Increasing parallelism in climate models via additional component concurrency

Jörg Behrens, Hendryk Bockelmann
Deutsches Klimarechenzentrum (DKRZ)

Joachim Biercamp, Philipp Neumann, Reza Heidari (DKRZ),
Karl-Hermann Wieners, Leonidas Linardakis (MPI-M)

PAL MOD GERMAN CLIMATE MODELING INITIATIVE

CHALLENGES

- identify and quantify fundamental processes of Earth's climate trajectory and variability during the last glacial cycle
- simulate with comprehensive Earth System Models (ESMs) from the peak of the last interglacial up to the present – 130k years
- assess possible future climate trajectories beyond this century

Physical System

Biogeochemistry

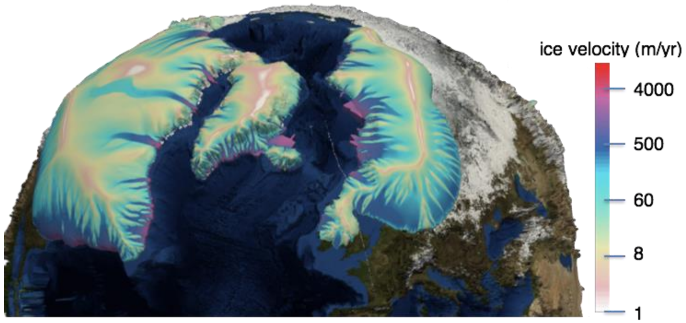
Synthesis and Analysis of Proxy Data

Optimization of Quality and Performance

Optimization of Quality and Performance

Additional workload resulting from improved physical & biogeochemical processes like

- Feedbacks between continental ice sheets, sea level & large scale ocean circulation
- Dust sources, transport and deposition
- Variable land sea mask



F. Ziemen, N. Röber

Requirements (atmospheric component ECHAM only)

- LR (T63L47, 1.9°, 147km at 45°) desired
- CR (T31L47, 3.8°, 295km at 45°) tolerable for higher throughput
- 500-1000 SYPD needed to simulate 130k years in a reasonable amount of time

Approaches

- Novel numerical concepts (e.g. parallelization in time)
- Improved technical concepts (e.g. component concurrency)

ESiWACE: Centre of Excellence in Simulation of Weather and Climate in Europe

WP1 Governance & engagement

WP2 Scalability

Global high resolution model demonstrators

→ **ICON**, IFS, EC-Earth, NEMO

→ **DYAMOND initiative**

WP3 Usability

WP4 Exploitability

WP5 Management & dissemination

Meet us!

- ICT, 4-6 Dec 2018, Vienna
- EGU, 7-12 Apr 2019, Vienna
- PASC, 12-14 June 2019, Zurich
- ISC HPC, 16-20 June 2019, Frankfurt

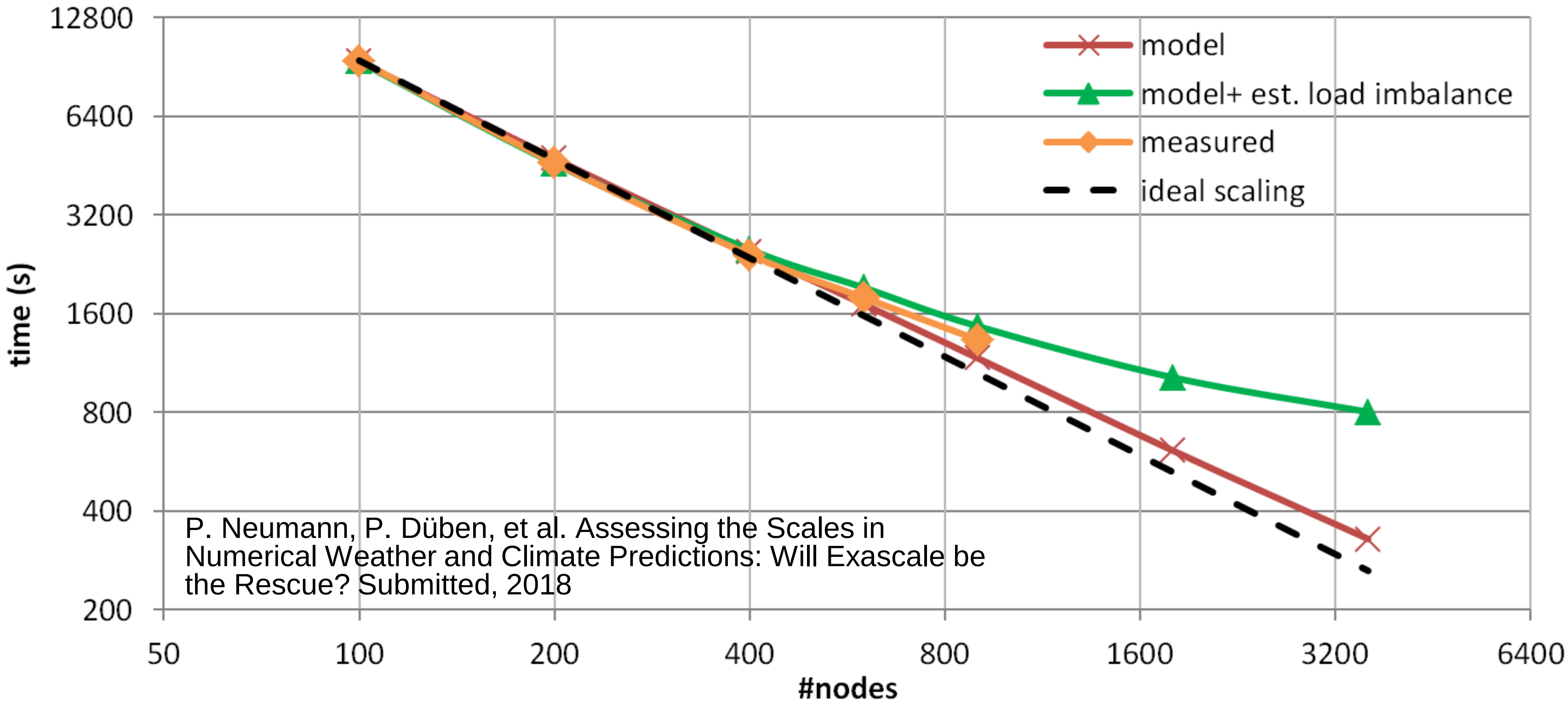
Coordinator: DKRZ

WEATHER CLIMATE HPC

DYnamics of the Atmospheric general circulation Modeled On Non-hydrostatic Domains (DYAMOND)

- Goal: Intercomparison of global high-resolution models
- Participation list: ICON, NICAM, MPAS, FV3, SAM,
  NASA GEOS5, UM, ARPEGE-NH, IFS-H
- Data management and support through DKRZ/ESiWACE
- More information: www.esiwace.eu/services/dyamond

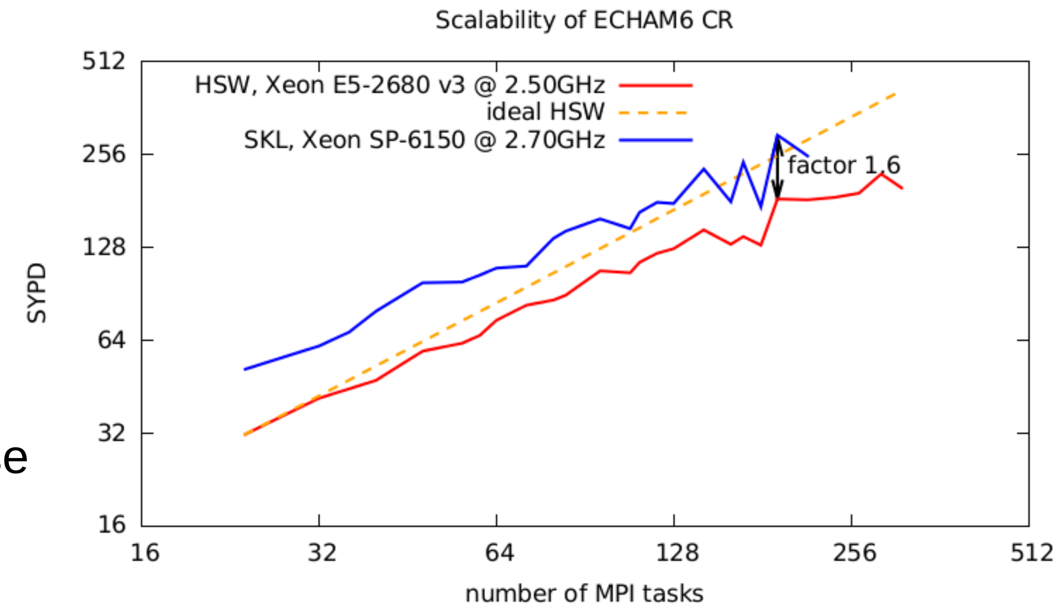
Scalability limit of ICON at high resolution

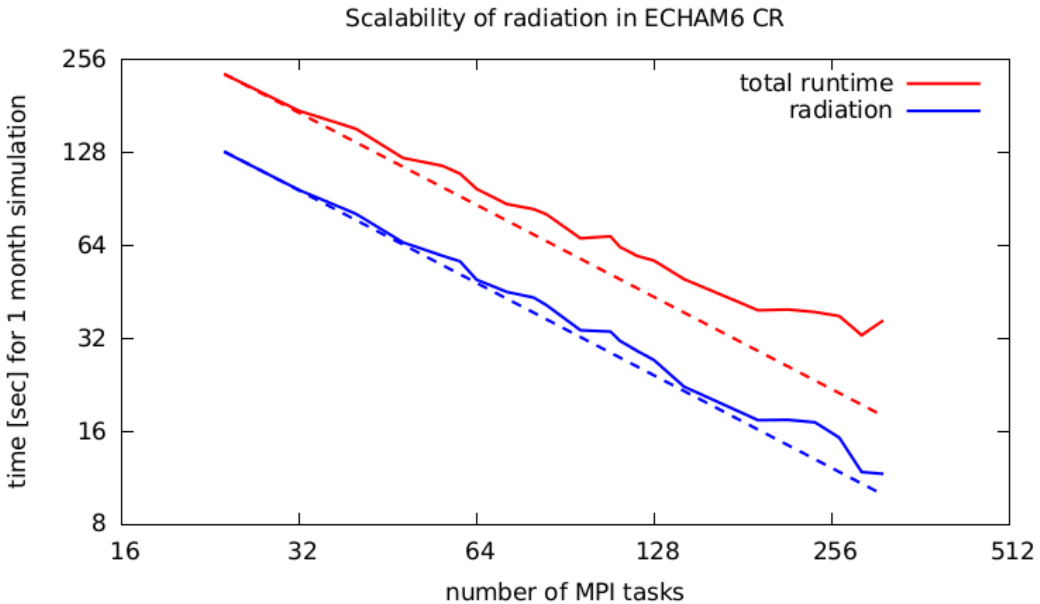


Goal: 1 SYPD throughput
Extrapolation of ICON R2B9 DYAMOND to 1km:
17x too slow, assuming infinite number of (Broadwell) nodes
→ need for radical performance improvement at all levels

Issues at coarse resolution

- Scaling via domain decomposition reaches its limit
- New CPU based hardware will no longer give jump in performance
- Switching to GPU based systems require too much effort for legacy codes
- GPUs do not perform well on coarse grids





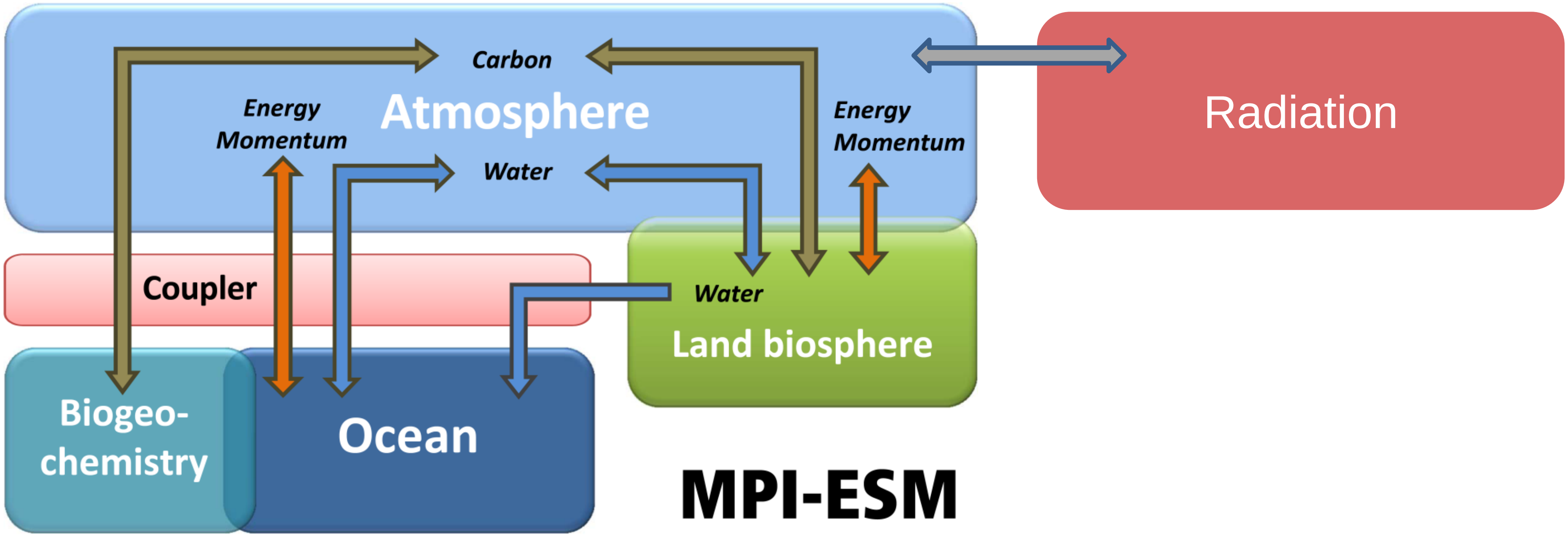
But still components exists that do scale !

Approach to Performance Improvement

- ESiWACE:
  - Single precision
  - OpenMP-based concurrency of radiation and wave model in IFS
  - DSL for performance portability (including GPUs)
  - HPC services to support wider community at performance tuning
  - evaluate concurrency on homogeneous & hybrid architectures (CPU,GPU), ICON:radiation as prototype, evaluate generalization [MPIM,DKRZ,MSWISS]

- PalMod:
  - single precision
  - flexible concurrent radiation using YAXT
  - Novel numerical methods

Component Concurrency



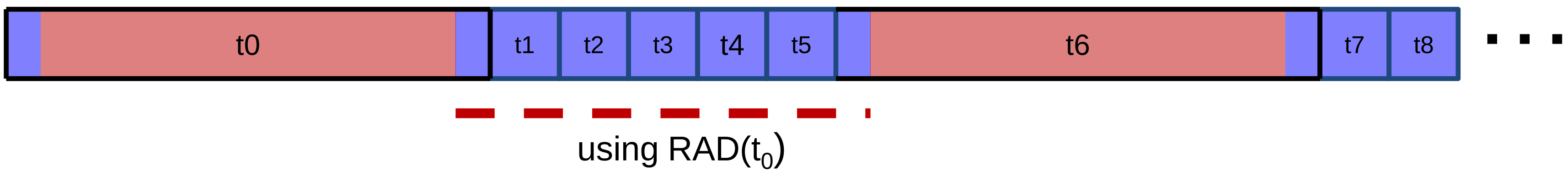
Michael Böttinger, DKRZ

based on:

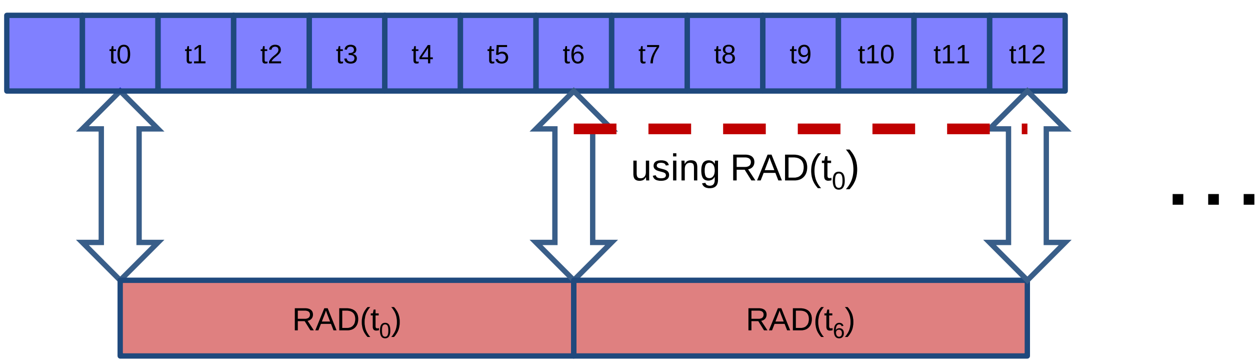
- IFS: ECMWF investigated MPI based concurrent radiation (Mozdzynski, Morcrette)
- Coarse-grained component concurrency in ESM (Balaji at al)

Concurrent radiation: time delay

sequential: ATM(t_0) → RAD → ATM(t_1) … ATM(t_{NRAD})

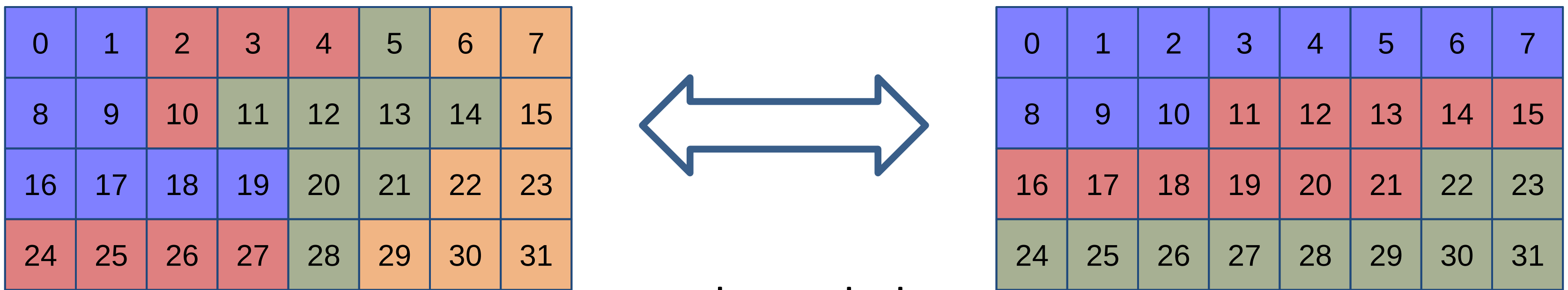


asynchronous: ATM(t_0) → ATM(t_1) … → / → RAD → ATM(t_{NRAD}) … ATM($t_{2 \times NRAD-1}$)



YAXT communication library: overview

YAXT redistributes data between decompositions



Decomposition A

color coded MPI task

Decomposition B

Usability:

- No explicit message passing required
- User only supplies decompositions + data layout

Performance:

- Exploits MPI performance potential
- Applies collective communication optimization

YAXT: general aspects

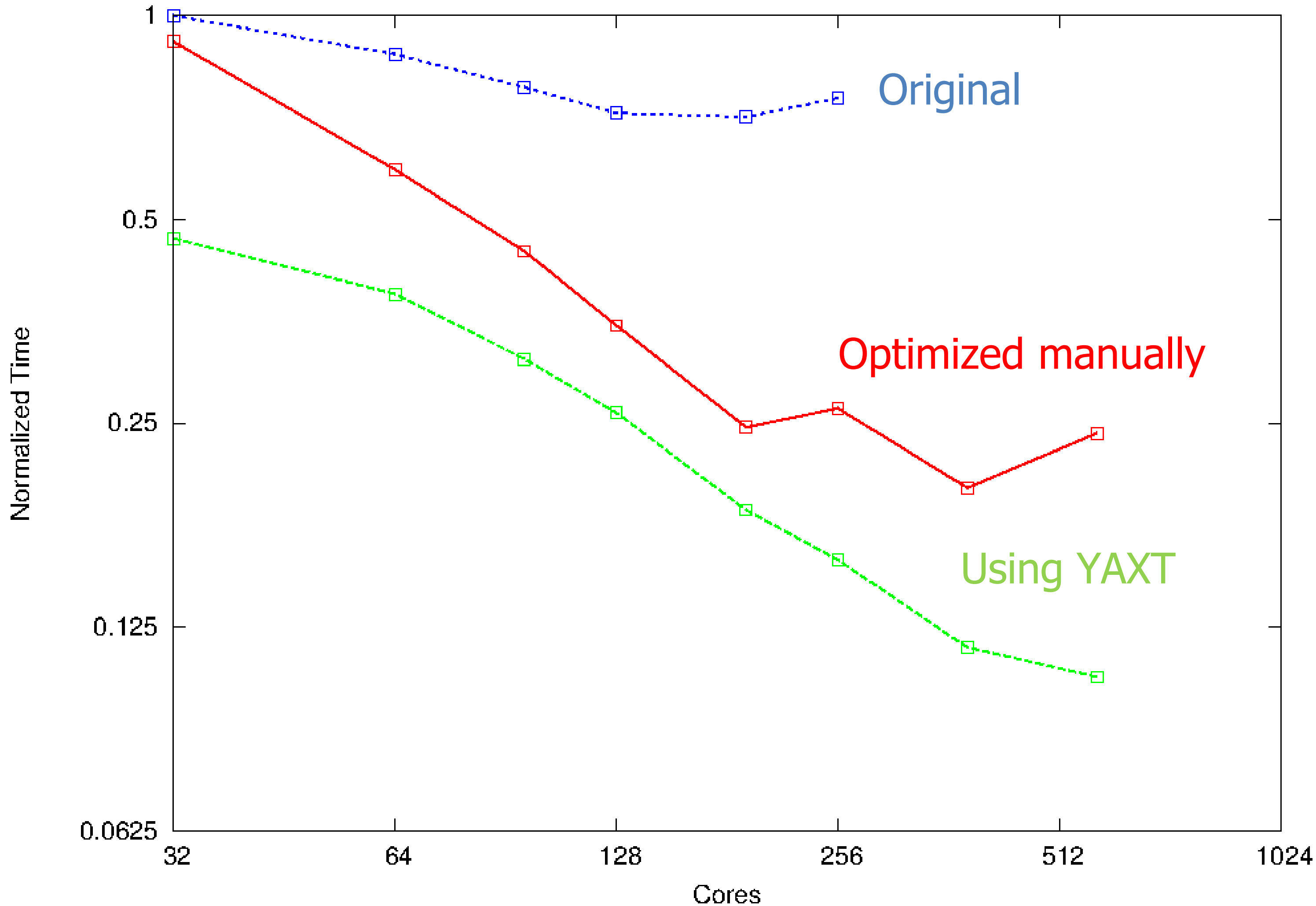
- Purpose:
  - Reduce complexity of writing MPI applications
  - Exploit difficult to use performance potential of MPI:
    - Data layout description using MPI Derived Data Types (DDT)
    - Supports aggregation of communication

- Concept:
  - Data abstraction: global index definition
    - Decomposition = distribution of indices
  - Separation between decomposition and data layout
  - Each process only requires local knowledge
  - YAXT provides communication objects to change decompositions

- Performance:
  - Library on top of MPI, performance depends on quality of MPI [DDT] implementation
  - Cooperation with BULL/ATOS to improve derived datatypes in OpenMPI

Performance example: ECHAM Transposition gp->ffsl

T63L47 (synchronized measurement on prev. Pwr6 system)



YAXT: general aspects (cont.)

- Related tools (all in Fortran):
  - Unitrans (ScalES project), MCT, PILGRIM
- YAXT is maintained by DKRZ
  - Dev. Team: Thomas Jahns, Moritz Hanke, Jörg Behrens
- Access:
  - Documentation: https://doc.redmine.dkrz.de/yaxt/html/
  - Download:
    https://www.dkrz.de/redmine/projects/yaxt/wiki/Downloads

Concurrent Radiation: communication aspects

single-phase communication:

- ATM tasks talk directly to RAD tasks
- ➢ Communication costs at ATM depends on decompositions at both ends
- ➢ Average communication costs for RAD and ATM
- Current test implementation:
  - Identical decompositions at ATM and RAD
  - Only single task to single task communication

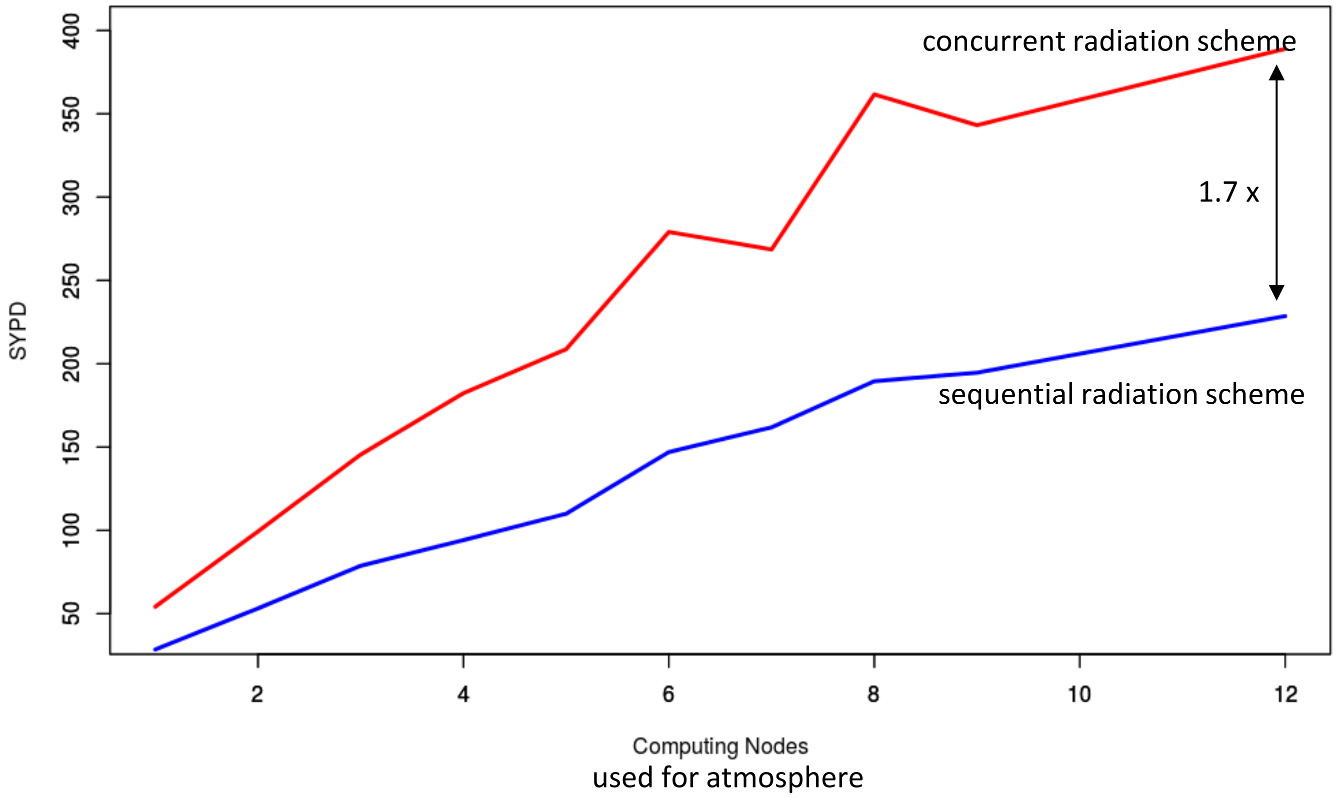
two-phase communication:

1. ATM tasks talk to a similar intermediate decomposition at RAD
2. RAD performs an internal transposition to reach final decomposition

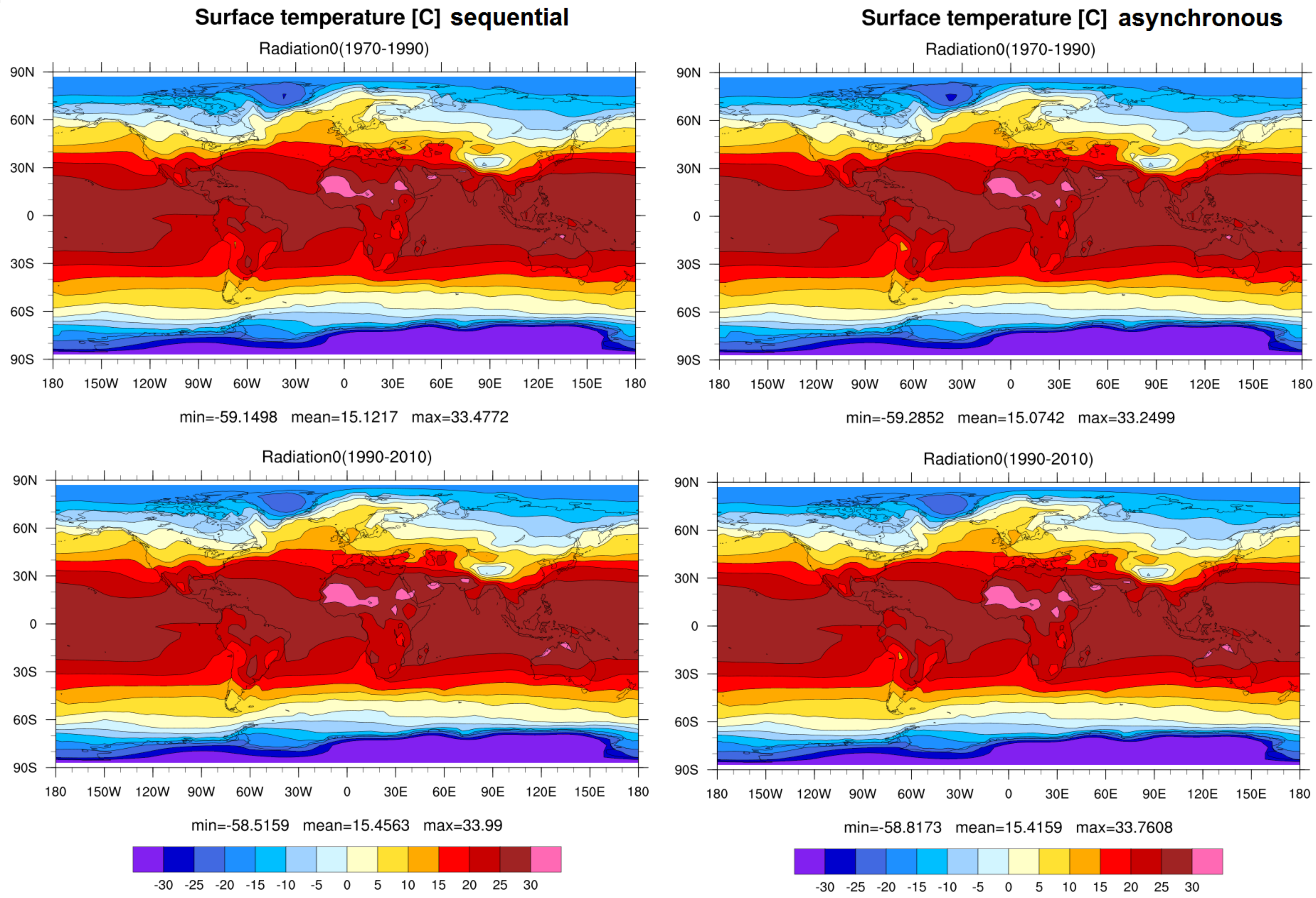
- ➢ Minimal communication costs for ATM
- ➢ Increased overhead for RAD

First Performance Results

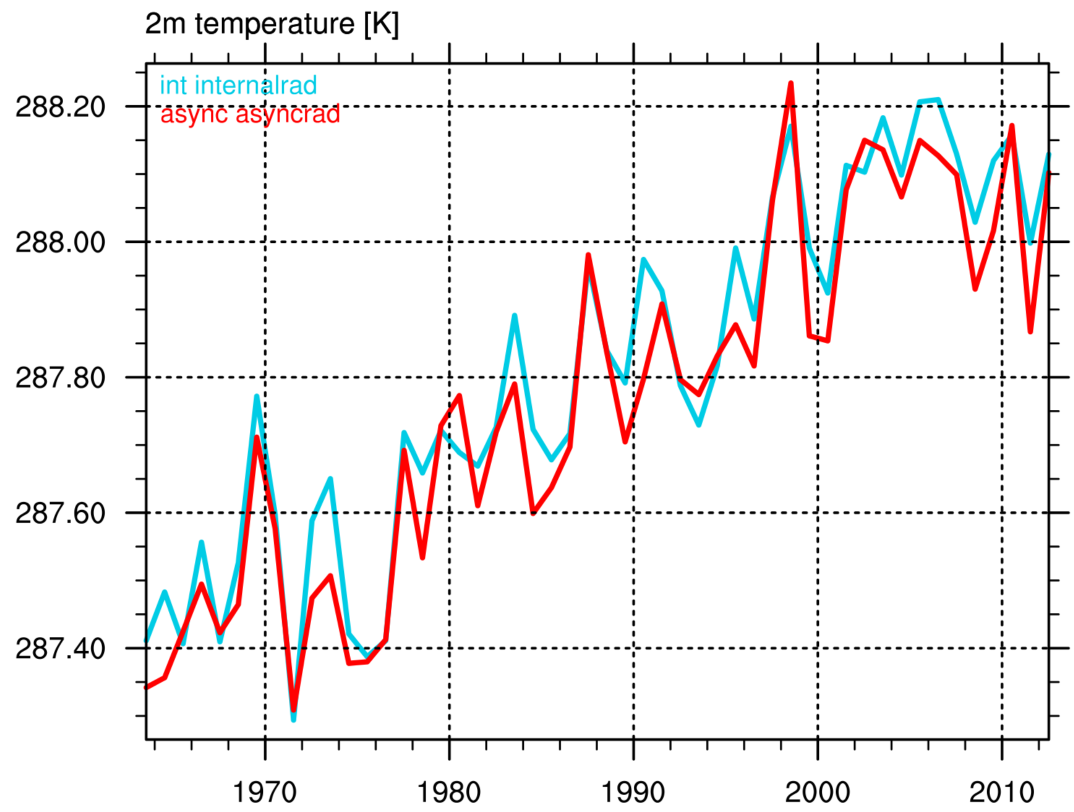
Comparison of sequential and concurrent radiation scheme in ECHAM6 at coarse resolution (T31L47)

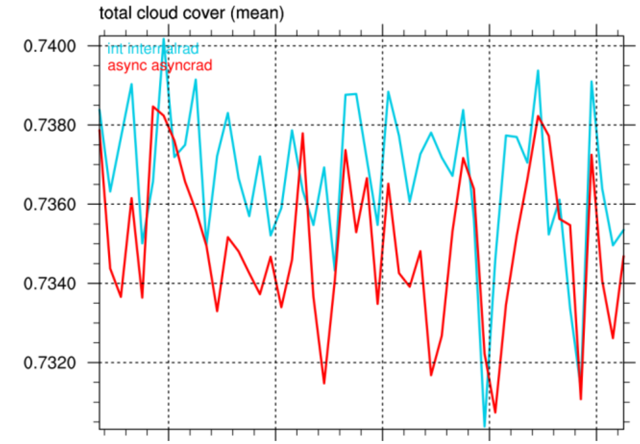


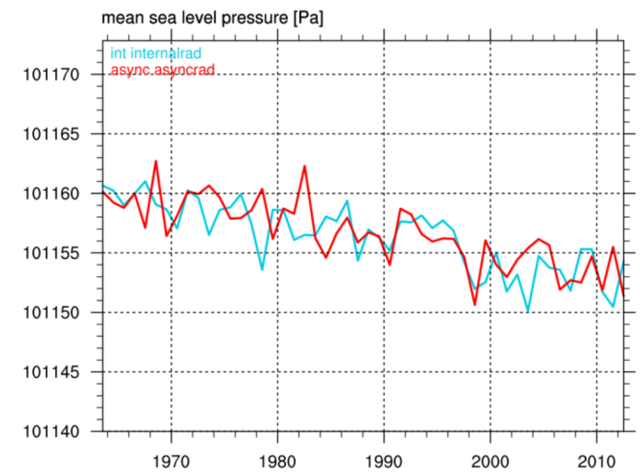
First Comparison of Simulation Results



First comparison of simulation results







Outlook

- Review and scientifically verify tolerable lag between ATM and RAD
- Further improve asynchronous scheme
  - Evaluate (dynamic) load balancing for radiation tasks
  - Align compute load in ATM and RAD to reduce waiting phases
- Technical optimization
  - Communication aggregations
- Extent component concurrency to other processes, e.g. passive tracer

Acknowledgement

PALMOD | GERMAN CLIMATE MODELING INITIATIVE

funded by

Federal Ministry of Education and Research

More information about PalMod*:*

www.palmod.de

esiwace

CENTRE OF EXCELLENCE IN SIMULATION OF WEATHER AND CLIMATE IN EUROPE

The ESiWACE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 675191

This material reflects only the author's view and the Commission is not responsible for any use that may be made of the information it contains.

References

Balaji at al [2016]

Balaji, V., Benson, R., Wyman, B., and Held, I.: Coarse-grained component concurrency in Earth system modeling: parallelizing atmospheric radiative transfer in the GFDL AM3 model using Flexible Modeling System coupling framework, Geosci. Model Dev., 9, 3605-3616, doi:10.5194/gmd-9-3605-2016, 2016

Mozdzynski, Morcrette [2014]

Mozdzynski and Morcrette, Reorganization of the Radiation Transfer Calculations in the ECMWF IFS, Technical Memorandum, ECMWF 2014