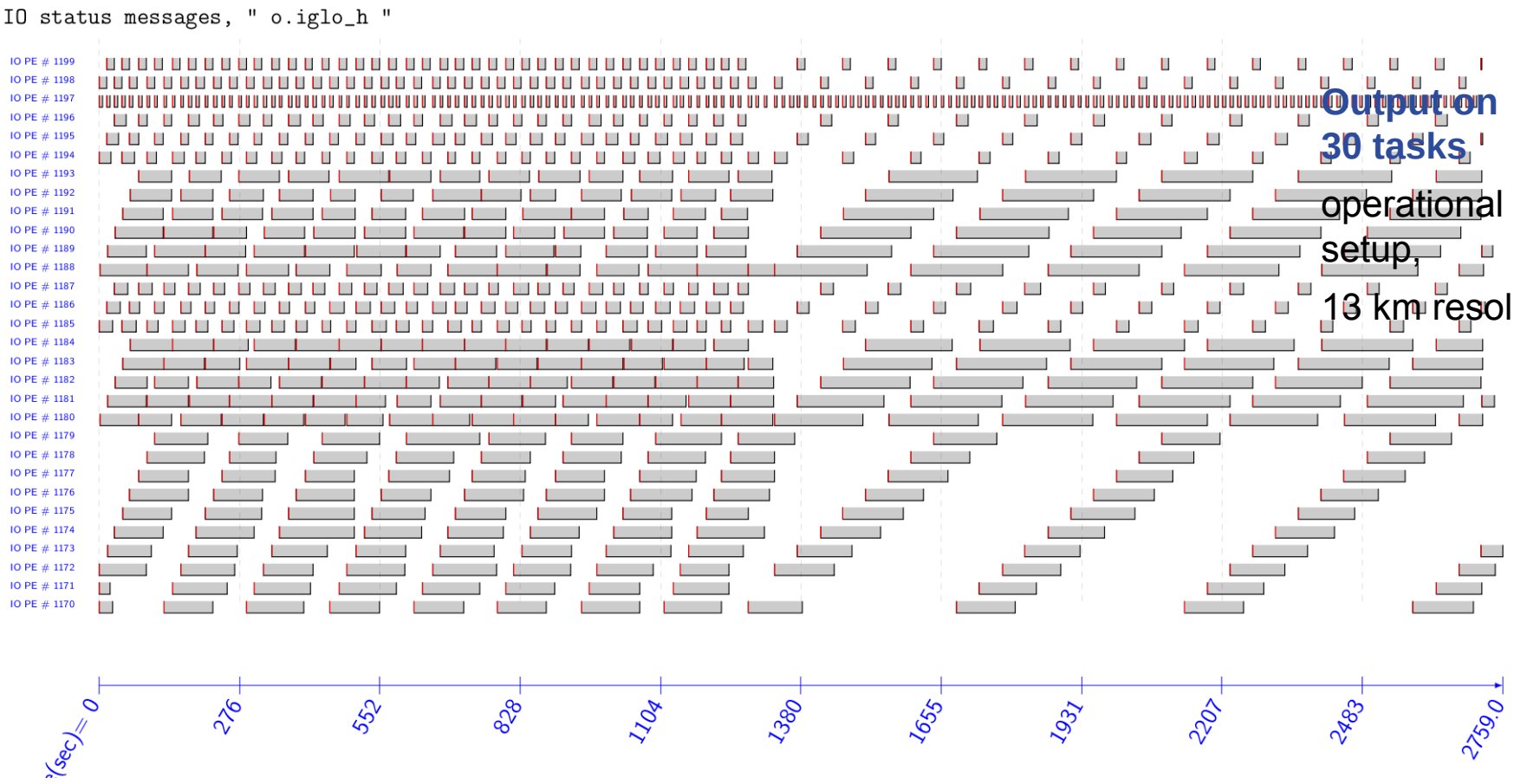
results obtained on
Phase 0 system (April ´14)

13 km global resolution

24 h forecast, xcw.dwd.de,
with reduced radiation grid

ICON I/O Servers

Deutscher Wetterdienst
Wetter und Klima aus einer Hand



Output on 30 tasks

operational setup,

13 km resol.

- **computation and I/O overlap**
- **fast system layer: *Climate Data Interface***
- **WMO GRIB2 standard (ECMWF's `GRIB_API`)**

(Ensemble) Data Assimilation

Data Assimilation at DWD: Algorithms

Deutscher Wetterdienst
Wetter und Klima aus einer Hand
DWD

- Global deterministic Data Assimilation, for GME/ICON:
  Variational Analysis (3D-Var PSAS, operational)
  - Minimization of (non-linear) cost-function

- Convective Scale:
  - Deterministic Analysis: Nudging Scheme (operational)
  - **K**m-Scale **EN**semble **D**ata **A**ssimilation (KENDA):
    LETKF (Local Ensemble Transform Kalman Filter, *testing*)

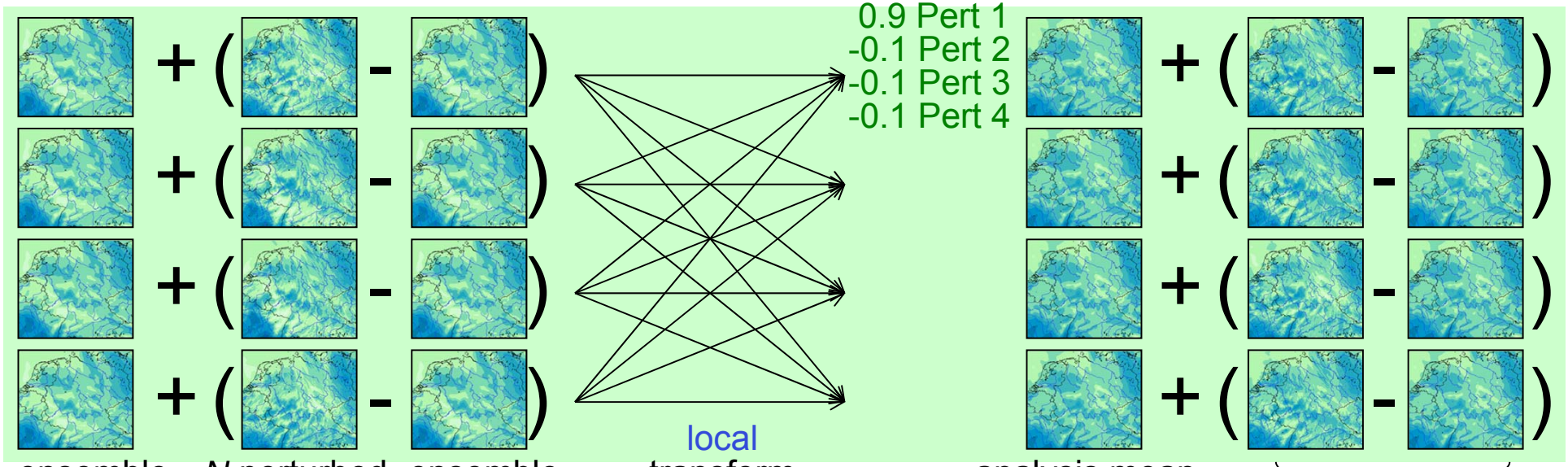
- Global Ensemble Data Assimilation for ICON (IEDA):
  - LETKF (*testing*)
  - VarEnKF (Hybrid Variational/LETKF system, *under development*)

Selected HPC Challenges in Data Assimilation

- I/O of model fields and observational data
  - Single-node bandwidth typically only several hundred MByte/s
  - Aggregate bandwidth by scattering different files over nodes (most easily achievable for ensemble methods)
- For given architecture: try to find suitable algorithm
  - Variational data assimilation: solver needs efficient implementation of linear operators $B(x)=\mathbf{B}\bullet x$ for high dimensions ($n=10^8$)
    - In suitable basis (spectral or wavelet), **B** often a sparse matrix
    - Cache-based machines: **CSR** & **CSC** storage order reasonably efficient; small improvements possible with suitable permutations (CSRPERM, CSCPERM)
    - Vector processors (e.g. NEC-SX): **J**agged **D**iagonal **S**torage order allows for longer vectors, additional optimizations (Sunil R. Tiyyagura, Uwe Küster, Stefan Borowski, ICCS '06)

Local Ensemble Transform Kalman Filter (LETKF, Hunt et al. 2007)

Deutscher Wetterdienst
Wetter und Klima aus einer Hand



flow-dep background error cov.

$$\mathbf{P}^b = (N-1)\,\mathbf{X}^b\,(\mathbf{X}^b)^T$$

in the (*N*-1) -dimensional (!) sub-space S spanned by background perturbations :

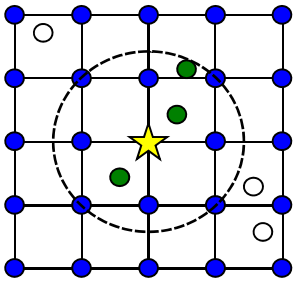
$x^{(a(i))} = \bar{x}^b + X^b w^{(a(i))}$ set up cost function $J(\mathbf{w})$ in ensemble space,
explicit solution for minimisation (Hunt et al., 2007)

LETKF (km-scale) for COSMO

Deutscher Wetterdienst
Wetter und Klima aus einer Hand
DWD

Summary of method: Use implementation following Hunt et al., 2007

- Basic idea: perform analysis in the space of ensemble perturbations
  - computationally efficient (!) in terms of operation count
  - explicit localization: separate analysis at each grid point, using only nearby observations: local data, well parallelizable
  - analysis ensemble members are locally linear combinations of first guess ensemble members
  - Weight calculation: symmetric square-root of *NxN* matrix at each analysis grid point, e.g. using eigenvalue decomposition
    - On NEC-SX, DSYEV from Mathkeisan was too slow (many bank conflicts, short vectors, few MFlop/s); RS from ancient EISPACK (iterative algorithm) already factor 5 faster; further speedup by factor 10 by vectorizing over matrix index ("loop pushing") even with additional indirect addressing. Are there better ways?
    - LAPACK's DSYEV on cache-based machines reasonably fast



And Beyond

Up to 2018

- The contract with Cray lasts until December 2016 (with the option to extend it up to 2 years until December 2018)

- If we would get new hardware after 2016, what could it be?
  - Intel based: Haswell / Broadwell
  - other scalar CPUs: Sparc, others,…
  - vector CPUs

- Or even:
  - Intel Knights Landing (?)
  - GP GPUs (?)

- But is our software able to use Knights Landing or GPUs?

Accelerators

Deutscher Wetterdienst
Wetter und Klima aus einer Hand
DWD

- COSMO-Model
  - Much work has been invested in Switzerland (CSCS; MCH) to port the full COSMO-Model to GPUs. A prototype will be ready by the end of the year.
  - Most of the code has been ported using OpenACC directives, but:
    - only few compilers available for OpenACC at the moment;
    - will Intel understand OpenACC for Knights Landing?
  - The dynamics has been completely re-written by defining a meta language (DSL): STELLA (a „Stencil Library“).

- ICON
  - By mid (end) of 2015, ICON and the COSMO-Model will share the same physics packages. Most of them are then implemented using OpenACC.
  - Work on ICON dynamical core is ongoing to implement OpenACC directives (work by CSCS)

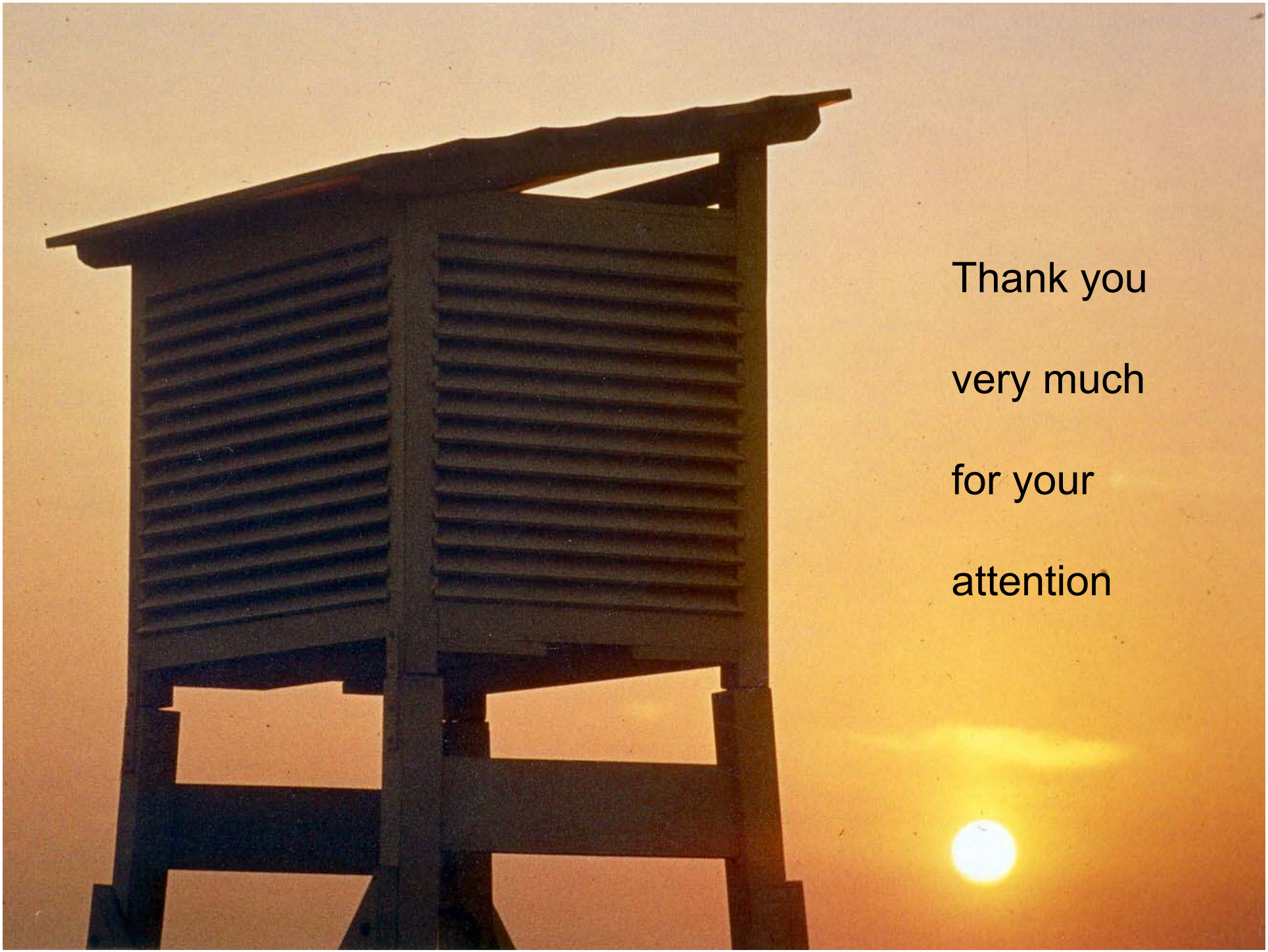
Conclusions and Outlook

Deutscher Wetterdienst
Wetter und Klima aus einer Hand
DWD

- Porting DWD applications to Cray XC30 was successful.
- Phase 1 (update in progress!) will run 45 members of COSMO-DEbig ensemble: Upgrade by a factor of 3.
- There are still a few problems with the memory management (Cray compiler) and performance of some collective operations (Cray is working on it).
- Work is going on to adapt the operational and the experimenting system to the heterogeneous Phase 1 system.

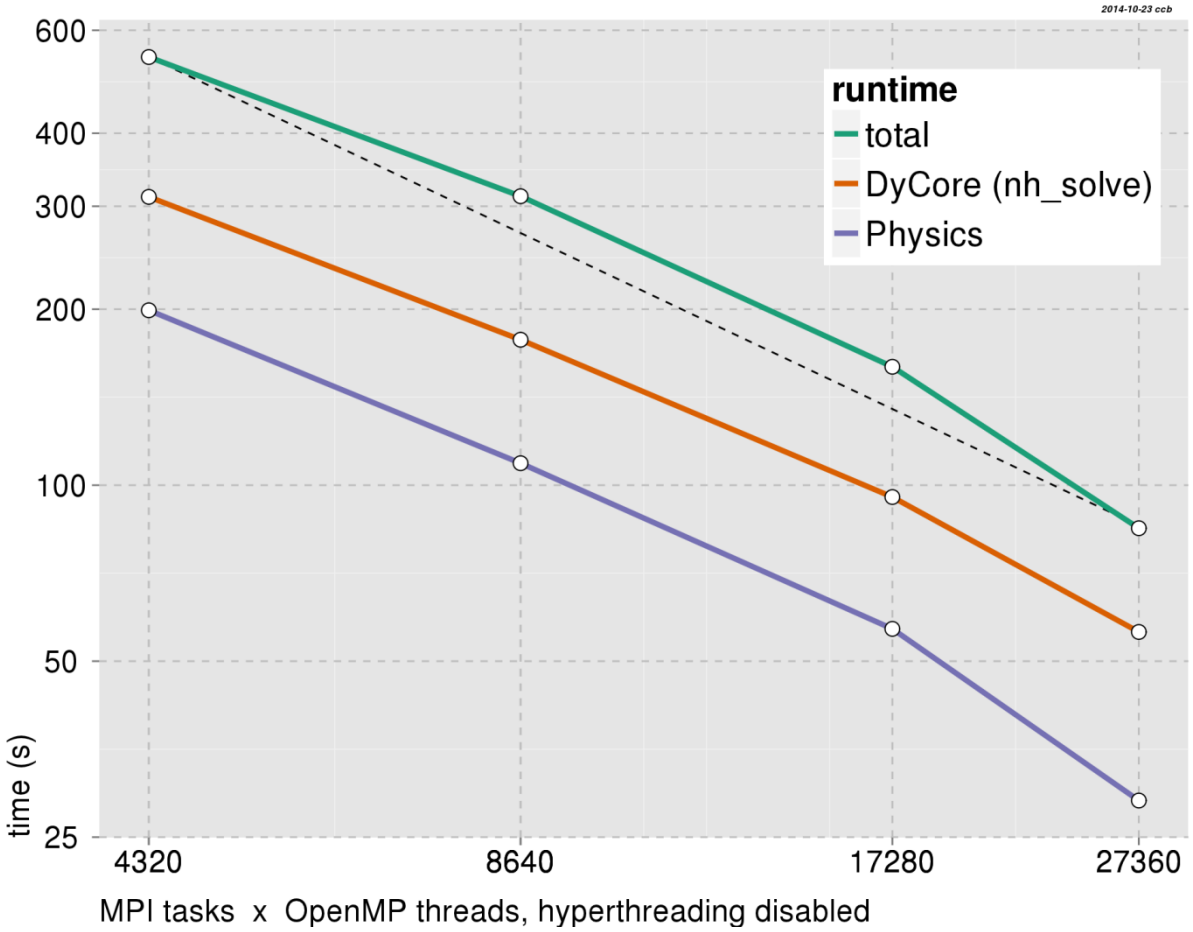
- Will OpenACC and OpenMP have a common future?
- First experiences have been gathered in COSMO to run the model on GPUs, but we are not yet ready to use accelerators right now.
- See the presentation by our colleague Xavier Lapillonne (from Switzerland), later this week:

  „Are OpenACC directives the easiest way to port Numerical Weather Prediction applications to GPUs?“


Thank you
very much
for your
attention

ICON Parallel Scaling on ECMWF´s XC30 („ccb“)

Deutscher Wetterdienst
Wetter und Klima aus einer Hand



Real-Data Test Setup

(date: 01.06.2014)

5 km global resolution

20,971,520 grid cells

hybrid run, 6 threads/task

1000 steps forecast,
w/out reduced radiation grid,
no output