ICOMEX: Icosahedral-grid Models for Exascale Earth System Simulations

Early results of the G8 Exascale Projects

Thomas Ludwig - *University of Hamburg/DKRZ*, **Günther Zängl** - *Deutscher Wetterdienst*,

Masaki Satoh - ***University of Tokyo***, **Hirofumi Tomita** - ***Japan Agency for Marine-Earth Science and Technology***,

Leonidas Lindarkis - *Max Planck Institute for Meteorology*, **John Thuburn** - *University of Exeter*,

Thomas Dubos - *École polytechnique*

Overview

- Introduction
- Scientific objectives, progress and recent results
  - WP1: Model intercomparison and evaluation (PIs Masaki Satoh, Hirofumi Tomita)
  - WP2: Abstract model description scheme (PI Leonidas Linardakis)
  - WP3: GPUs for atmospheric models (PI Thomas Dubos)
  - WP4: Implicit time integration schemes (PI John Thuburn)
  - WP5: Parallel internal postprocessing (PIs Thomas Dubos, John Thuburn)
  - WP6: Parallel I/O (PI Thomas Ludwig)
  - WP7: Collaboration with vendors (PI Thomas Ludwig)
- Project coordination (PI Günther Zängl)
- Summary

Introduction

- What is ICOMEX?
  - Consortium of four international model development groups focusing on icosahedral-grid Earth system models (NICAM, ICON, MPAS, DYNAMICO)

- Main strategic goals of ICOMEX
  - Select a few key issues relevant on the path towards Exascale computing
  - Develop – as far as possible – generic solutions for these issues
  - These solutions are first developed / tested in one of the modeling systems participating in the program (NICAM, ICON, MPAS, DYNAMICO)
  - In the final project phase, the solutions are also made accessible to the project partners and subsequently to the scientific community

WP 1: Model intercomparison and evaluation

PI: Masaki Satoh - *University of Tokyo*,

Hirofumi Tomita - *Japan Agency for Marine-Earth Science and Technology*

Model Inter-comparison

To provide basic information for the other groups

- **Computational aspects**
  - Performance in one node
    - Sustained / peak performance ratio [%]
  - Performance over nodes
    - Weak and strong scalability
- **Scientific aspects**
  - Numerical error, convergence, climatological behavior in Aqua-planet experiments, etc.

To exploit synergy effects

- Regularly intercomparing the developing model codes on the wide variety of computing platforms available to the project partners

The target experiments

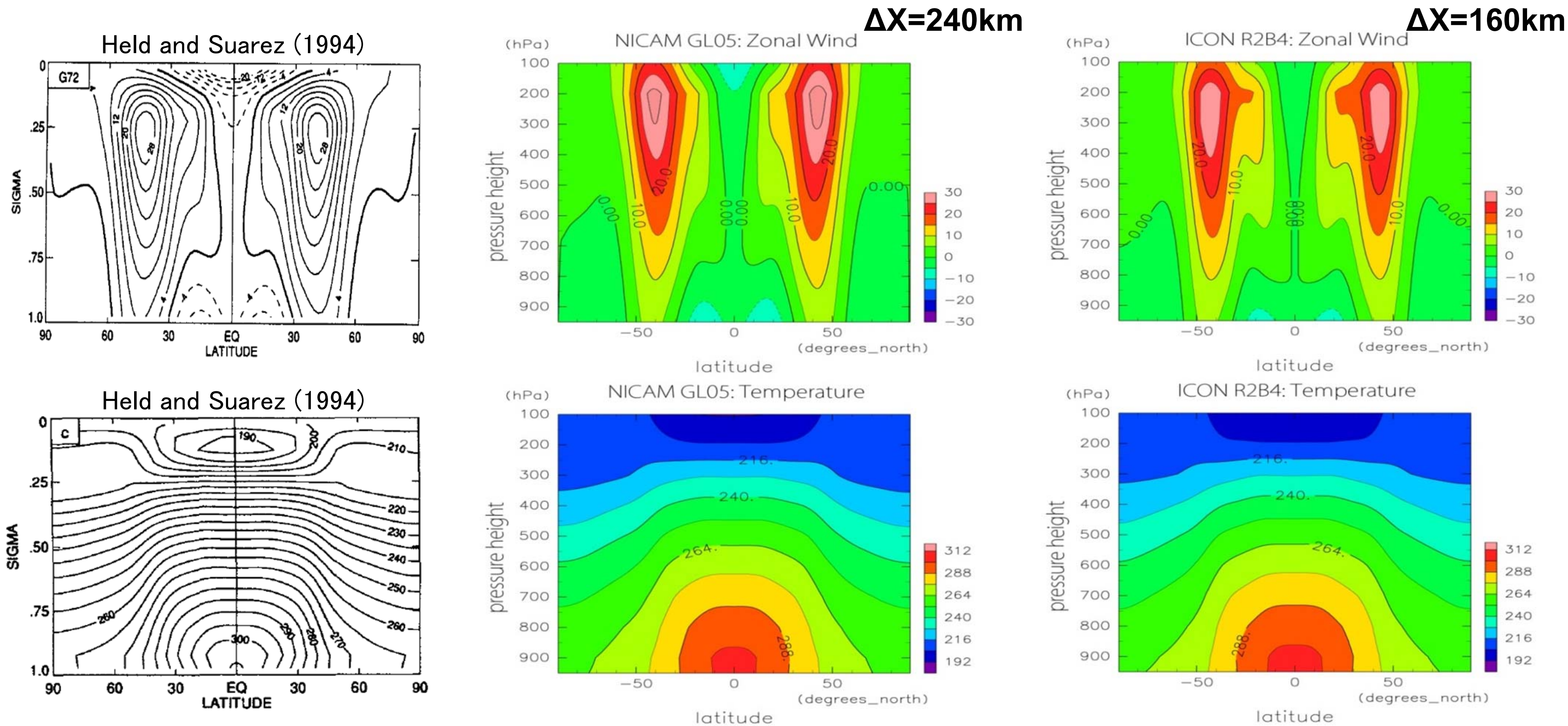
- **Deterministic test:**
  - Baroclinic wave test (Jablonowski and Williamson, 2006),
    - Resolution: 240km (glevel5 in NICAM) ~ 30km (glevel8)
- **Statistical test:**
  - Held & Suarez (1994) Test Case
    - Wave activity intensity with the same resolution
    - Check of conservation, effective model resolution, energy spectrum
  - Multi-year Aqua-planet studies including full physics (Neale and Hoskins, 2000)
    - Reveals behavior of physics parameterizations and quality of physics-dynamics coupling
    - Results usually strongly depend on cumulus parameterization
- **A 30-year AMIP run**

  (experiment 3.3 of the CMIP5 experimental suite)

∗ First evaluations for as-is codes:
NICAM, ICON, MPAS, DYNAMICO

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Preliminary Result
Held and Suarez (1994)

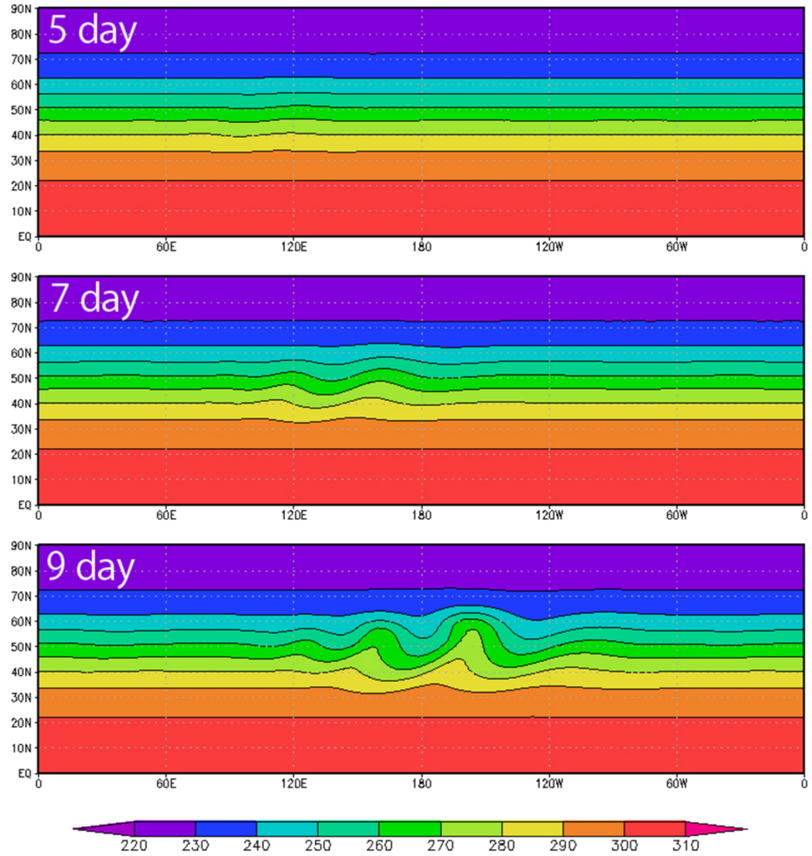


Integration Period: 1300 days (first 300 days are for spin-up)
Platform: Intel Xeon (Westmere) Cluster

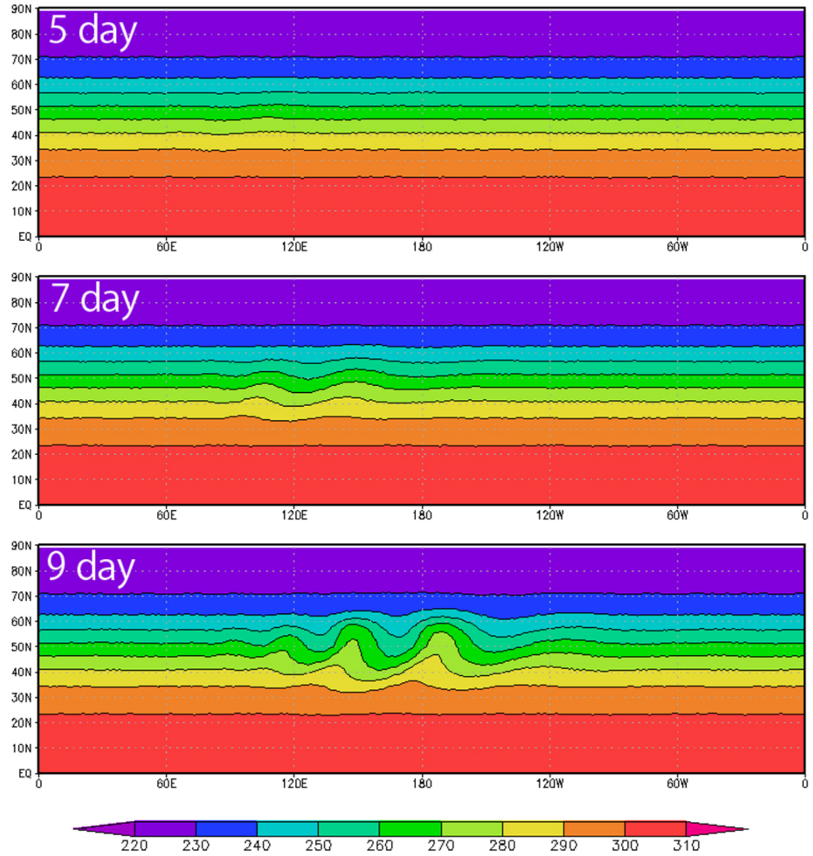
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Preliminary Result
Jablonowski Baroclinic Wave Test

ΔX = 120km NICAM GL06RL01



ΔX = 80km ICON R2B05



Temperature at 1.5km height, day 11

Platform: Intel Xeon (Sandy bridge) Server

Computational Performance on the K computer

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Grid Level	Resolution	PE	DT (sec)	for 10 days	Efficiency
5	240 km	5	1200	3 min	4.8%
6	120 km	20	600	4 min	5.0%
7	60 km	40	300	13 min	5.4%
8	30 km	160	150	26 min	5.3%
9	14 km	640	75	53 min	5.4%
10	7 km	2560	36	100 min	5.3%
11	3.5 km	5120	18	450 min	5.9%
12	1.75 km	5120	9	3000 min	6.0%

- Test case is Jablonowski-Williamson baroclinic wave.
- Number of vertical Layers is 40 levels with constant 600m distance.
- These performances were measured including initialize and data I/O sequences.

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Summary and Future

The ICOMEX project: Synergy effects are needed to push the existing icosahedral-grid models towards exascale computing.

- Iterative identification and mitigation of performance bottlenecks
- NICAM and ICON have already been run on a variety of different platforms

Held-Suarez test case was carried out as a statistical test using NICAM and ICON.

Baroclinic wave test was carried out as a deterministic test.

- Reference results at even higher resolution are needed to conduct convergence tests and calculate errors

NEXT…

- Try to run the other participating models, MPAS and DYNAMICO, on the same platforms
- APE test case - consistency of physical Processes is one of the difficulties.

WP 2: Abstract model description scheme

PI: Leonidas Lindarkis - *Max Planck Institute for Meteorology*

ICON DSL: Overview

Goals:

- Abstract description of arrays/loops by extending Fortran into a DSL.
- Use a parser to create pure Fortran code.

Targets:

- Performance
  - Generate “architecture depended” memory access patterns
  - Facilitate architecture specific optimizations (vectorization)
- Explore the possibility to express parallelization uniformly
- Software Engineering: Express models in a more “natural way”
  - Improve productivity, code readability, robustness, maintenance

ICON DSL data abstraction

```
SUBROUTINE div3d( vec_e, ptr_patch, ptr_int, div_vec_c, …)

! Define where the variable lives on the grid, instead of dimensions
REAL(wp), ON_EDGES_3D, INTENT(in) ::  vec_e
REAL(wp), ON_CELLS_3D, INTENT(inout) :: div_vec_c
INTEGER,  CELLS_CONNECT_TO_EDGES,  POINTER :: iidx, iblk
...
 DO jc = i_startidx, i_endidx
       DO jk = slev, elev
!The parser reorders the indexes according to architecture-specific rules
      div_vec_c(jc,jk,jb) =  &
           vec_e(iidx(jc,jb,1),jk,iblk(jc,jb,1)) * ptr_int%geofac_div(jc,1,jb) + &
           vec_e(iidx(jc,jb,2),jk,iblk(jc,jb,2)) * ptr_int%geofac_div(jc,2,jb) + &   …
      ENDDO
 ENDDO

END SUBROUTINE div3d
```

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

ICON memory access patterns

```
! System A (Vector machine)
  DO jc = i_startidx, i_endidx
   DO jk = slev, elev
    div_vec_c(jc,jk,jb) =  &
     vec_e(iidx(jc,jb,1),jk,iblk(jc,jb,1)) * ptr_int%geofac_div(jc,1,jb) + &
     vec_e(iidx(jc,jb,2),jk,iblk(jc,jb,2)) * ptr_int%geofac_div(jc,2,jb) + &
     vec_e(iidx(jc,jb,3),jk,iblk(jc,jb,3)) * ptr_int%geofac_div(jc,3,jb)

! System B (Cache-based machine)
DO jc = i_startidx, i_endidx
    blk1 = iblk(jc,jb,1)
    idx1 = iidx(jc,jb,1)
    blk2 = iblk(jc,jb,2)
    idx2 = iidx(jc,jb,2)
    blk3 = iblk(jc,jb,3)
    idx3 = iidx(jc,jb,3)

    DO jk = slev, elev
     div_vec_c(jk,jc,jb) =  &
      vec_e(jk, idx1, blk1) * ptr_int%geofac_div(1,jc,jb) + &
      vec_e(jk, idx2, blk2) * ptr_int%geofac_div(2,jc,jb) + &
      vec_e(jk, idx3, blk3) * ptr_int%geofac_div(3,jc,jb)
```

Inner loop on inner index, no indirect indexing

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

First performance results

Experiment R2B4 on an IBM Power6

Cores	32	64	128	192
No DSL time/(cell*iter)	1.574e-06	7.012e-07	3.574e-07	2.777e-07
DSL time /(cell*iter)	1.390e-06	6.008e-07	3.230e-07	2.504e-07
No DSL iterations/sec	635479	1426037	2798150	3601217
DSL iterations/sec	719527	1664402	3096318	3993947
Speed-up	13%	17%	11%	11%

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

ICON DSL loop abstraction (in progress)

- Math-like syntax using elements/subsets)

```
subset,  on_cells_3D :: all_cells
element,  on_cells_3D :: cell
element, edges_of_cell_2D :: edge

for cell in all_cells do
   div_vec_c(cell) = 0.0_wp
   for edge in cell%edges do
         div_vec_c(cell) =   div_vec_c(cell)  + vec_e(edge) * ptr_int%geofac_div(edge)
   end do
end do

! compact sum
for cell in all_cells do
   div_vec_c(cell) = sum[in cell%edges] (vec_e * ptr_int%geofac_div)
end do
```

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

WP 3: Feasibility study for using GPUs for atmospheric models

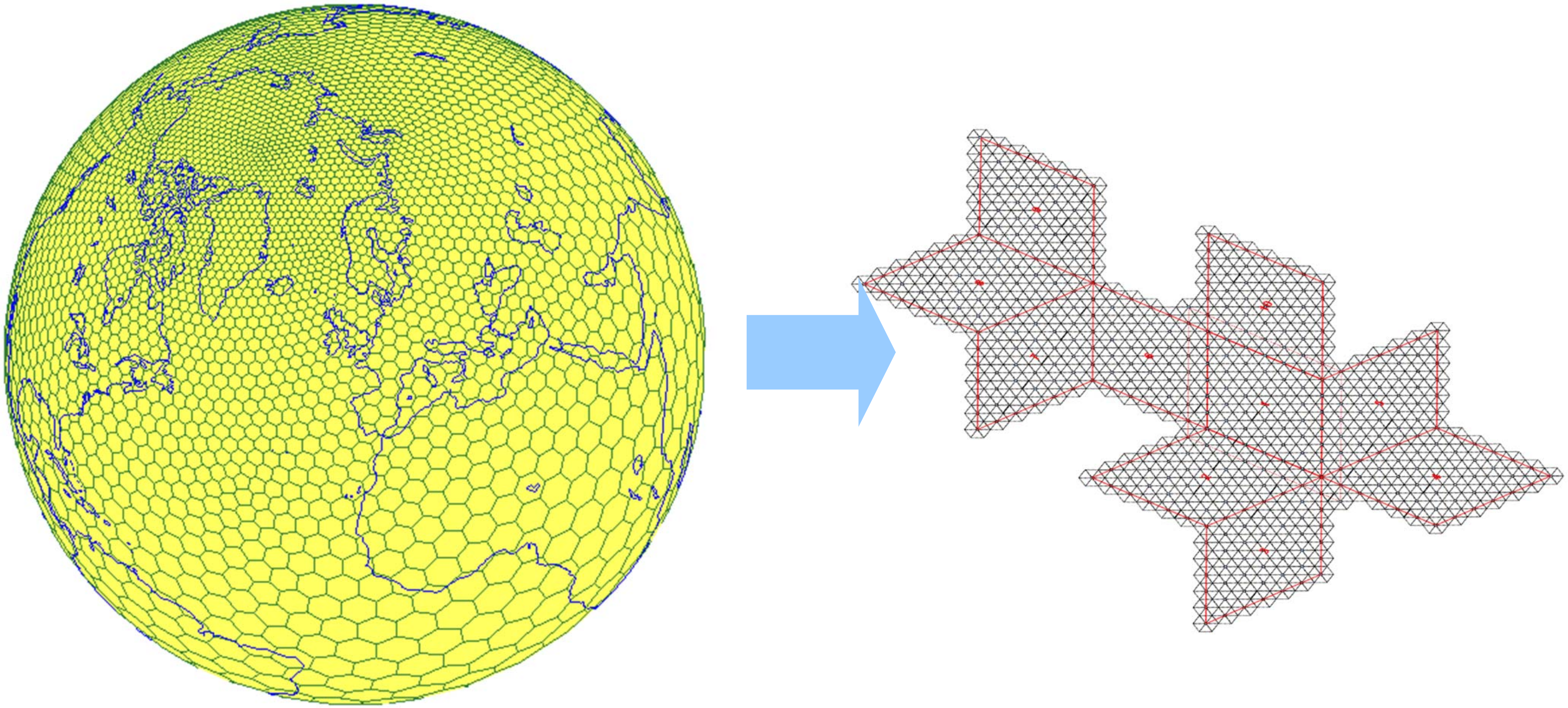
PI: John Thuburn - *University of Exeter*,

Thomas Dubos - *École polytechnique*

Scientific objectives

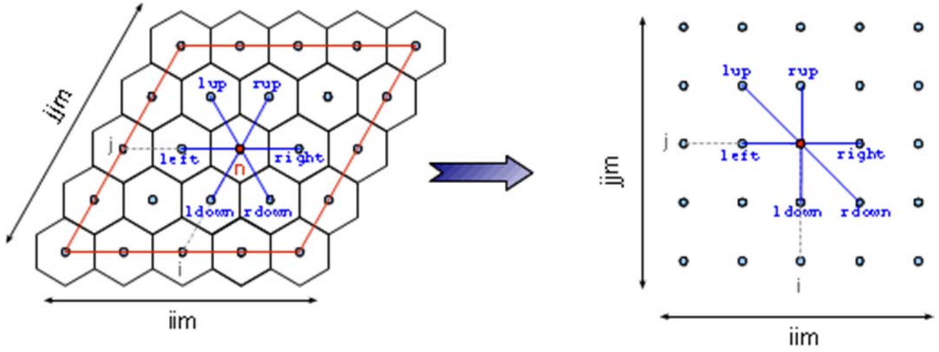
- How much GPU performance can be extracted by:
  - Low-level implementation (CUDA, OpenCL)
  - High-level implementation (HMPP, OpenAcc)
- Assess performance with DYNAMICO core
  - ~1000 performance-critical lines
  - Recent GPU-friendly design
  - Hydrostatic but compute patterns representative of ICON, NICAM, MPAS
- Identify efficient programming patterns
  - Expressed at high-level
  - Suitable for multi-component, open modelling systems

DYNAMICO: a hydrostatic core on a structured icosahedral grid

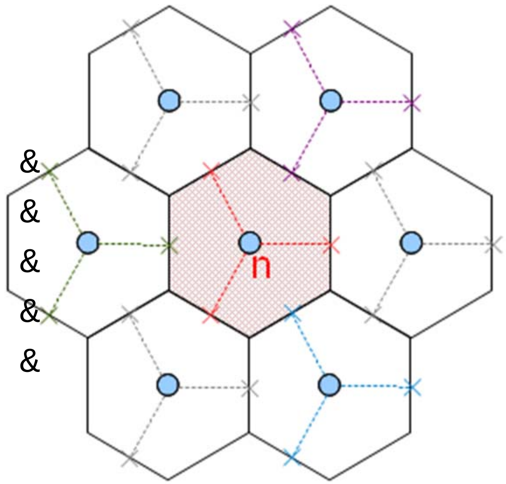


Structured data layout

- Data stored in rectangular arrays
- Direct access to neighbours
  - via constant offsets
- No special case for pentagons
  - handled by metrics
- Vertical direction in outer loops




```
DO j=jj_begin,jj_end
  DO i=ii_begin,ii_end
    n=(j-1)*iim+i
    dhi(n)=-1./Ai(n)*(ne(n,right)*ue(n+u_right)*le(n+u_right) + &
                      ne(n,rup)*ue(n+u_rup)*le(n+u_rup) +         &
                      ne(n,lup)*ue(n+u_lup)*le(n+u_lup) +         &
                      ne(n,left)*ue(n+u_left)*le(n+u_left) +      &
                      ne(n,ldown)*ue(n+u_ldown)*le(n+u_ldown) + &
                      ne(n,rdown)*ue(n+u_rdown)*le(n+u_rdown))
  ENDDO
ENDDO
```

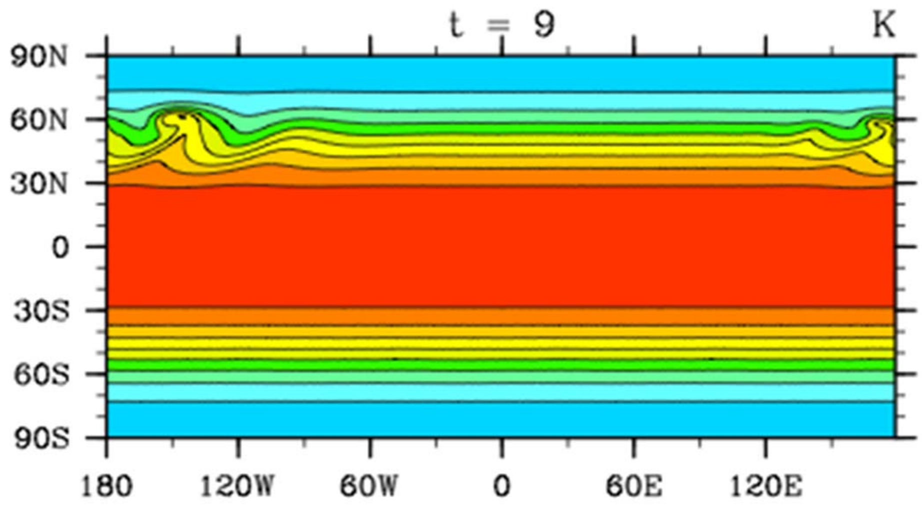


Status and plans

- DYNAMICO core
  - No-physics dry core, ready summer 2012 (was planned spring 2012)
  - Passed short-term test cases during DCMIP workshop in Aug. 2012 (sufficient for this WP)
  - Code streamlining under way before GPU experimentation

- GPU implementations/assessment
  - Start Jan. 2013 on NVIDIA hardware
  - Low-level : CUDA Fortran
  - High-Level : OpenACC

- Extra plans if time/resources allow
  - Couple with physics parameterizations
  - Experiment with XeonPhi hardware



WP 4: Implicit time integration schemes

PI: John Thuburn - *University of Exeter*

Implicit time integration schemes

Goals:

- Stable and accurate scheme allowing longer time steps
- Efficient and scalable elliptic solver for icosahedral grids

$$\Phi^{n+1} - \Phi^{n} + \Phi^{*}\overline{\nabla \cdot \mathbf{u}} + (NL) = 0$$

$$\mathbf{u}^{n+1} - \mathbf{u}^{n} + \overline{\nabla \Phi} + (NL) = 0$$

where $\overline{\psi} = \alpha \psi^{n+1} + (1-\alpha)\psi^{n}$

Requires less ad-hoc dissipation than (split) explicit or HEVI schemes

But implicit schemes require the solution of an elliptic problem for the unknowns at each time step

$$\alpha^2 \Delta t^2 \nabla \cdot (\Phi^* \nabla \Phi^{n+1}) - \Phi^{n+1} = \mathrm{RHS}$$

Can we solve such problems efficiently enough on massively parallel machines to make implicit time schemes worthwhile?

Explore multigrid methods to solve the elliptic problem

- Unlike Krylov subspace methods (such as CG), there is only local communication at each iteration.

- The elliptic problem has an intrinsic length scale $L = (\Phi^*)^{1/2}\,\Delta t$
We only need to coarsen until $\Delta x \sim L$, typically 3-4 levels.
Processors don't run out of work.

- A Jacobi smoother is effective and conservative, and keeps the possibility of strong bit reproducibility.

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Some initial results

For a single multigrid sweep on a typical test problem...

Underrelax param	Residual	Error
0.5	1.22E-3	3.44E-4
0.6	7.12E-4	2.30E-4
0.7	4.52E-4	1.62E-4
0.8	3.02E-4	1.18E-4
0.9	2.09E-4	8.80E-5
1.0	1.52E-4	6.76E-5

Number of levels	Residual	Error
1	0.56	0.56
2	1.29E-5	1.28E-5
3	1.12E-7	1.00E-7
4	1.12E-7	1.00E-7
5	1.12E-7	1.00E-7

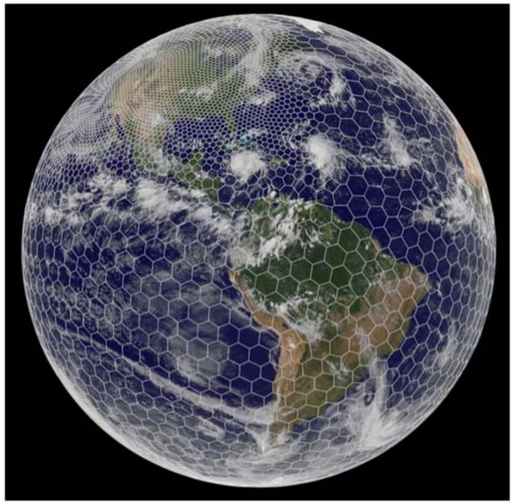
On a hexagonal Voronoi grid, the optimal under-relaxation parameter is close to 1

$\Delta x \sim L$ is a good criterion to determine the number of levels needed
(Here 3 is enough)

Next steps

Parallel implementation and optimisation (e.g. duplicating flops to reduce communication)

Restriction and prolongation operators for locally refined grids



Implementation within MPAS – a new atmospheric model developed at NCAR / Los Alamos

WP 5: Parallel internal postprocessing

PI: John Thuburn - *University of Exeter*,

Thomas Dubos - *École polytechnique*

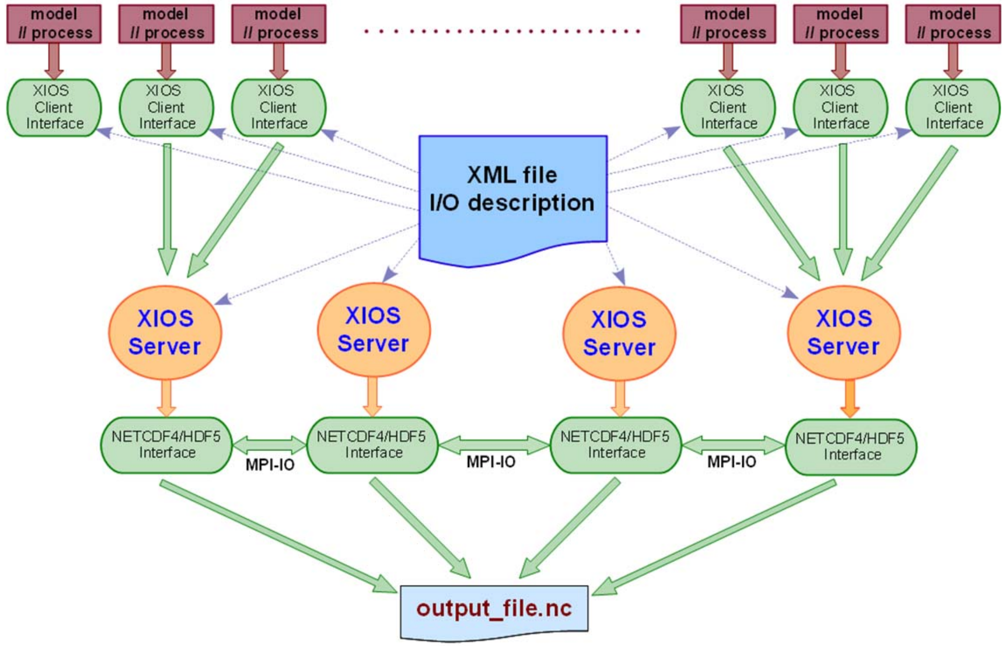
Objectives

- Major bottleneck: massive amounts of data to be produced by exascale climate simulations
- Scientific usage does not require all data at high spatio-temporal resolution
- Approach: reduce outputs by performing common post-processing on-line
  - Temporal average / min / max
  - Extraction of region of interest (clipping)
  - Grid coarsening, transfer to user-friendly grids (lon-lat)
- Design must be parallel from the start
- Approach : start with XIOS (XML I/O Server) and develop missing key-functionalities

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

XML I/O Server

- Parallel design
- User-friendly (XML) description of outputs
- Features temporal average/min/max, clipping
- Production-grade tool used by NEMO ocean model
- Needs extension to unstructured grids
- ICOMEX development: flexible interpolation tool to/from unstructured spherical meshes

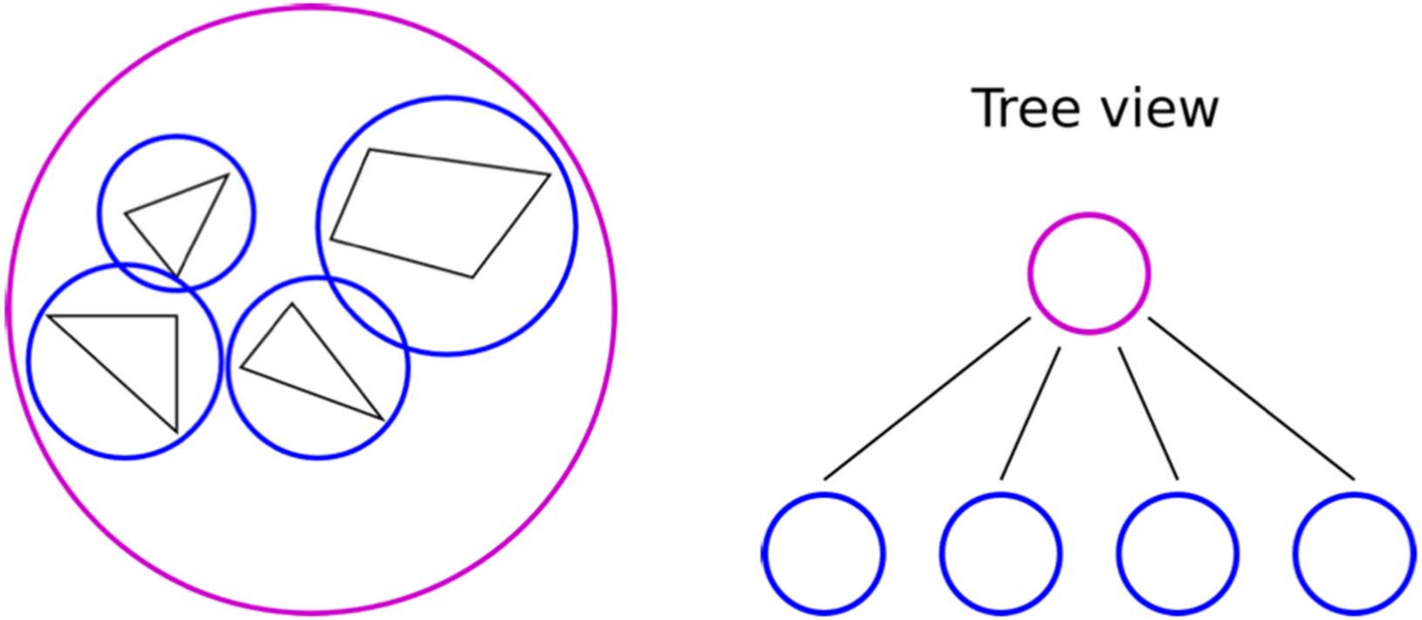


Interpolation to/from spherical meshes

- Desired properties
  - Arbitrary spherical meshes
  - Exactly conservative
  - Second-order accuracy
  - Explicit and local : no global linear system to solve, no iteration
  - Algorithmic efficiency
  - Parallelism of performance-critical parts
- Criteria not met by existing libraries: Jones (1999) (SCRIP), Farrell et al. (2005), Ullrich et al. (2009)
- Evaggelos Kritsikis hired in March 2012
  - Conservation guaranteed by using a supermesh and careful treatment of sphericity
  - Supermesh construction based on fast tree-based search
  - Tree construction costs O(N log N) for each mesh
  - Accuracy obtained by finite-volume style piecewise linear reconstruction

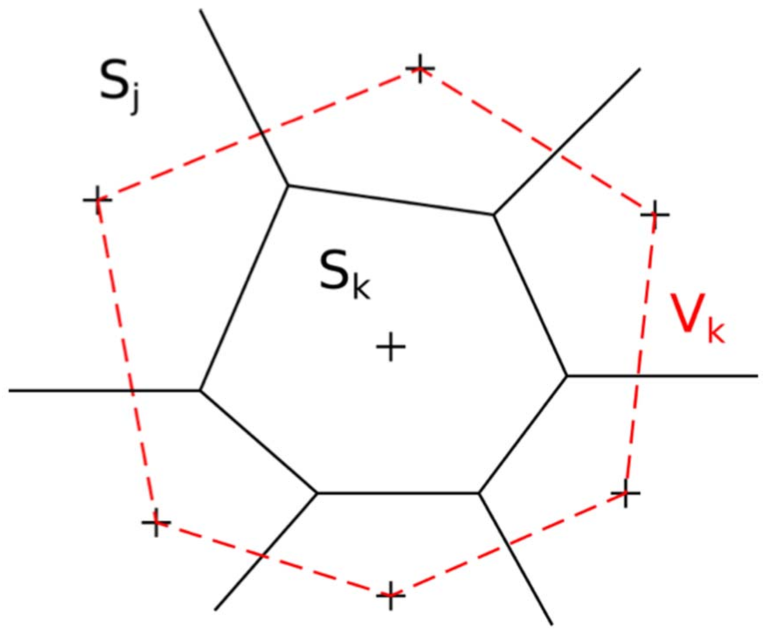
Interpolation to/from spherical meshes

- Tree is computed once per simulation (pre-processing step)
- Tree view of unstructured meshes
  - Cells are inserted in a hierarchy of «nodes»
  - Nodes are characterized by their circumcircle => fast search algorithm
  - Nodes are split when they become too large

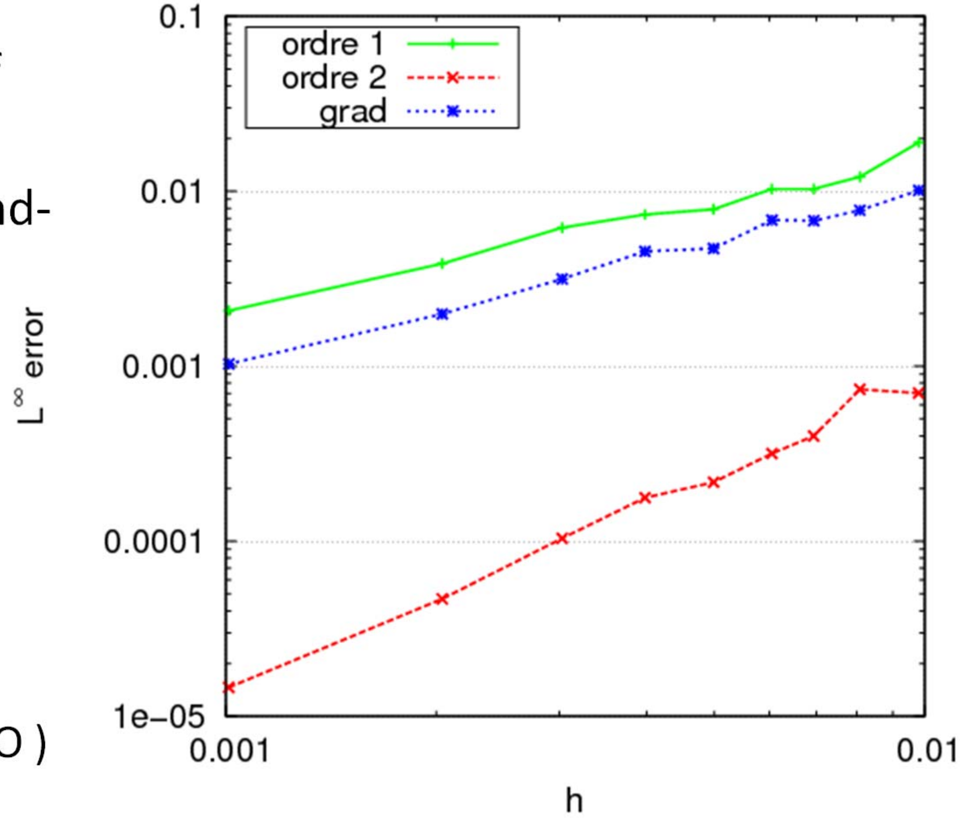


Interpolation to/from spherical meshes

- Piecewise linear reconstruction
  - Data locality: only nearest neighbours
  - Explicit first-order gradient reconstruction
  - Second-order one-point quadrature formula at centroids of supermesh cells
- Interpolation weights are computed once per simulation (pre-processing)
- Data locality => inherently parallel interpolation



- Tree construction time closer to O(N) than theoretical O(N log N)
- Conservation is exact within round-off error
- Piecewise-linear interpolation is second-order accurate as expected



- Plans for 2012/2013
  - Extend XIOS to unstructured grids
  - Integrate interpolation with XIOS
  - Provide as standalone library
  - Interpolation of vector fields defined on staggered grids (MPAS, ICON, DYNAMICO)

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

WP 6: Parallel I/O

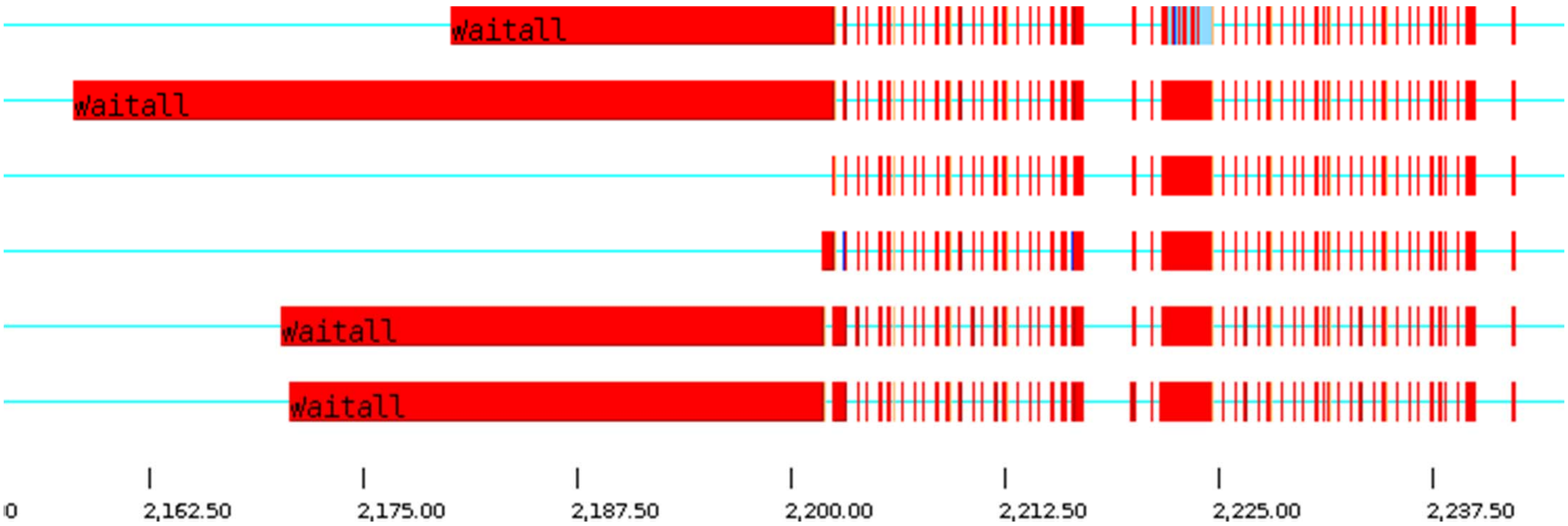
PI: Thomas Ludwig - *University of Hamburg/DKRZ*

Scientific objectives

- Analysis of access patterns and behavior of
  - Applications
  - I/O middleware
  - System
- Assess performance on the different I/O layers
- Localization of bottlenecks in hardware and software
  - By comparison of theoretical and measured performance
- Develop optimization strategies
- Creation of a scalable benchmark which mimics model I/O on these layers
- The ICON model is clear application driver
  - Analysis and modifications to middleware and system help everyone

Subgoal: Tracing of MPI and I/O routines

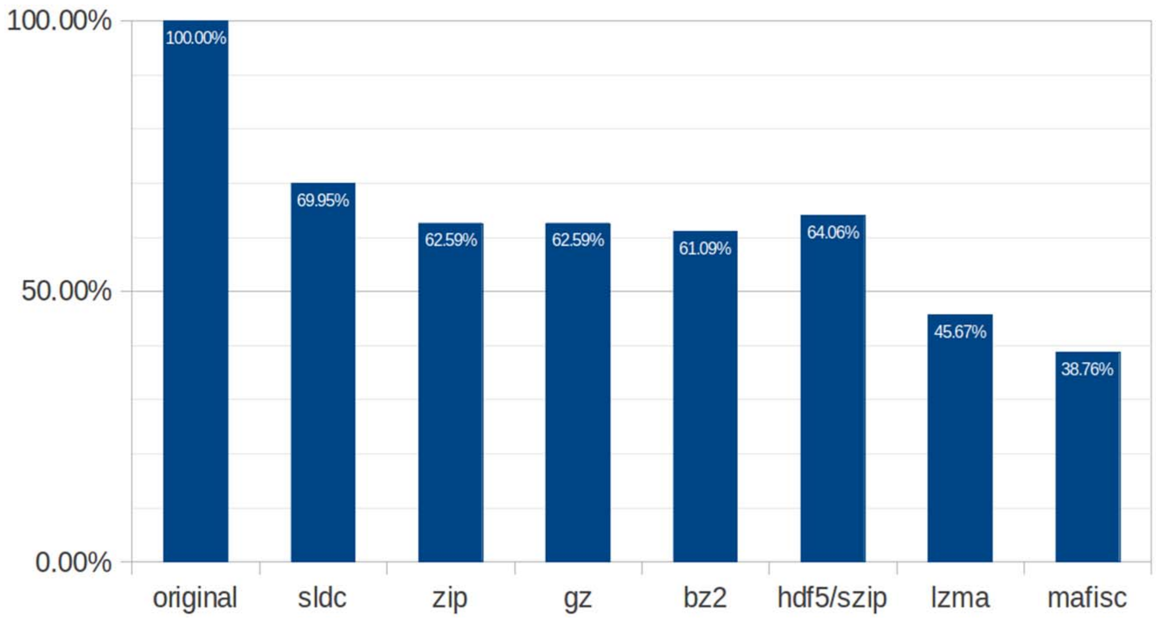
- Instrumentation of MPI and (internal) I/O routines
- Helps during analysis



Mafisc Data compression

- General lossless compression for scientific data
- Highest compression rate in tests (16% better than uninformed compression)
- Improved on-disk format for long-term archival
- Patch for HDF5

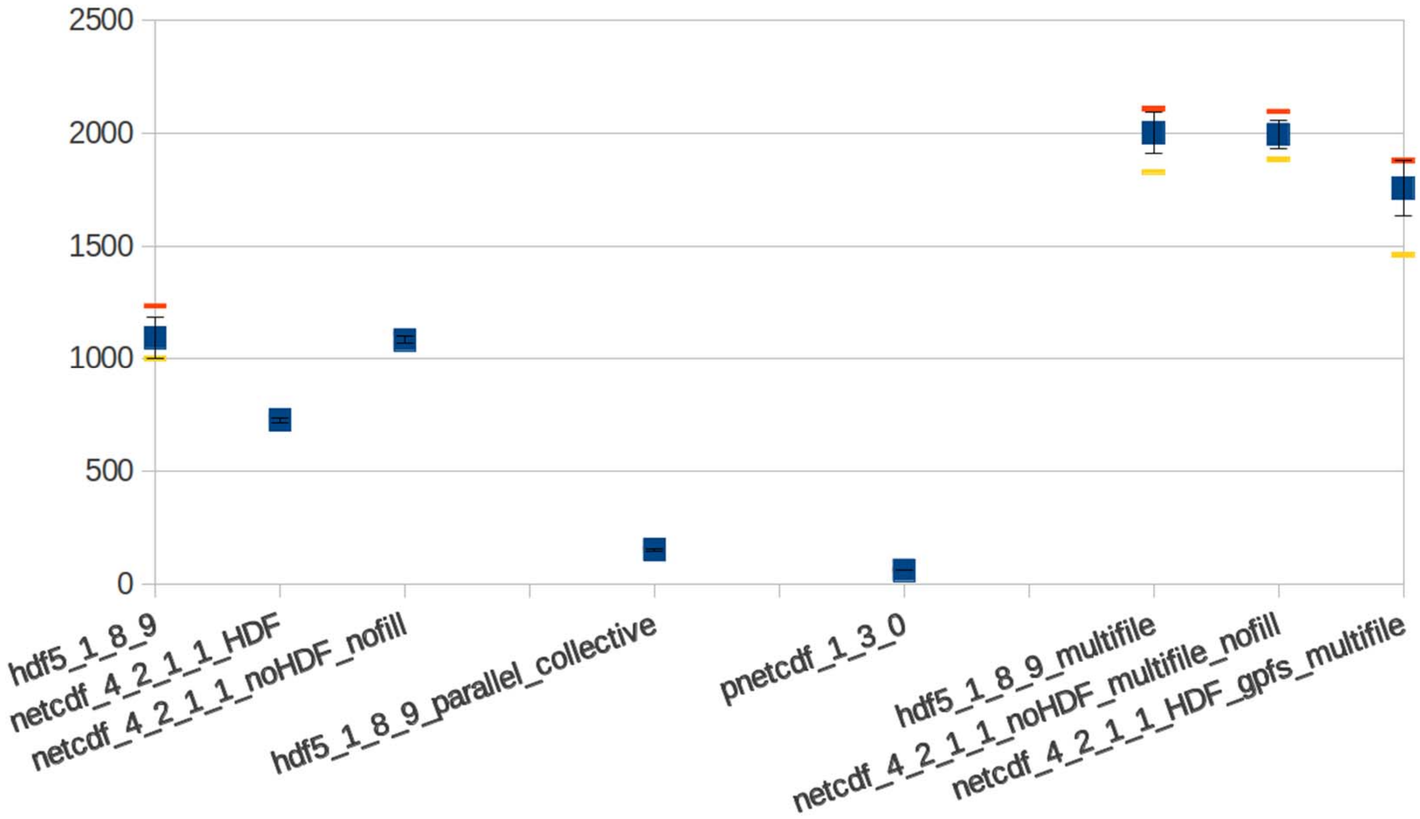
  to support MAFISC



Benchmarks

- To localize bottlenecks of the current ICON-I/O, a set of (simple) benchmarks were written
  - Resembles current ICON–Output
  - Ported to HDF5, NetCDF, and pNetCDF
  - Versions for sequential, parallel and parallel multifile access
  - Applied to ten different library builds
- Resulting performance spread across three orders of magnitude
  - Identified several performance issues in the interplay between DKRZ‘s GFPS file system, NetCDF, pNetCDF and HDF5
- In progress: Parameterizable benchmark

Selected benchmarking results



Optimization

Localized performance issues in NetCDF and pNetCDF

- Patch for NetCDF to improve performance:
  - Available at http://wr.informatik.uni-hamburg.de/research/projects/icomex/cachelessnetcdf
- Improves performance by a factor of 3.2

pNetCDF issue was found in the underlying MPI/IO−library

- This library is not open source.
  - We can neither investigate further nor can we develop a fix for it.
- The vendor has been informed.
  - Lengthy discussion
  - Necessary modifications to pNetCDF will be made to extract performance

WP 7: Collaboration with vendors

PI: Thomas Ludwig - *University of Hamburg/DKRZ*

Goals

- This WP addresses co-design and knowledge transfer
- Vendors => ICOMEX consortium
  - Guidance on efficient code level structure, especially for future platforms
  - Allows climate codes to be ready for future technology
- ICOMEX consortium => Vendors
  - Information on specific needs of climate codes
  - Allows vendors to develop products to address these needs

Current status

- The codes, documentation of the consortium are available for vendors on our Redmine
- So far: Discussions of demands occur mainly during project meetings
- Communication of I/O bottlenecks and patches
  - NetCDF–patches and issues communicated to the developers
  - Detailed description of MPI/IO–issue communicated to IBM
  - MAFISC source code communicated to NetCDF and HDF5 developers
    - Need for external filter modules communicated to HDF5 developers
    - Additional patch for an external module loader

In progress / Future work

- Communication of benchmark code (I/O and model cores)
  - Allow vendors to test compiler developement, scalability etc… with stripped applications
- Setting up of a forum to allow more direct communication
  - Integration into ENES Portal anticipated
  - Collaboration with DKRZ for sustained activity that helps earth-science in the long-term
  - Initial conceptual sketch will be developed within ICOMEX

Project coordination

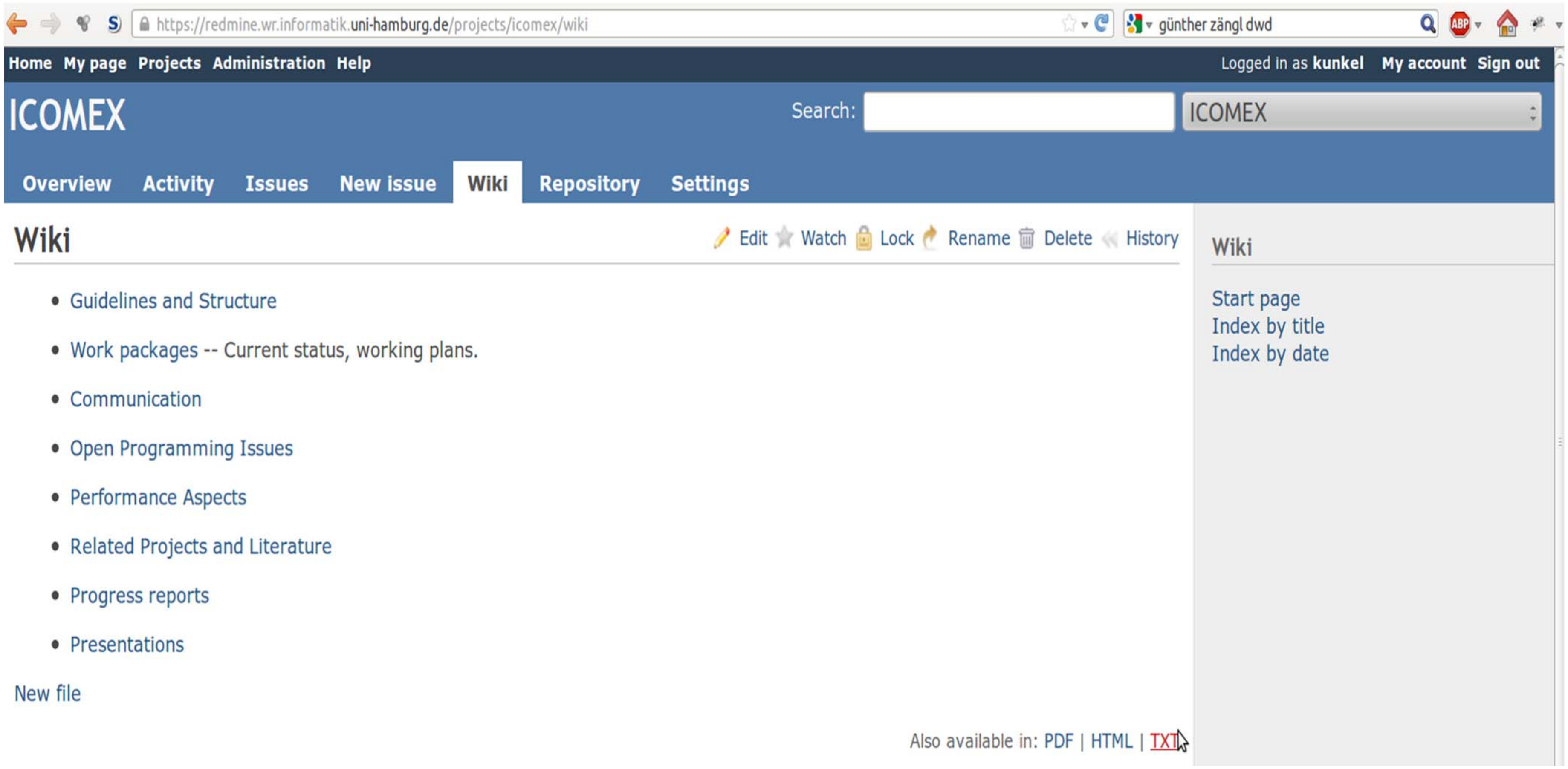
Günther Zängl - *Deutscher Wetterdienst*

Communication / Coordination

- Available communication tools
  - Redmine project management
    - Wiki, Issue trackers
  - Mailing lists: internal and external including vendors
  - Phone conferences have not been found necessary so far
    - Different time zones leave only narrow time slots; mailinglists and Redmine provide sufficient communication channels
  - Provides daily updated mirror of svn servers used for model development
    - Currently used for ICON only, mirrors for other models are in progress.
- Annual project meetings at DWD

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Redmine



Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Progress

- Most projects started late due to difficulties in recruiting appropriate scientists
  - Last project scientist for WP3/WP5 started only in summer 2012!
  - Progress is within schedule relative to the start date of for most WPs
  - WP1 already started 6 months before the official start of ICOMEX; other WP's were not yet ready at that time
- Difficulties with supercomputers
- Modular project structure mitigates fluctuations in progress speed/starting dates
  - WP2 to WP6 do not have much interdependencies within the main project phase

Path to extreme scale

- ICOMEX enables research in key issues required for scalable earth-system models
  - Alternative algorithms (implicit solvers)
  - Improved code quality and portability (DSL)
  - Scalable I/O
  - Model quality
  - Future architectures
- ICOMEX allows to conduct a combination of basic and applied research
- ICOMEX has connections to broader needs of the scientific research community
  - Generic solutions are made available and can be adjusted by different fields

Benefits of international cooperation

- Access to international resources are possible (e.g. evaluation on K-computer)
- Rapid information exchange across countries
  - Spreading best-practice of hardware/software
  - Consortial members of each country are typically integrated in national efforts
  - Multipliers of knowledge
- We know that there is more potential than currently used
  - We seek for approaches and strategies to exploit the potential better

Summary

- What do we have achieved so far?
  - Enhanced international collaboration among global model developers
  - Work on selected key problems on the way to Exascale computing has started
  - Several performance bottlenecks have already been identified; solutions have been developed for part of them

- Early conclusions – how to proceed towards Exascale computing?
  - Much more resources will be needed to thoroughly prepare our models for the upcoming challenges, including extensive participation of experienced senior scientists
  - Access to supercomputing resources for model development / optimization should be simplified