Porting and Optimizing the COSMOS coupled model on Power6

Luis Kornblueh
Max Planck Institute for Meteorology

November 5, 2008

Outline

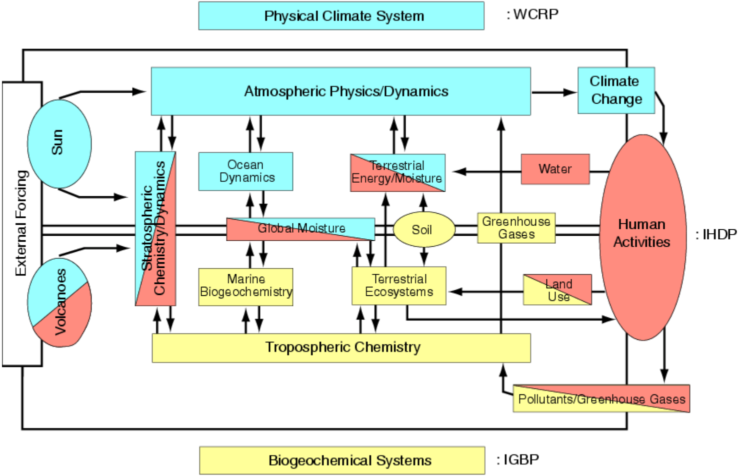
1. Introduction
2. ECHAM5
3. Performance
4. The COSMOS coupled model
5. Conclusion

Earth System Research - the general objective

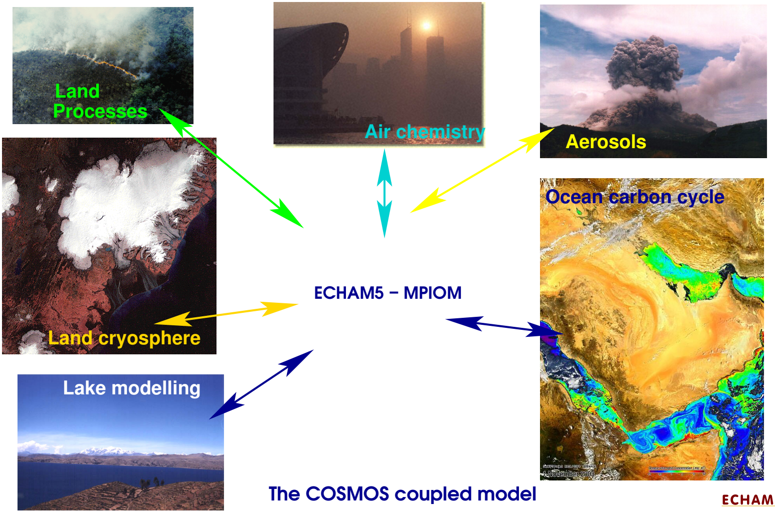
Why to do it ...

- To understand how physical, chemical, and biological processes, as well as human behavior contribute to the dynamics of the Earth system, and specifically how they relate to global and regional climate changes.
- To observe, monitor, analyze, understand, and predict in order to better manage the Earth system.

The system to look for



Standard model extensions ...



Land Processes

Air chemistry

Aerosols

Ocean carbon cycle

Land cryosphere

ECHAM5 – MPIOM

Lake modelling

The COSMOS coupled model

The climate model ECHAM5

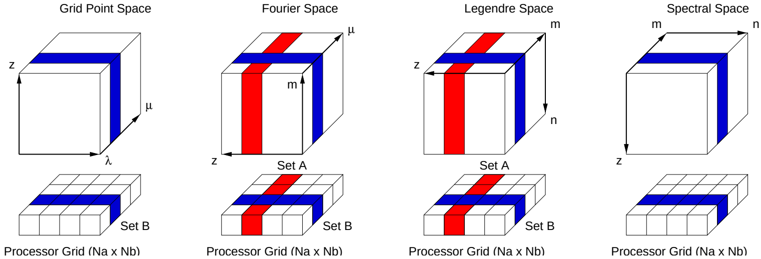
- Spectral dynamical core
- Semi-implicit leapfrog time differencing
- Flux-form semi-Lagrangian transport of passive tracers
- Shortwave and longwave radiation schemes
- Stratiform clouds based on micro-physical prognostic equations
- Convection solved by a mass-flux scheme (Tiedtke/Nordeng)
- Subgrid-scale induced gravity wave drag
- Vertical diffusion (subgrid-scale turbulence closure by TKE)
- *Solving the surface energy balance equation*
- *Mode based aerosol physics*
- *Extensive multi-tile land-surface model including dynamik vegetation*

Model states

Decomposition challenge

- spectral space
  - Legendre space
  - Fourier space
- grid point space
- transport flux-form space
- grid point space
  - Fourier space
  - Legendre space
- spectral space

Double transposition strategy



This is subject to extensive optimization work: transpose by ESSL, reworked communication for optimization with IBMs MPI

Redistribution for tracer transport

Scaling improved

- Resort in latitudinal direction (Na)
  - higher wind speed (transporting) in east-west direction, hardly predictable Courant number
  - low wind speed in north-south direction, Courant number < 1
- and vertical levels and tracers (Nb)

Transposition back into grid point space before calculation of vertical transport.

Communication improvements for high core counts (Thanks to ET, CRAY)

Loop-level parallelization

OpenMP usage

- Physics block - OpenMP orphaning
- Remaining model - OpenMP on loop level *done for all remaining model parts (Thanks to SB, NEC - now University of Cologne)*
- Performance problem on NEC: OS jitter - *no analysis on AIX yet*

Vectorisation/Cache blocking

Flexibility

- Physics block - high level strip mining (VL/average optimal cache blocking), *nproma SX: n × VL, nproma PWR6: 72 - fully L2 resident*
- Remaining model - blocking and *fast and dirty* code
- Performance problem 1: tracer transport as usual
- performance problem 2: *non-mathematical* formulated part *made some first try and applied for a project to tackle this problem together with FHG SCAI*
- However reaches 2 GFlops on SX-6 on a small resolution (T63L31) run on a single CPU *and without any changes 1 GFlops on PWR6*
- *Points out that a lot of special vector code is very good for the in-order architecture of the Power6*

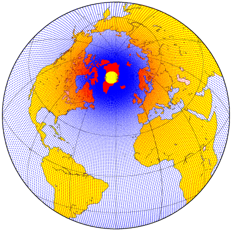
Intranode performance

core binding - a lot of options

- *not analyzed in detail yet, but big improvements in scaling due to SMT usage suggests that the single CPU performance can be improved substantially. There seems to be less OS jitter.*
- *OpenMP is not explored by now but is coming soon (has already been used on SX)*

MPIOM

- Primitive Equations
- C-Grid, z-level, partial cells
- mixed upwind/centered difference advection scheme (other schemes possible)
- Hibler sea ice model incl. snow and fractional ice cover *replacement in testing*
- Sub-gridscale parameterisations
- Isopycnal diffusion
- Edddy induced tracer transports
- Slope convection
- Conformal mapping

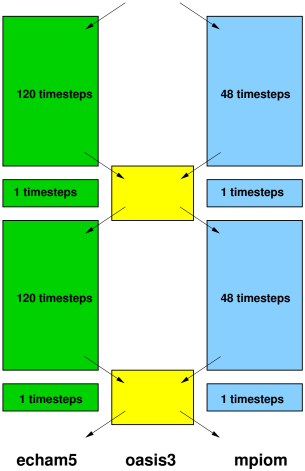


The ocean model development level

hybrid parallelization

- hybrid MPI - OpenMP
- very good vectorization level *reordering of loops and properly selected decomposition makes it L2 resident - however bandwidth requirement are very high and current performance reflects stream benchmark)*
- very poor global solver (2d linear equation system, SOR or direct LU) *improvments have been done*
- 3.75 GFlops on a singel SX-6 CPU, *but 700 MFlops on PWR6 at the beginning, improvements under way*

The coupling itself



Inherent load inbalance causes half of a total of 30 % loss in performance by coupling.

Improved performance for coupling

SMT based

Due to the fact that the requirements on bandwidth are so different a SMT based seems to be a promising way to go

I/O improvements

Latest experiments

- climate simulations of several 1000 years, in total 30000 years!
- data handling very complicated and workflows are very unstable

I/O improvements continued

Improvement by additional packing of the BDS data section in GRIB and using zlib on netCDF reals (dynamic range problem) delivers a packing factor of 2 (netCDF4)!

Properties

Resolution	grib-szip	gzip (external)
Source	MPI-M	GNU
E5 T42 L19	2.13	2.06
E5 T63 L31	1.78	1.35
E5 T106 L60	4.75	3.81
E5 T213 L31	3.06	2.41
mean	2.93	2.15

I/O improvements continued

Problems

- NASA and others szip solution patented - no cost for non-commercial organisations but would be great to get this into WMO standard for grib2! What to do? Getting the NASA patents licensed without fee for WMO and implement it indenpendently from szip?
- grib-szip: 84 MB/s throughput on a standard PC (maximum possible disk bandwidth)

Conclusion

Work to do

- We have to work more on achieving high performance on the component models and the coupled model . . . - machine will be installed in Januray 2009.
- optimize and parallelize post-processing tools
- Work on stable workflow handling
- Improve model infrastructures by using Fortran 2003 features.
- Explore CAF as soon its available

Vote for . . .

. . . using advanced Fortran

- 2003 strong need to support the development of high level model infrastucture - this requires the OO features provided by the actual standard!
- 2008 ask your compiler vendor to support CAF at least to provide an intra-node implementation