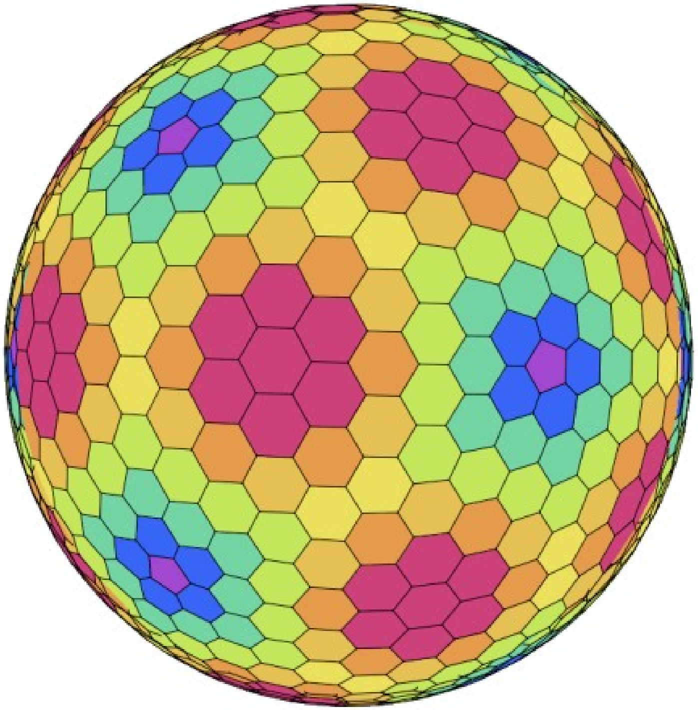
ICOMEX

ICOsahedral-grid **M**odels for
EXascale Earth System Simulations

Julian M. Kunkel

G8 Initiative – Final Review Meeting



Masaki Satoh, Hirofumi Tomita

Günther Zängl, Leonidas Linardakis, Thomas Ludwig (Julian Kunkel)

Thomas Dubos

John Thuburn

Outline

WP1: Model Intercomparison and Evaluation

WP2: Abstract Model Description Scheme

WP3: Feasibility Study for Using GPUs
(postponed / researcher left)

WP4: Implicit solvers for massively parallel computing

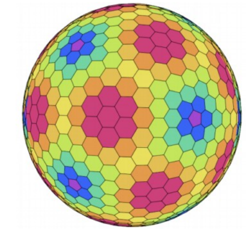
WP5: Parallel Post-Processing

WP6: Parallel I/O

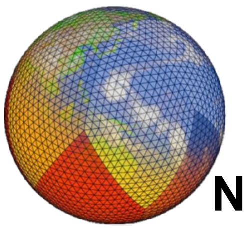
WP7: Vendor Communication

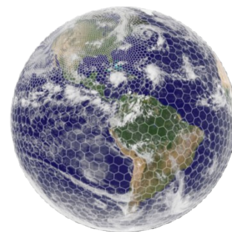
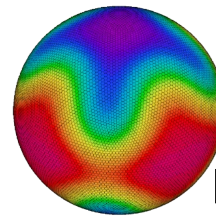
Lessions Lerned and Future Perspective

Background and Motivation



Before G8: groups work independently on different models with icosahedral grids


NICAM

ICON

MPAS

DYNAMICO

Icosahedral grids with *differences in numerics and grid structure*

- NICAM, Structured hexagonal A-grid
- ICON, Unstructured triangular C-grid
- MPAS, Unstructured hexagonal C-grid
- DYNAMICO, Structured hexagonal C-grid

Diversity and competition is beneficial to overcome hurdles, however:

- Expensive replication of effort, slow progress
- Funding structures were hurdles to effective collaboration

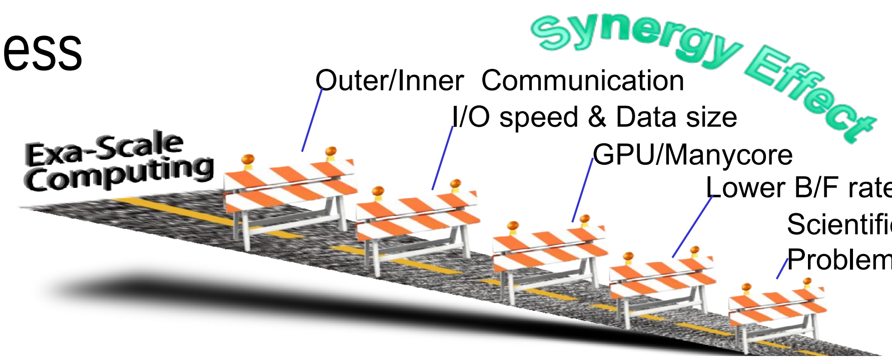
Objectives of ICOMEX

Idea: **Collaboration to solve roadblocks toward the exa-scale computing**

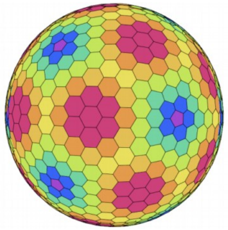
Goals: Improve computational, I/O performance and scalability

Approach

- Each group addresses a key problem and derive generic solutions
  - Exchange information and insights
- Model intercomparision
  - Helps validating correctness
- Learn best-practises



WP1: Model Intercomparison and Evaluation

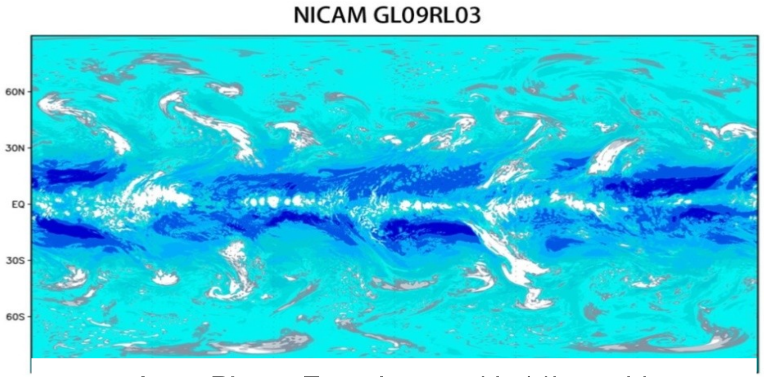


Motivation

- To exploit the synergy effects perform an comparison of the model codes
- Assess both Computational and Scientific aspects

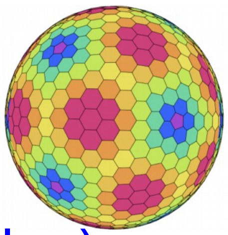
Approach

- Evaluation on the models in 4 experiments from meteorological and climatological scientific aspects.
- Performance comparison on the models from computational scientific aspects.



Aqua Planet Experiment with 14km grid space; an example of test case with full physics

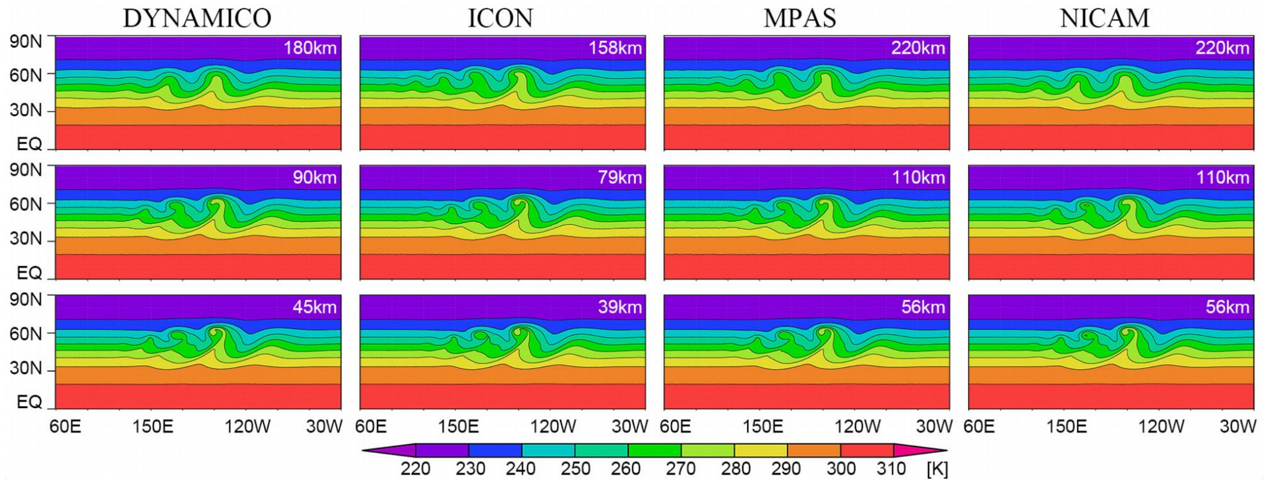
Selection of Conducted Experiments



ex1). Meteorological Aspect: Baroclinic Wave Test (i.e. mid-latitudinal low)

→ *All four models simulated reasonable wave structure!!*
In higher horizontal resolution, finer structures are simulated.

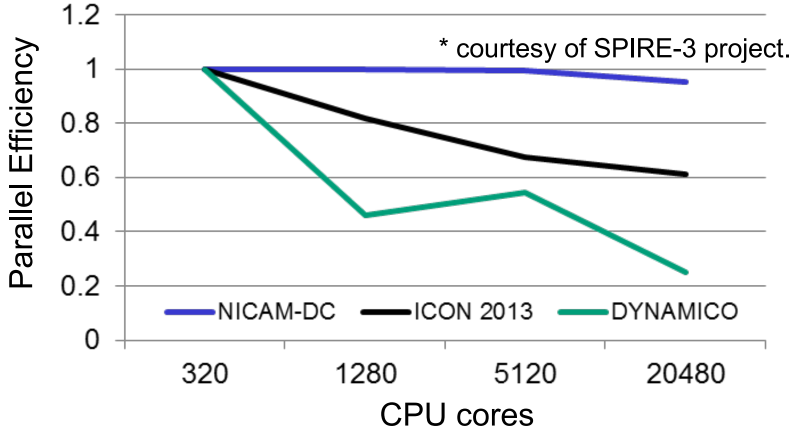
Temperature field around 1.5km height from sea surface after 9 days time-integration in models



ex2). Computational Performance: Weak Scaling on K computer

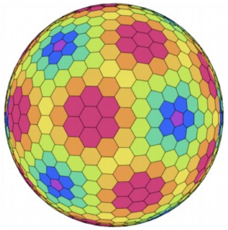
→ *NICAM was already tuned for K computer, and achieves weak-scaling up to 655,360 cores (not shown).*

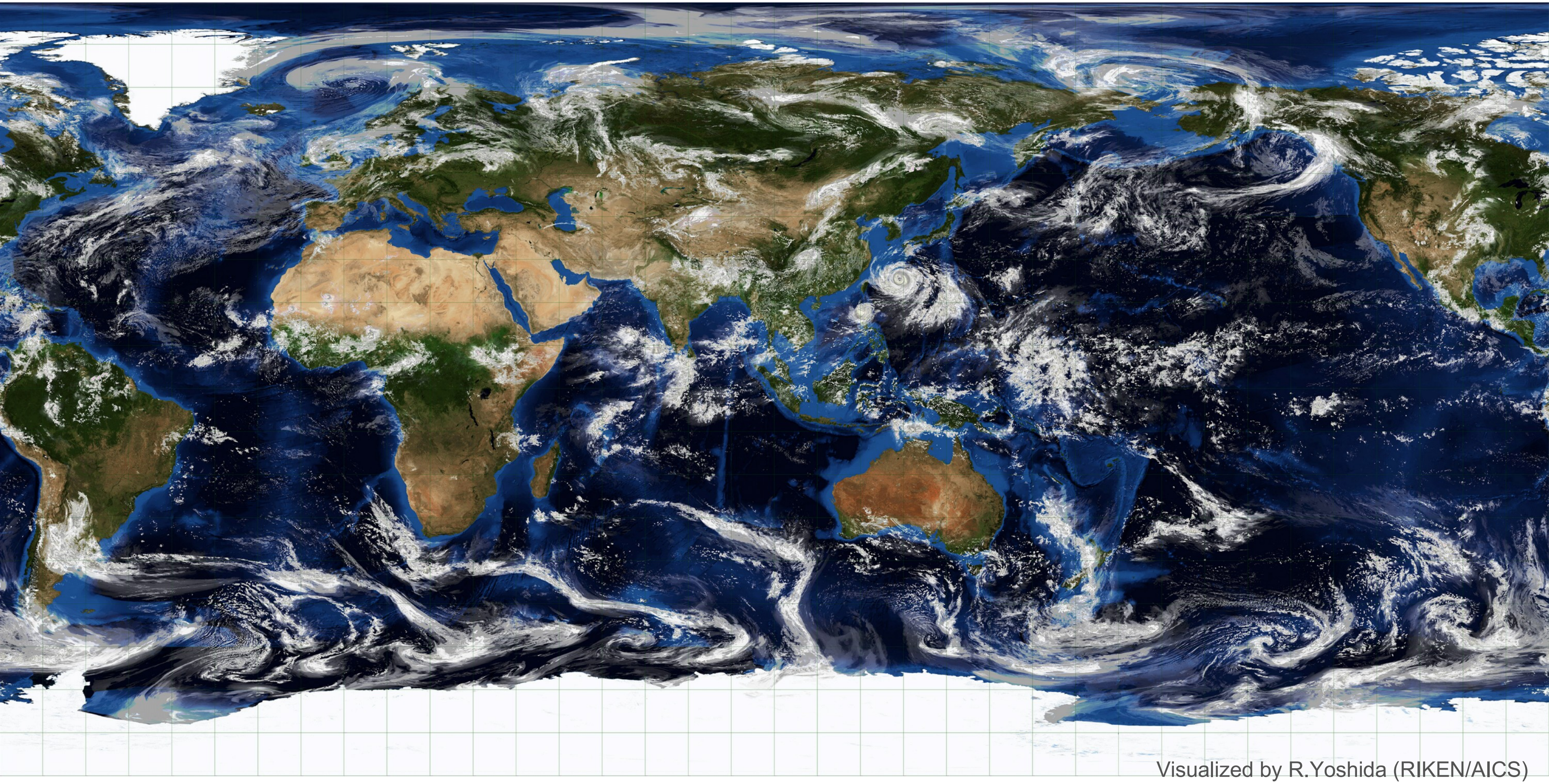
→ *AS-IS codes of ICON and DYNAMICO show reasonable weak-scaling with large number of cores.*



Towards Exascale Climate Simulations

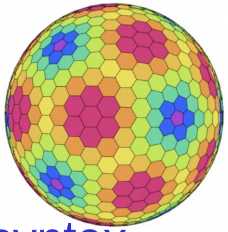
Global sub-km simulation by NICAM (Miyamoto et al., 2013 GRL)
20480 PEs(163840cores) of the K computer with 0.2 PFLOPS





Real simulation: 25 – 26, Aug, 2012
dx=870m, 97layers, dt=2sec
870 EFLOP for 24hour simulation
8TB for restart file, total output was 160TB for 24hour simulation

WP2: Abstract Model Description Scheme



Memory abstraction: abstraction of arrays and loops via memory unaware syntax

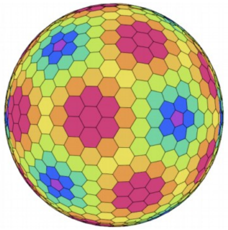
Goals:

- Express the model in a “natural way”
- Reduce CS info unrelated to the model
- Generate “architecture depended” memory access patterns
- Facilitate architecture specific optimizations

Approach:

- Extend Fortran to include subset notation
- Use Source-to-Soucre translation
  - Simplifies the gap problem between general languages and architectures. Allows the use of other architecture-unaware language approaches
  - <u>Bottleneck</u>: no mature Source-to-Source tools

WP2: Abstract Model Description Scheme



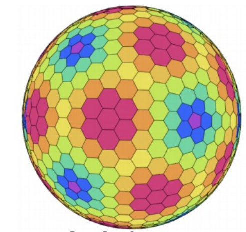
Example of subset usage:

```
Subset,  on_cells_3D :: all_cells
Element, on_cells_3D :: cell
Element, edges_of_cell_2D :: edge

! sum over a subset
for cell in all_cells do
    div_vec_c(cell) = sum[for edge in cell%edges] (vec_e(edge) *
         ptr_int%geofac_div(edge))
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
```

Status and Outlook



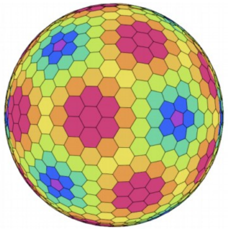
Status:

- Preliminary results for the ICON nh-dycore show up to 20% speedup on traditional architectures (pwr6, Westmere)
- Evaluate optimal memory layouts for simple operators on Accelerators (in progress)

Plans:

- Design NICAM, DYNAMICO dialects with the collaboration of the ICOMEX group (community feedback is important)
- Create 'lite' DSL that does not require sophisticated Source-to-Source tools. Less powerful but more likely to be implemented in the production codes in midterm
- Seek collaboration with DSL Source-to-Source tools initiatives. Not climate specific, but potentially can provide a very powerful framework.

WP4: Implicit solvers for massively parallel computing



Motivation

- Many operational models (such as the Met Office) use 3D implicit time integration schemes to achieve excellent stability, accuracy, and robustness
- 3D implicit schemes require the solution of an elliptic problem at each time step; feared to be expensive in communication
  - ➔ Originally ICOMEX models used *horizontally explicit vertically implicit* (HEVI) time integration schemes

Goal

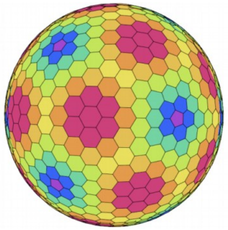
Demonstrate the feasibility of a 3D implicit scheme on massively parallel computers

➔ Make the first clean comparison of cost/accuracy between 3D implicit and HEVI in one model

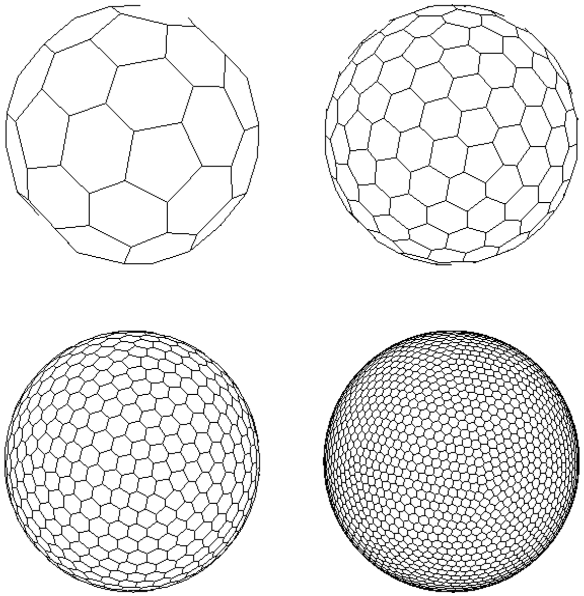
Approach

- We are implementing a Strang-Carryover scheme in MPAS (including Helmholtz solver)
  - Slow terms are treated with a RK3 step, similar to original scheme
  - Fast terms are treated with a trapezoidal step
  - Linear system for the unknowns gives a (elliptic) Helmholtz problem

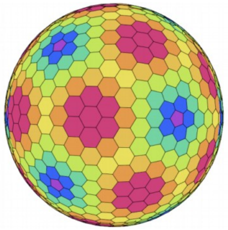
Multigrid Helmholtz solver



- For our specific problem, a bespoke solver (rather than a general purpose package) is advantageous; e.g. vertical line solve to handle vertical stiffness
- Advantage of multigrid
  - Only local communication at each iteration
  - Does not inhibit bit-reproducibility
- Helmholtz problem is well conditioned
  - Only a shallow grid hierarchy is needed (3-4 levels)
- Utilize experience
  - ENDGame and GungHo projects

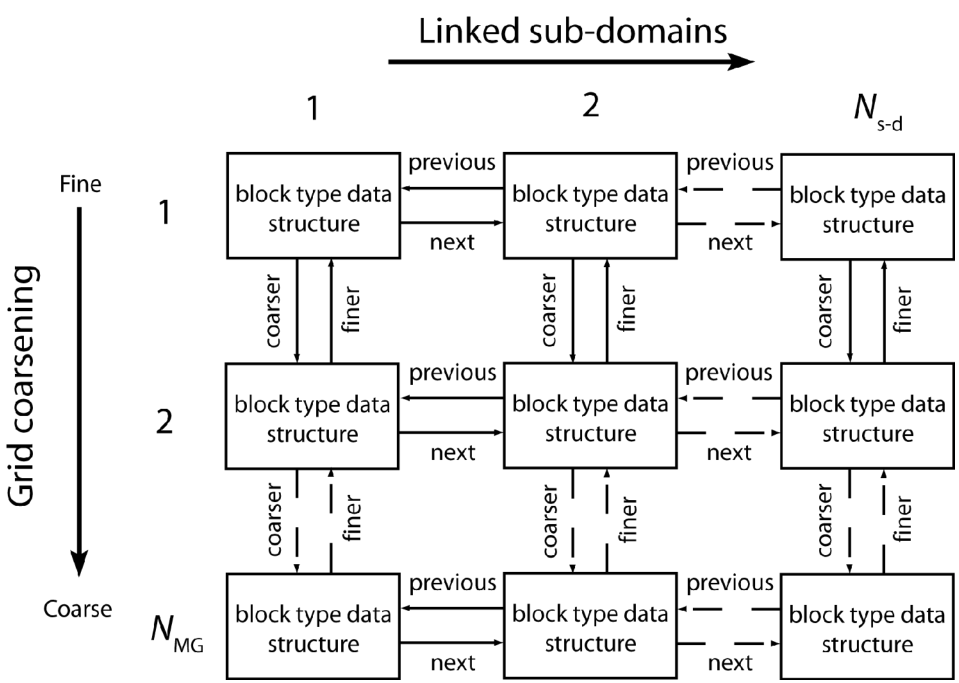


Implementation Details

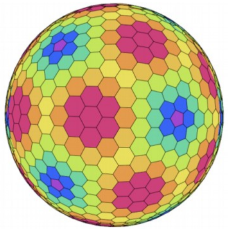


- Extend MPAS data structure to handle multi-resolution fields
- New subroutines to implement new time integration scheme, including multigrid solver. Analysis of data flow to determine halo exchanges
- Communication burden is estimated to be similar to the original HEVI scheme

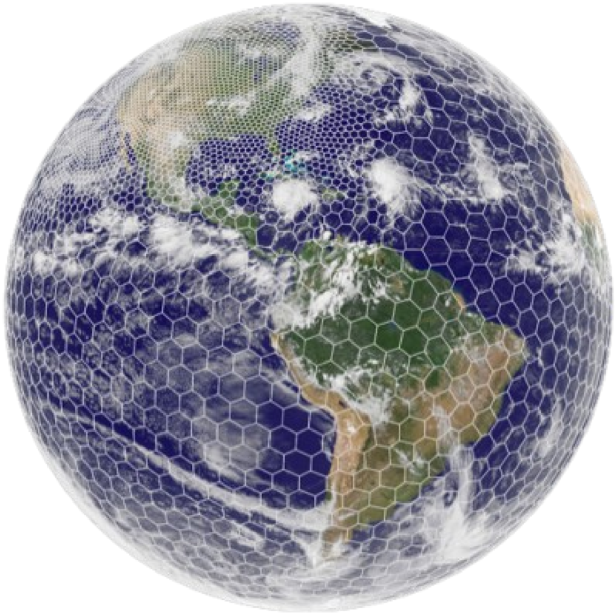
Hope to have a working version in the next few weeks, then compare (I) model results and (ii) parallel performance with the original HEVI scheme



Lessions Learned



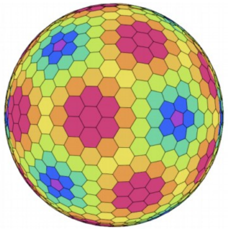
- Importance of working closely with main code owners/developers
  - International collaboration very important!
- Ideas and scientific understanding should be transferable to other models
  - Code is unlikely to be



- The multigrid data structure can have other applications:
  - Quickly output low-resolution output
  - Data assimilation
- We plan to apply the results of WP5 to use the multigrid solver on locally refined grids

(This WP will finish in Feb 2015)

WP5: Parallel Online Post-Processing



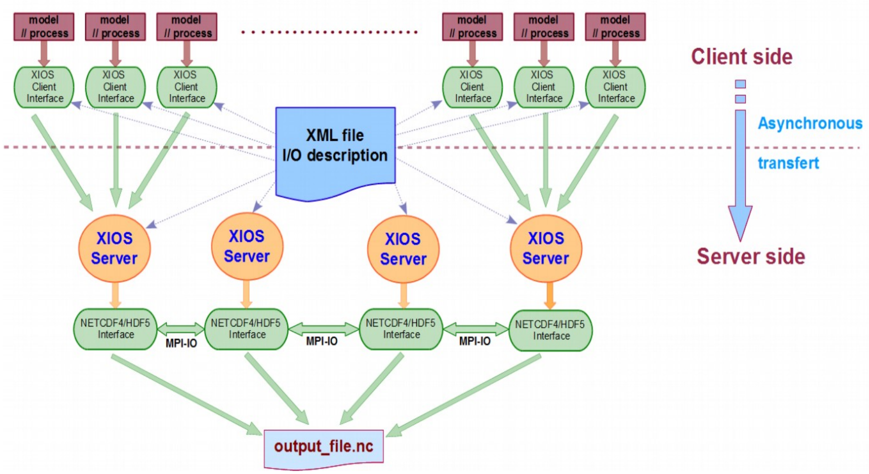
Motivation

- Business-as-usual not sustainable at extreme resolutions with current solutions (parallel asynchronous I/O – XIOS, XML I/O Server)

Goal: Address bottleneck caused by massive I/O

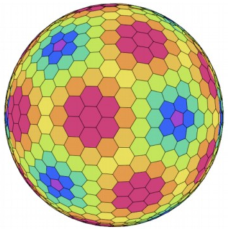
Approach

- Online post-processing to limit I/O demand
  - retaining scientifically important information
  - Local/temporal post-processing already provided by XIOS
- In ICOMEX: Extend XIOS with a critical feature
  - Remapping to non-native grids
  - Icosahedral grid => Coarser grid, lon-lat grid



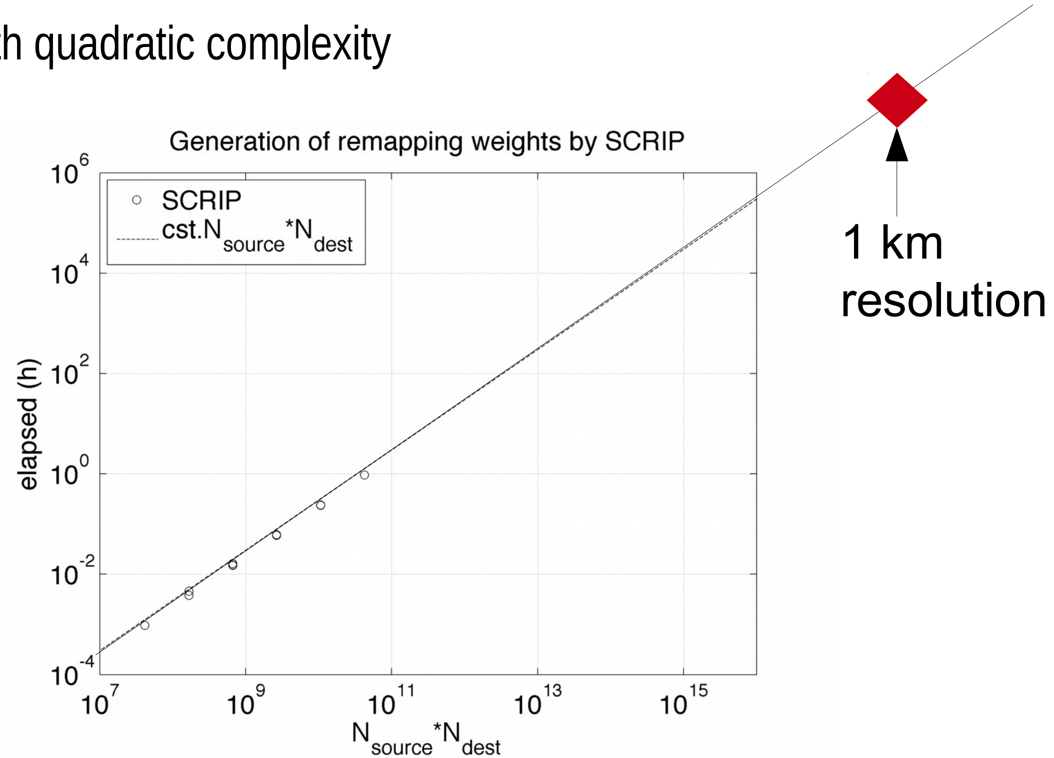
  - Features:
    - Flexible, accurate (second-order, conservative), linear complexity, scalable

WP5: Parallel Online Post-Processing



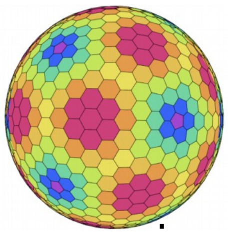
Existing solutions (SCRIP) not scalable

- brute-force algorithm with quadratic complexity



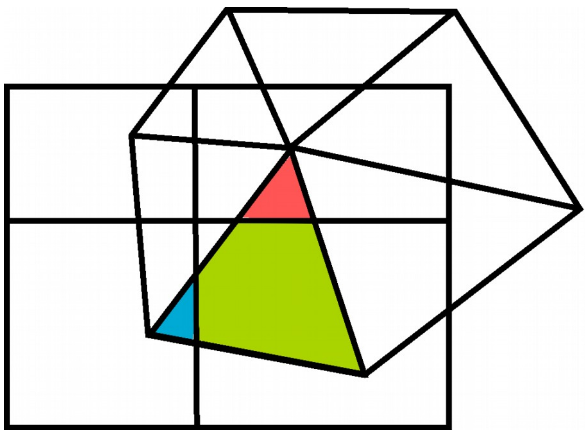
Time needed by SCRIP to generate remapping weights. 1km resolution corresponds to $5*10^8$ grid cells.

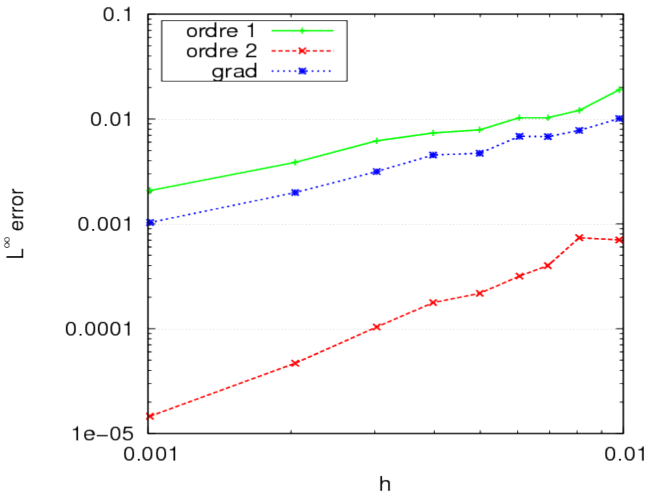
Achievements of the Remapping Scheme



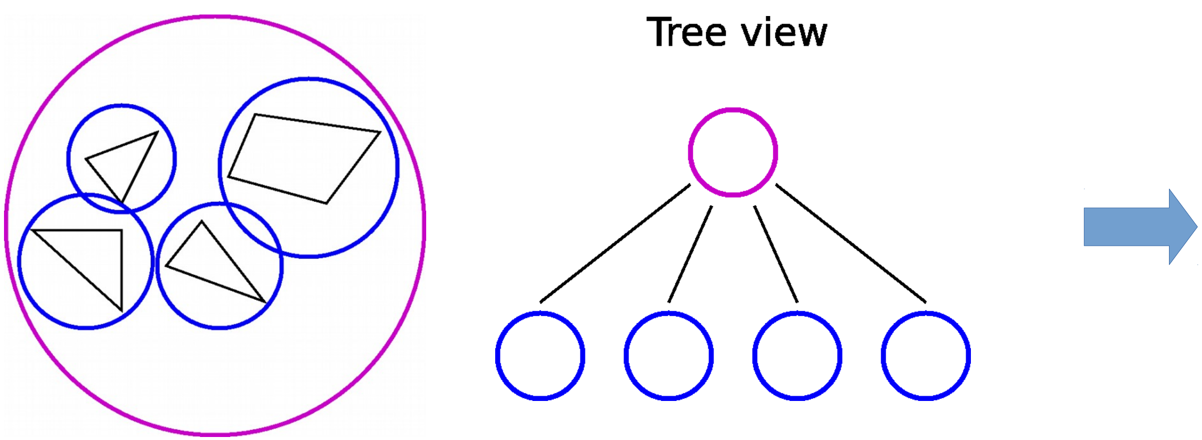
Flexible: Supermesh from polygonal and lat-lon meshes

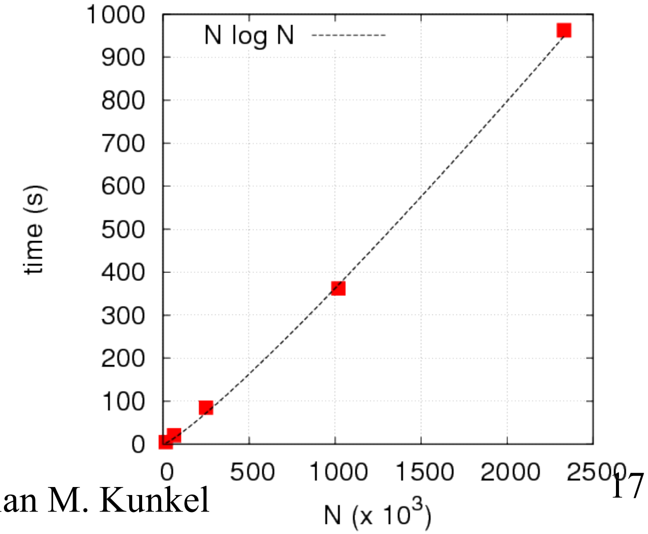
Accurate: Conservative 2nd order remapping



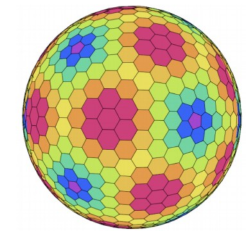


Efficient: Tree-based search for supermesh construction in O(N logN) time



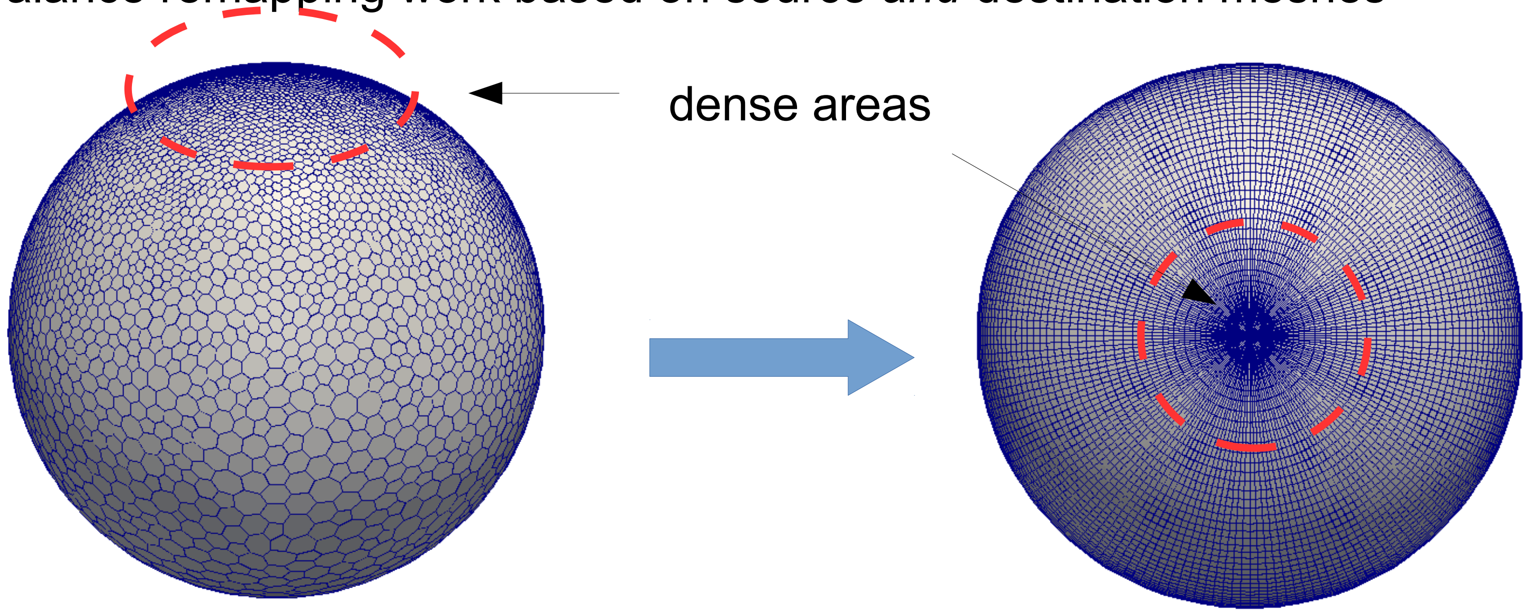


Achievements of the Remapping Scheme



Scalable parallel remapping

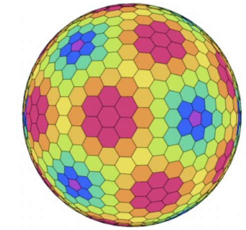
- Balance remapping work based on source *and* destination meshes



Work in progress

- Parallelize partitioning, tree-building and supermesh construction
  - to be achieved by Dec 2014
- XIOS now handles unstructured meshes
- Deliver remapping library(ies) for other ICOMEX models at project end

WP6: Parallel I/O



Goals: Analysis and optimization of parallel I/O

Approach

1. Analysis of ICON I/O (as archetype for other models)
2. Creation of an ICON similar benchmark
3. Evaluation of I/O performance on all involved layers
4. Localization of bottlenecks
5. Performance optimization

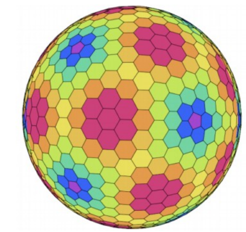
- Orthogonal effort: Storage format optimization

Performance loss due to suboptimal interactions between file systems and I/O layers

Modelling of Parallel I/O

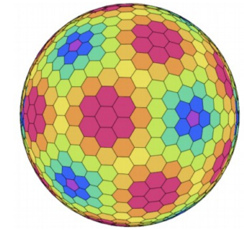
- Qualitative assessment of I/O architectures
  - Before any code is written!
- Conclusions
  - Asynchronous approaches are preferable
  - Independent parallel I/O is preferable
- Burst buffer concepts are essential
  - Currently implemented by domain scientists!

Achievement: Compression



- Pushed lossless limits with MAFISC preconditioner
  - Roughly 10% better compression ratio than best other algorithm
  - Slower than other algorithms
- Economic evaluation, example DKRZ tape archive:
  - **Good compression is more important than speed**
  - Fastest algorithm: 45068 €/a, best algorithm: 81494 €/a
  - **MAFISC has best economics: 91857 €/a**
- MAFISC is contributed to the community => available as an HDF5 plugin
- **Scientists need to consider lossy compression!**

http://wr.informatik.uni-hamburg.de/research/projects/icomex/mafisc



Optimization to NetCDF: NoCache Patch

- NetCDF shows bad performance with large records
    - Culprit: NetCDF cache + data initialization
- Prepared patch to deactivate NetCDF cache
    - **> 3x improvement on DKRZ supercomputer**
- Problem communicated with NetCDF community

http://wr.informatik.uni-hamburg.de/research/projects/icomex/cachelessnetcdf

Optimization to HDF5: Multifile Patch

- Observation
  - best performance with parallel writing of independent files
- Patch for HDF5 accesses multiple files transparently
  - Eliminates need for synchronization
  - Reconstruction of data upon read
  - **10x faster parallel writing (measured on one node)**
- Patch communicated with HDF5 community
- Conclusion: Scientists should not worry about data layout

http://wr.informatik.uni-hamburg.de/research/projects/icomex/multifile
hdf5

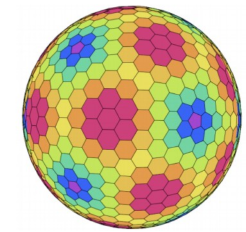
WP7: Vendor Communication

Goal: Communicate bottlenecks and requirements to vendors, best practices

Approach

- Invitations of vendors to meetings
- Bridged the gap to vendors/groups developing I/O middleware
    - IBM, HDF5, NetCDF
    - Satellite effort: integrated reqs. into Exascale10 initiative
- We developed a concept for better vendor communication
    - Classical bilateral approach is suboptimal
    - But: the implementation would be a project of its own

Lessions Learned During the Project



- International communication and coordination is important
  - Huge potential to share/re-use approaches and results
  - G8-initiative is perfectly suited to overcome organisational hurdles

- More interdisciplinary effort involving computer science is needed
  - e.g. co-design with storage system developers

- Challenges to overcome
  - Code portability
  - Performance portability
    - Inefficiencies in deployed software stack
  - Appropriate abstraction to formulate models

- Opportunities for international funding
  - Joint development of key components
  - Establishing of useful standards

Conclusions and Future Plans

G8 project boosted activities

- Comparing model (scientific) performance (CMIP5 in a limited scope)
- Running models on the K computer
- Exchanging best-practices well defined topics
- Some components have already been evaluated / adopted by other groups
- Involving vendors requires to create enough interest

Collaboration was a success BUT we'll need to strengthen it

- Increase communication
- Add further countries
- Involve data centers directly
    - They know their systems best, bridging the gap to vendors
    - Apply for computing time for the project
  - Strengthen collaboration with US institutions e.g. NCAR
- We'll try to build and exchange more components between the models
- Some claim: Funding was not enough to include the critical mass of people
  - => Go for larger projects