Karlsruhe Institute of Technology
KIT-Campus Alpin
IMK-IFU: Atmospheric Environmental Research
NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH
Universität Augsburg University

Contact
Dom Heinzeller
KIT, IMK-IFU
heinzeller@kit.edu

Aerosol-aware, convection-resolving climate modelling

Regional and global modelling approaches with WRF and MPAS

Preparing for Exascale: Convection-permitting, global atmospheric simulations

D. Heinzeller, M.G. Duda, H. Kunstmann, 2016: Geosci. Model Dev., 9, 77-110, http://www.geosci-model-dev.net/9/77/2016

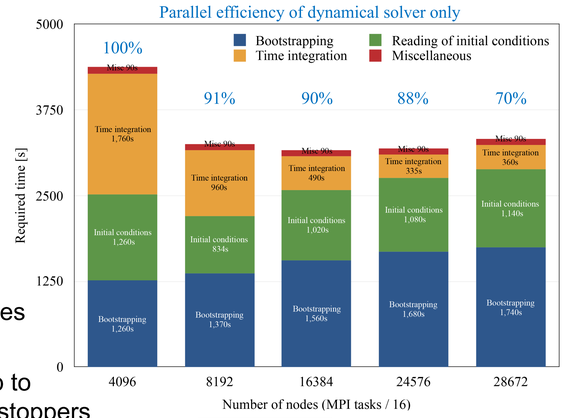
MPAS Model for Prediction Across Scales

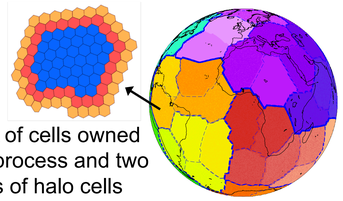
Convection-permitting global model applications are the next grand challenge in NWP and on the horizon of next-generation, massively parallel HPC systems.

Extreme scaling experiment with MPAS on FZJ JUQUEEN (IBM Bluegene /Q) in 2015:

- Uniform 3km mesh, 65,536,002 cells
- 41 vertical levels, double precision
- 1hr model integration, no file output
- Initial conditions: 1.1TB pnetCDF CDF5
- Min. 4096 nodes, 65TB memory; max: 28,672 nodes
- Fastest run: 6.3 x real-time, 1.6 Mio CPUh/24h

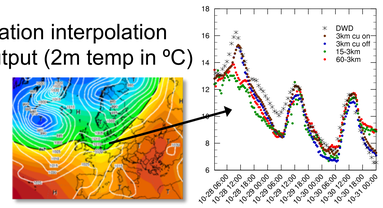
The dynamical solver of MPAS-Atmosphere scales up to 400,000 MPI tasks (160 owned cells per task). Show-stoppers at extreme scale are file I/O and model setup (bootstrapping).





Block of cells owned by a process and two layers of halo cells

Station interpolation output (2m temp in ºC)



Step 1. Addressing the file I/O performance

- SIONlib I/O layer (http://www.fz-juelich.de/jsc/sionlib) for massively parallel I/O in addition to existing I/O formats
- Post-processor core for converting to netCDF, regridding to lat-lon grids and interpolation to station locations
- Reading/writing in SIONlib format requires to use the same number of MPI tasks and the same graph partition
- Information encoded in SIONlib data can be used to skip parts of the bootstrapping at model startup

The SIONlib I/O layer addresses file I/O and model setup costs.

	Timer name	pnetCDF, CDF5	SIONlib
1	total time	3585	2117
2	initialise	1176	244
3	bootstrapping	540	168
3	stream input	612	52
2	time integration	1580	1658
2	stream output	818	204

Timing results for a uniform 2km mesh with 147,456,002 cells on LRZ SuperMUC on 2048 nodes x 16 MPI tasks x 1 OpenMP task (131TB memory). Integration for 10min model time using a conservative 5s time step.

```
120G Apr 10 11:28 diag.2013-10-27_12.00.00.nc
120G Apr 10 11:28 diag.2013-10-27_12.05.00.nc
120G Apr 10 11:28 diag.2013-10-27_12.10.00.nc
3.0T Apr 10 11:28 history.2013-10-27_12.00.00.nc
3.0T Apr 10 11:28 history.2013-10-27_12.10.00.nc
```

Step 2. Reducing MPI communication overhead

- Hybrid MPI+OpenMP parallelisation to speed up bootstrapping and decrease halo exchange time in dynamical solver of MPAS-A
- Optimisation for latest many-core architectures
- Combine with SIONlib I/O layer improvements for maximum performance at extreme scale
- Threading of one additional routine in solver
- Avoid repeated creation and destruction of threads

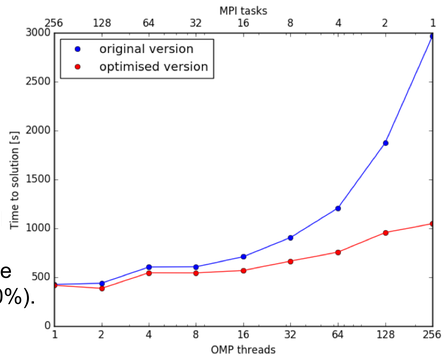
Nodes x MPI x OMP	Solver/h [s]	Init [s]	I/O [s]
4096 x 16 x 1	350	830	33
4096 x 8 x 2	323	258	39
4096 x 2 x 8	407	222	21
4096 x 1 x 16	689	162	5

1hr model integration of 3km mesh on LRZ SuperMUC

Original code	Optimised code
`! call non-threaded function`	`!$OMP parallel`
`...`	`...`
`!$OMP parallel do`	`!$OMP do schedule(static,1)`
`do thread=1,nThreads`	`do thread=1,nThreads`
`  ! call threaded function`	`  ! call threaded function`
`end do`	`end do`
`!$OMP end parallel do`	`!$OMP end do`
`...`	`...`
`!$OMP parallel do`	`!$OMP master`
`do thread=1,nThreads`	`! call non-threaded function`
`  ! call threaded function`	`!$OMP end master`
`end do`	`...`
`!$OMP end parallel do`	`!$OMP end parallel`

Original vs. optimised hybrid code on Intel Xeon Phi KNL for a uniform 240km mesh (10242 cells).

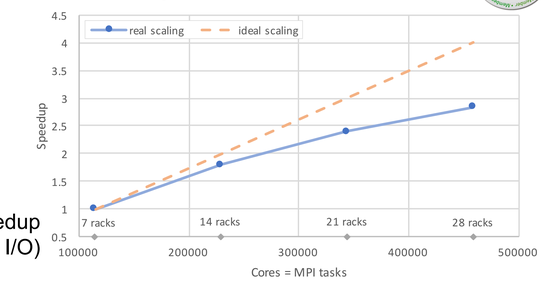
Note: Using optimised compiler flags for the Intel KNL, performance gains are smaller (≈20%).



Putting it to the test: extreme scaling experiment on FZJ JUQUEEN in 2017

- Jablonowski & Williamson (2006) baroclinic wave
- 2km mesh, 147,456,002 cells, single precision
- 26 vertical levels, 10min model integration, Δt=5s
- File input 1.8TB, output 4TB (SIONlib)
- Dynamics and file I/O only, no physics: more stringent test of dynamical solver
- MPI only: threading model not supported on JUQUEEN

Time integration speedup (dynamics and file I/O)



Conclusions

- Convection-permitting global simulations are within reach of current and next-generation HPC systems
- Efficient parallel I/O, code preparation for novel many-core architectures and HPC-specific adaptation are key to success

KONWIHR JÜLICH FORSCHUNGSZENTRUM SIONlib