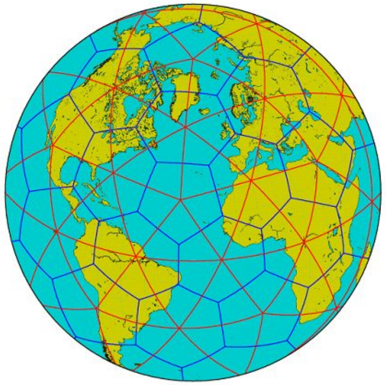
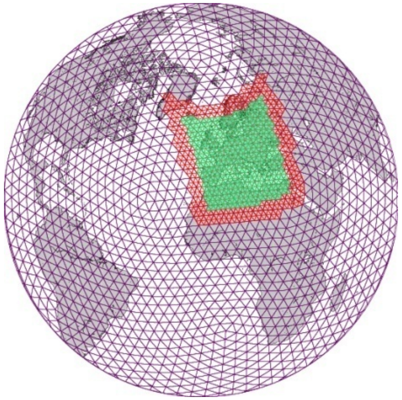




ICON

The Icosahedral Nonhydrostatic modelling framework

Key aspects for computational efficiency and scalability

ECMWF Workshop on Scalability, 14.04.2014

Günther Zängl, on behalf of the ICON development team

Outline

- Introduction: Main goals of the ICON project
- Dynamical core and numerical implementation
- Efficiency and scalability
- Conclusions

Primary development goals

- Unified modeling system for NWP and climate prediction in order to bundle knowledge and to maximize synergy effects between DWD and Max-Planck-Institute for Meteorology
- Better conservation properties
- Nonhydrostatic dynamical core for capability of seamless prediction
- Scalability and efficiency on O(10^4+) cores
- Flexible grid nesting in order to replace both GME (global, 20 km) and COSMO-EU (regional, 7 km) in the operational suite of DWD
- Limited-area mode to achieve a unified modelling system for operational forecasting in the mid-term future

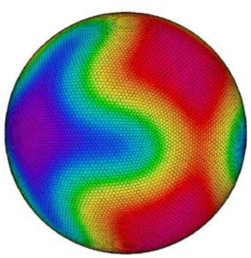
Related projects dealing with hpc aspects

HD(CP)²
High definition clouds and precipitation for advancing climate prediction

HD(CP)² (led by MPI-M, Hamburg):

High-definition clouds and precipitation for advancing climate prediction

Goal: simulations with 100 m mesh size over (almost) the whole of Germany



ICOMEX (led by DWD):

ICOsahedral-grid models for EXascale earth-system simulations

ICON-related subproject: DSL version of dynamical core

Model-independent subprojects: parallel I/O, parallel internal postprocessing

Thoughts on efficient time-stepping schemes in global models

- **Fact: ratio between sound speed and maximum wind speed approaches unity when the model resolution permits breaking gravity waves in the upper stratosphere / mesosphere**
- **Thus, split-explicit schemes such as widely used in mesoscale models may not be beneficial**
- **Semi-implicit schemes need to avoid a limitation by the advective Courant number (e.g. SISL)**
- **For ICON, we decided to use a HEVI (horizontally explicit – vertically implicit) scheme with time splitting between the dynamical core and tracer advection + physics parameterizations**

Model equations, dry dynamical core

(see Zängl, G., D. Reinert, P. Ripodas, and M. Baldauf, 2014, QJRMS, in press)

$$\frac{\partial v_n}{\partial t} + (\zeta + f)v_t + \frac{\partial K}{\partial n} + w\frac{\partial v_n}{\partial z} = -c_{pd}\theta_v \frac{\partial \pi}{\partial n}$$

$$\frac{\partial w}{\partial t} + \vec{v}_h \cdot \nabla w + w\frac{\partial w}{\partial z} = -c_{pd}\theta_v \frac{\partial \pi}{\partial z} - g$$

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\vec{v}\rho) = 0$$

$$\frac{\partial \rho\theta_v}{\partial t} + \nabla \cdot (\vec{v}\rho\theta_v) = 0$$

v_n,w: normal/vertical velocity component

ρ: density

θ_v: Virtual potential temperature

K: horizontal kinetic energy

ζ: vertical vorticity component

π: Exner function

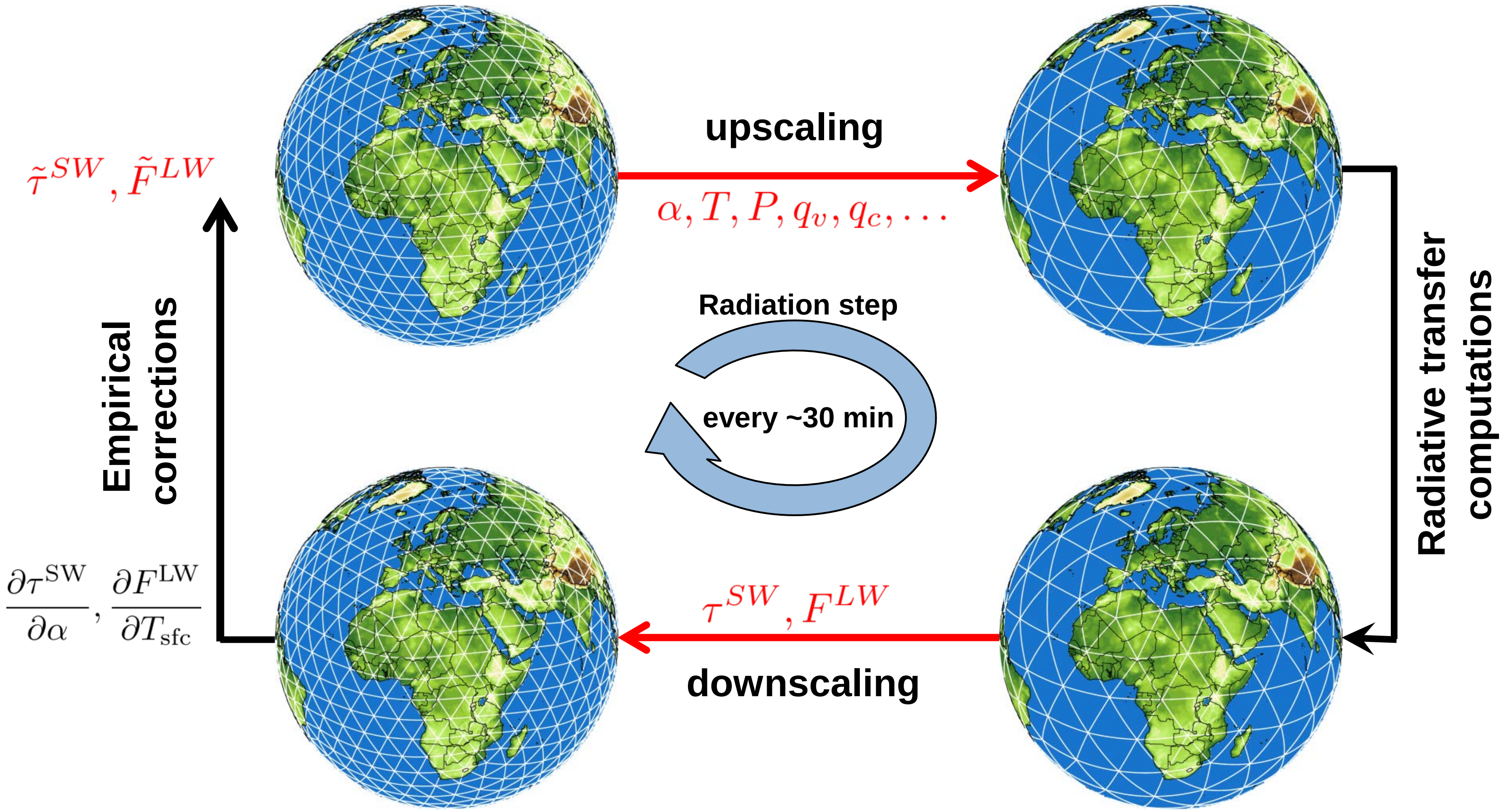
blue: independent prognostic variables

Numerical implementation

- Discretization on icosahedral-triangular C-grid
- Two-time-level predictor-corrector time stepping scheme
- Horizontally explicit-vertically implicit scheme; larger time steps (default 5x) for tracer advection / horizontal diffusion / physics parameterizations
- Tracer advection with 2nd-order and 3rd-order accurate finite-volume schemes with optional positive definite or monotonous flux limiters; index-list based extensions for large CFL numbers; substepping for QV advection above ~20 km (moisture physics is turned off above 22.5 km)
- No global communication except for diagnostics and I/O

Reduced radiation grid

- Hierarchical structure of the triangular mesh is very favourable for calculating physical processes (e.g. radiative transfer) with different spatial resolution compared to dynamics.



Code-level efficiency optimization

- Adjustable block length (‘nproma’)
- Memory storage order (cells,levels,blocks), but cpp-directive based possibility to switch from horizontal to vertical index for inner loop in indirectly addressed loops
- Option to use single precision for intermediate storage of derived quantities and some metric coefficients (dynamical core and transport scheme)
- Combined minimization of computations on halo points and number of communication calls (with priority on minimizing the latter)

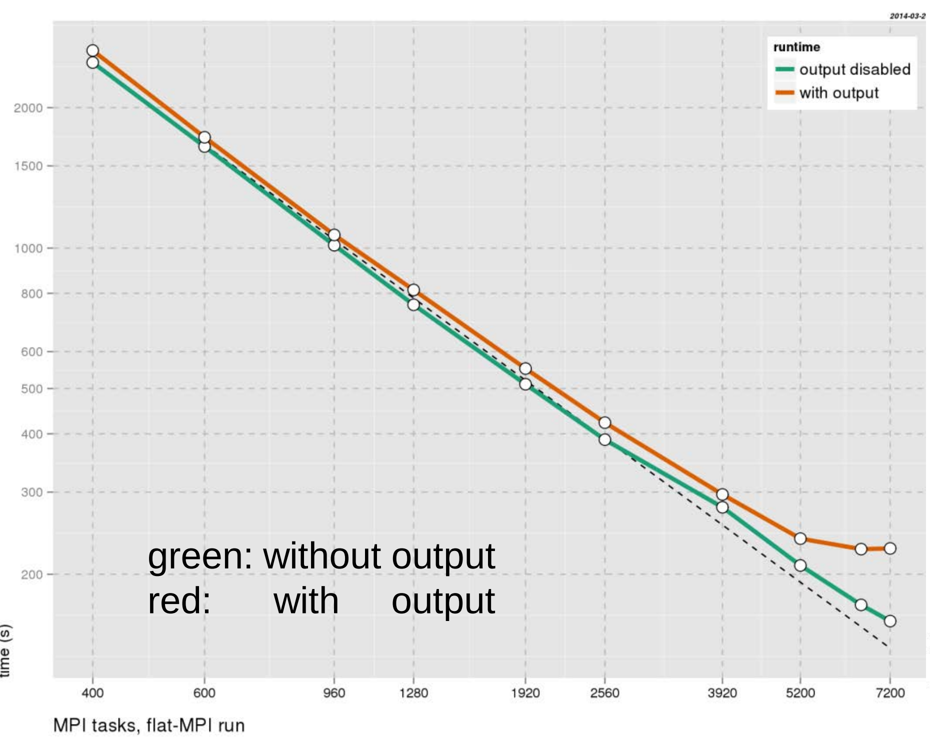
ICON vs. GME

- **GME: hydrostatic operational global model, icosahedral-hexagonal A-grid**
- **Semi-implicit leapfrog time-stepping scheme, time step limited by advective Courant number, iterative solver (SOR) for elliptic equation (thereby no global communication, but very frequent halo exchange)**
- **NEC SX-9: ICON runs a factor of 3-4 faster than GME for operational domain size (20 km / 60 levels)**
- **CRAY XC 30: ICON runs about a factor of 2 faster than GME (much faster communication network than SX-9, therefore better performance of GME)**

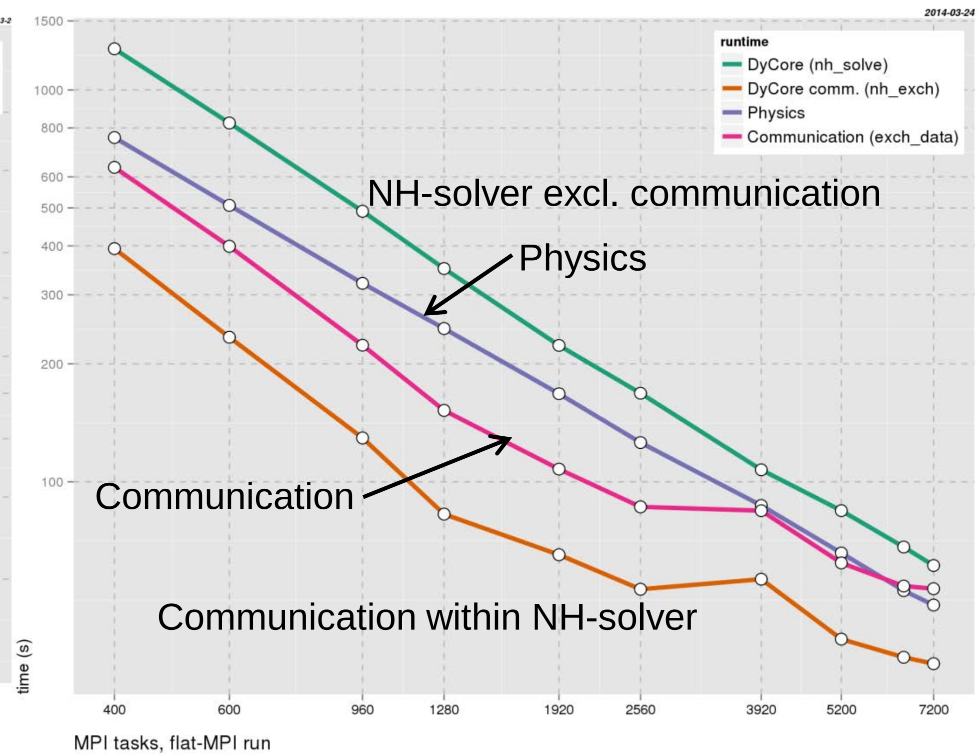
Scaling test

- Mesh size 13 km (R3B07), 90 levels, 1-day forecast (3600 time steps)
- Full NWP physics, asynchronous output (if active) on 42 tasks
- Range: 20–360 nodes Cray XC 30, 20 cores/node, flat MPI run

total runtime



sub-timers



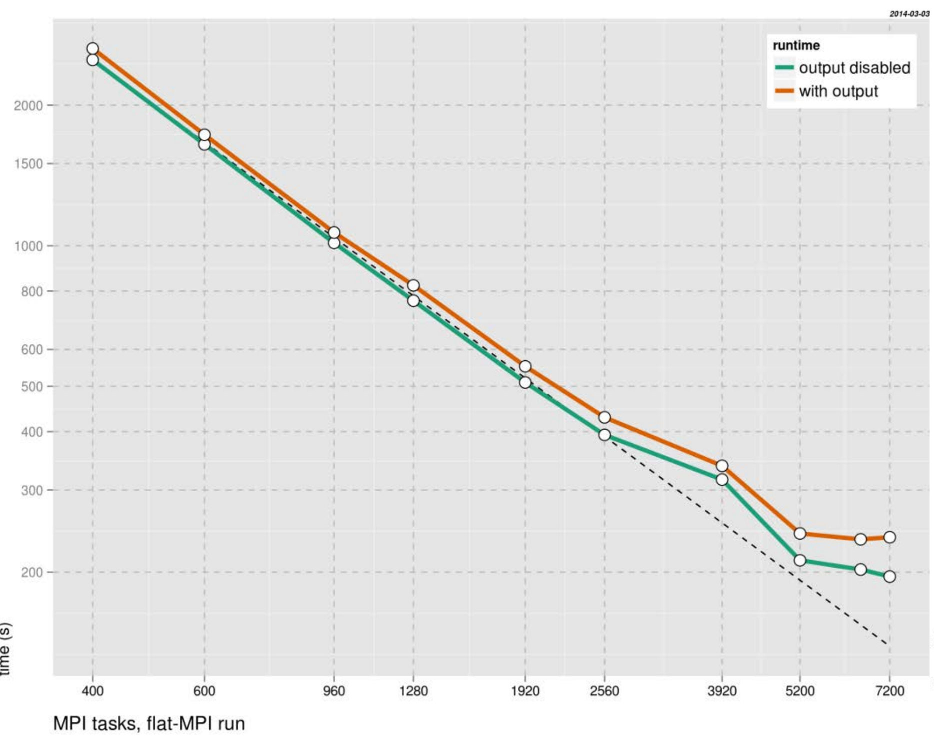
Thanks to Florian Prill!

Max-Planck-Institut für Meteorologie

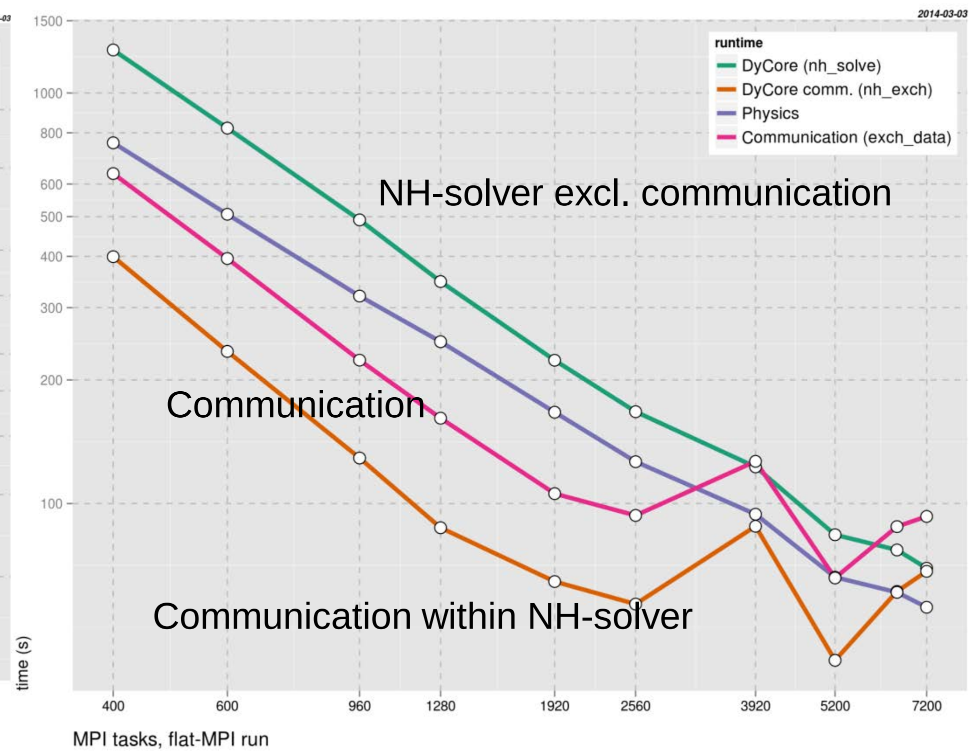
Deutscher Wetterdienst
Wetter und Klima aus einer Hand

Result of first try – before fixing some hardware issues …

total runtime



sub-timers

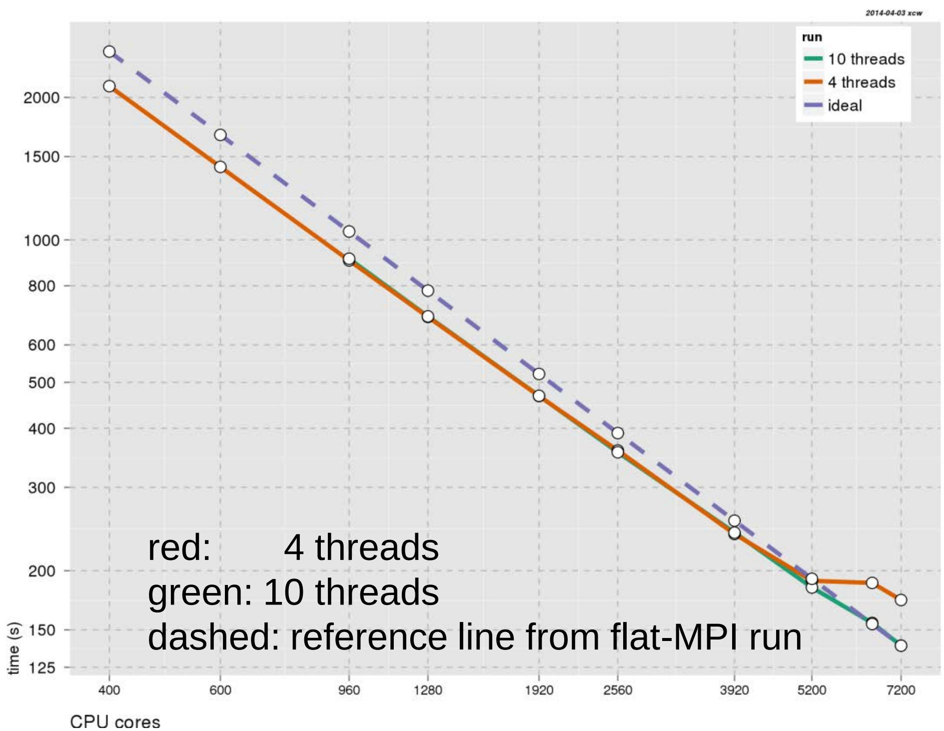


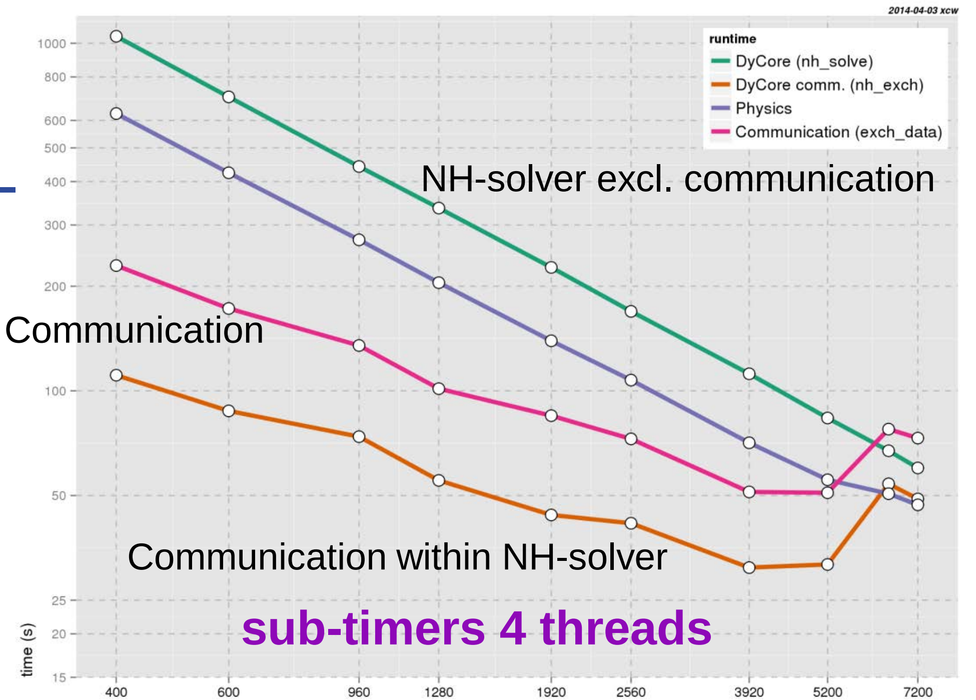
Max-Planck-Institut
für Meteorologie

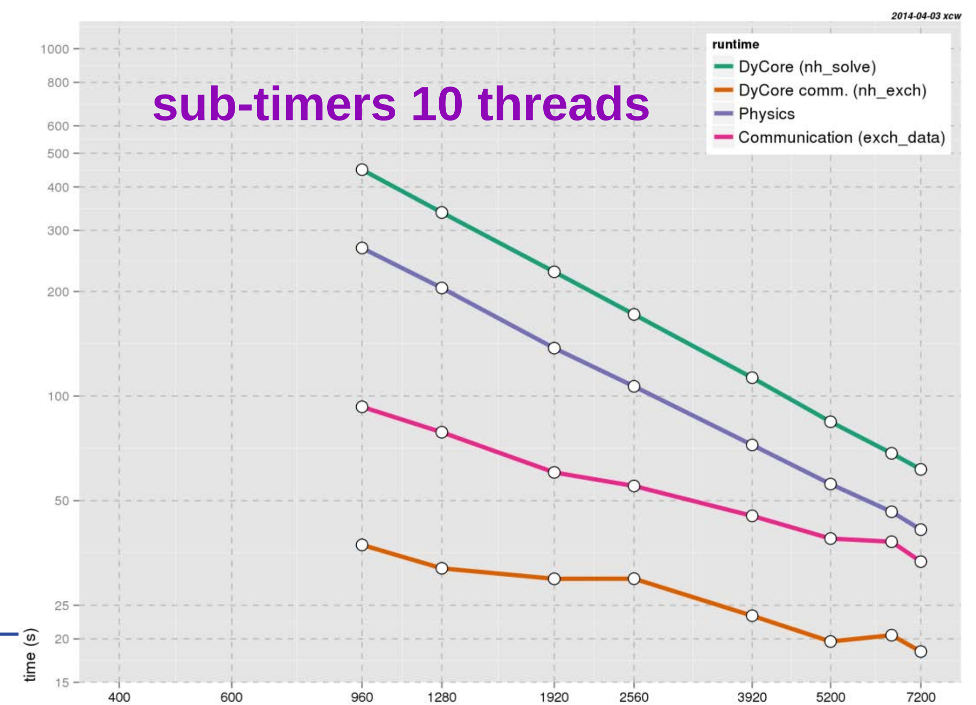
Hybrid parallelization:

4/10 threads with hyperthreading

total runtime (no output only)







[redacted]

Scaling tests on Cray XC30: important findings

- **Combined usage of hyperthreading and hybrid parallelization speeds up program execution by 10 – 15%**
- **Nearly identical results for 4 and 10 threads when using less than 75% of the machine, beyond that strange behaviour of communication times with 4 threads (does not occur with 10 km mesh size)**
- **Should be repeated from time to time to check for hardware issues…**

Major upcoming challenges

- Memory scaling I: remove remaining global fields used for computing the domain decomposition and communication patterns
- Memory scaling II: minimize usage of global fields in I/O
- Parallelization of I/O, hierarchical gather communication
- Performance improvement of GRIB2 I/O (uses ECMWF’s GRIB API)
- Later on: further improvement of compute scaling, e.g. by optimizing the domain decomposition, task placement, asynchronous halo communication

Conclusions

- **The computational efficiency and scalability constitute a major improvement over the hydrostatic GME**
- **Pushing the upcoming operational configuration (13 km, L 90) to the scaling limit requires a bigger machine than currently available at DWD**
- **Main issues to be solved in the near future: memory scaling, optimization and parallelization of I/O**
- **Further improvements of computational performance and scalability are less urgent**

Thank you for your attention!

Any questions?