ICOMEX

ICOsahedral-grid Models for EXascale Earth-system simulations

2nd IS-ENES workshop on HPC for climate models

Günther Zängl

31.01.2013

ICOMEX

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

- **One of six successful projects of the G8 call issued in 2009**
- **Six work packages representing four global modeling groups**
- **Main strategic goal: address a selection of key issues for future cutting-edge computing and develop generic solutions**
- **These solutions are first developed / tested in one of the participating modeling systems and will be made avaliable to the project partners (and afterwards to the scientific community) in the final project phase**

Project partners:

Günther Zängl (Lead PI, Deutscher Wetterdienst)
Hirofumi Tomita (RIKEN/AICS)
Masaki Satoh (U. Tokyo)
Thomas Ludwig (U. Hamburg/DKRZ)
Leonidas Linardakis(MPI-M)
John Thuburn(U. Exeter/MetOffice)
Thomas Dubos (IPSL/École Polytechnique)

ICOMEX: Participating models

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

- **NICAM : Hirofumi Tomita (RIKEN/AICS), Masaki Satoh (U. Tokyo)**
- **World's first global convection-permitting simulations**
- **Structured icosahedral-hexagonal A-grid**

- **ICON : Günther Zängl (DWD), Marco Giorgetta (MPI-M)**
- **Unstructured icosahedral-triangular C-grid**
- **Two-way nesting and limited-area option**

- **MPAS : John Thuburn (U. Exeter)**
- **Unstructured icosahedral-hexagonal C-grid**
- **Special care of energy and vorticity budgets**

- **DYNAMICO : T. Dubos (IPSL/École Polytechnique)**
- **Most recent development**
- **Structured icosahedral-hexagonal C-grid**
- **So far hydrostatic version only**

ICOMEX: Work packages

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

- **WP1: Model intercomparison and evaluation (Satoh/Tomita)**
- **WP2: Abstract model description scheme / domain-specific language (Linardakis)**
- **WP3: Feasibility study for using GPUs (Dubos)**
- **WP4: Implicit solvers for massively parallel computing platforms (Thuburn)**
- **WP5: Parallel internal postprocessing (Thuburn/Dubos)**
- **WP6: Parallel I/O (Ludwig)**

plus

- **WP7: Collaboration with hardware vendors (coordinated by Ludwig)**

Specific tasks

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

WP1:

- **Intercomparison of the participating models with respect to scientific and computational aspects**
- **Detect strengths and weaknesses of each participating model; exploit synergy effects from regular intercomparison**
- **Make high-performance computing platforms of the project partners mutually accessible**
- **Test cases: Jablonowski-Williamson baroclinic wave test, Aqua-planet experiments, 30-year AMIP runs**
- **Evaluation metrics: convergence tests, climatological behaviour of APE runs, weak and strong scaling tests on various platforms**

Selected result

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

Scaling of NICAM (dycore only) on the K computer

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.

Specific tasks

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

WP2:

- **Use domain-specific language (DSL) with automatic Fortran code generation in order to optimize model code for a variety of platforms**
- **Specifically: generate “architecture-dependent” memory layout and loop orders in order to optimize cache efficiency and/or vectorization**
- **Later stage: explore the possibility to express parallelization in a more abstract way**
- **Testbed: nonhydrostatic dynamical core of ICON**

Example


```
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
```

Example


```
! System A (Vector machine)
  DO jk = slev, elev
    DO jc = i_startidx, i_endidx
      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)
```

no indirect addressing depending in inner loop index

Result

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

Domain-specific language for ICON: synthetic test for the dynamical core, 20480 cells x 35 levels

Specific issue considered here: optimization of memory layout for IBM pwr6

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	6.008e-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%

Specific tasks

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

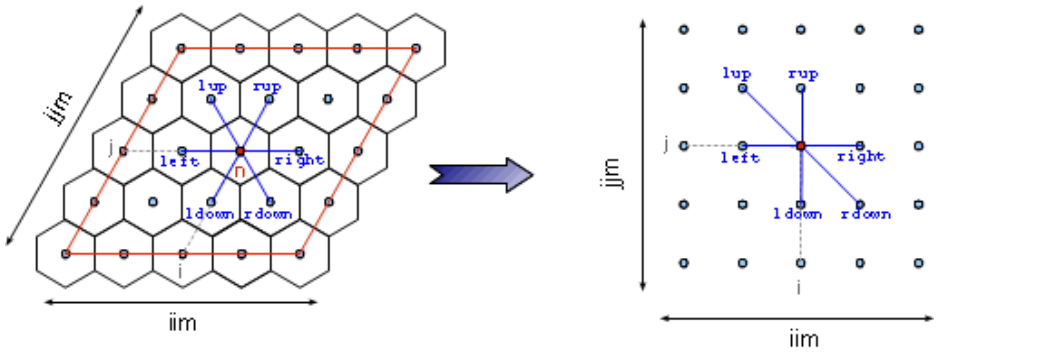
WP3:

- **GPU implementation of the hydrostatic dynamical core of DYNAMICO (besides conventional CPU implementation, which is done outside ICOMEX)**
- **Assess how much GPU performance can be extracted by:**
  - **Low-level implementation (CUDA, OpenCL)**
  - **High-level implementation (HMPP, OpenAcc)**
- **Identify efficient programming patterns**
- **Comparative testing and benchmarking**

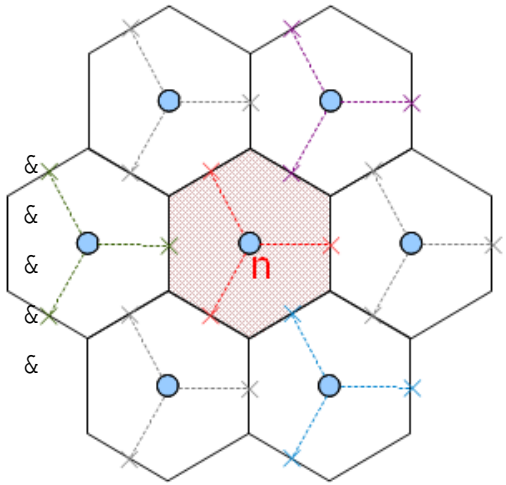
Structured data layout

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

- 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
```



Specific tasks

WP4:

- Development of an efficient and scalable multigrid elliptic solver for semi-implicit time integration on icosahedral grids

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

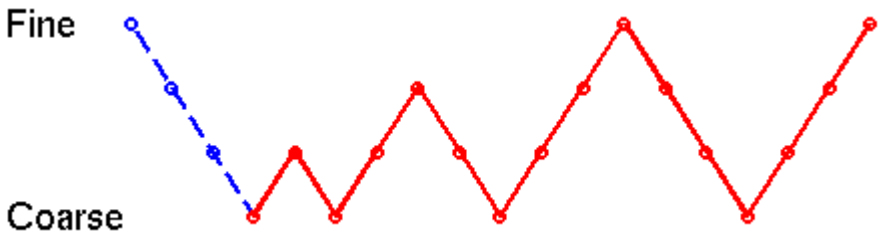
$$\mathbf{u}^{n+1} - \mathbf{u}^n + \overline{\nabla \Phi} + (NL) = 0$$

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

- Advantage of implicit time-stepping: higher numerical stability with less artificial diffusion and longer time steps
- Testbed code: MPAS

Multigrid methods

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.**

Initial result

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)

Specific tasks

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

WP5:

- Development of a parallel internal postprocessing library
- Motivation: I/O is one of the major bottlenecks on the way towards exascale computing because I/O bandwidth and disk performance will increase not as fast as compute power
- 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)
- Approach : start with XIOS (XML I/O Server) and develop missing key-functionalities
- First specific task (currently in progress): conservative remapping

Conservative remapping

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

- **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)**
- **Our approach**
  - **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**

Specific tasks

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

WP6:

- Analysis of access patterns and behavior of
  - Applications
  - I/O middleware
  - System
- Assess performance of 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

Benchmarks

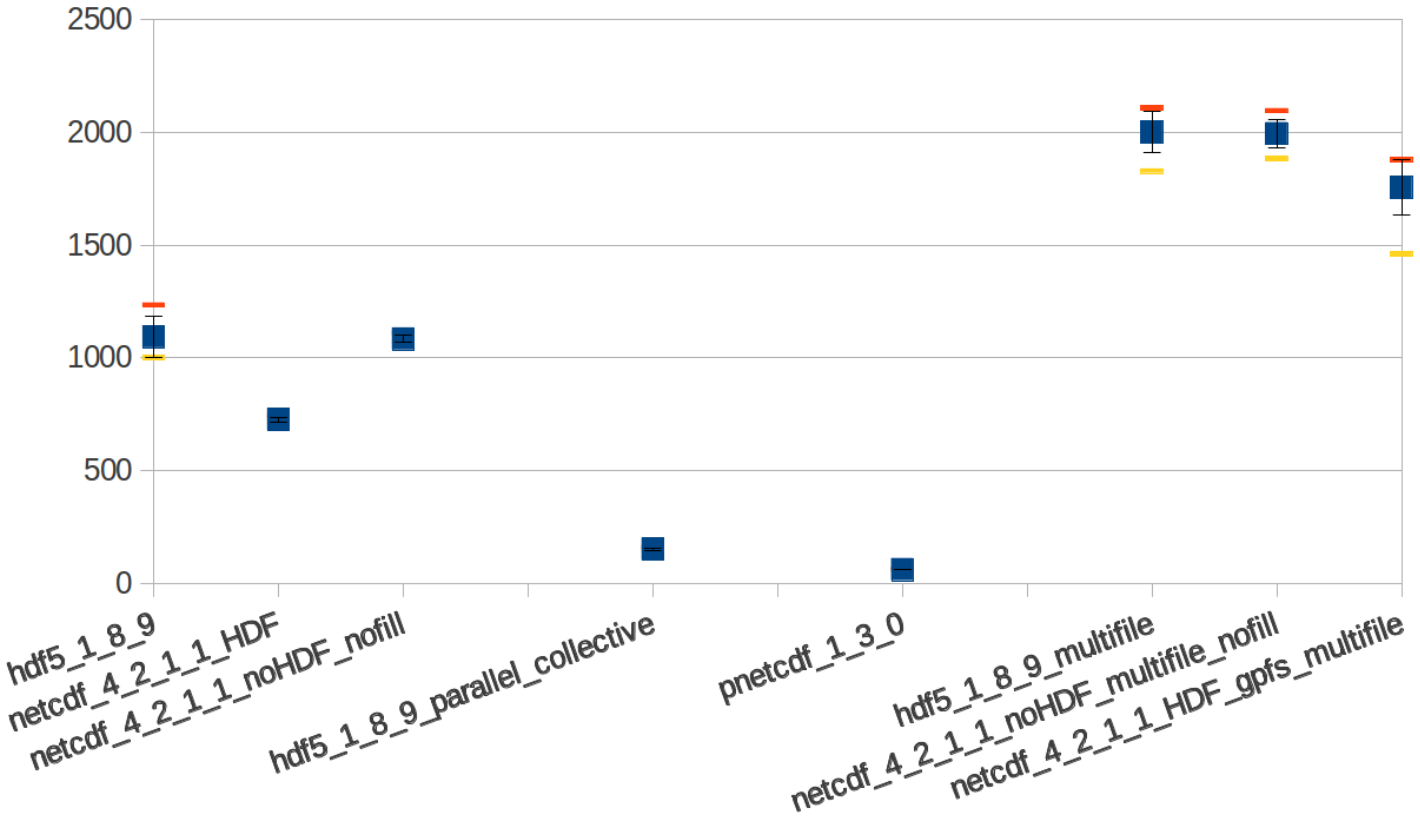
Deutscher Wetterdienst
Wetter und Klima aus einer Hand

- **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 bottlenecks on the interplay between DKRZ's GPFS file system -- NetCDF, HDF5**
- **In progress: Parameterizable benchmark**

Selected benchmark results

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

NetCDF write performance of different library builds on the IBM pwr6 @ DKRZ
in parallel runs with 4 processes (MiB/s)



Summary

- **Good work progress in most sub-projects …**
  - **given the fact that some positions could be filled only with substantial delay**
- **… but it is clear that much larger efforts are needed to really prepare our Earth-system models for Exascale computing**
- **From our own experience:**
  - **applying modern programming concepts is sometimes tedious due to compiler bugs or compilers not supporting the current Fortran standard**
  - **Extending the GRIB2 standard to support modern model design concepts like horizontally unstructured grids, generalized vertical coordinates or efficient tile-based methods for LSMs is even more tedious**