Contact Dom Heinzeller, *now with:*
University of Colorado/CIRES at
NOAA/ESRL/GSD, Boulder, CO, USA
dom.heinzeller@noaa.gov

Exascale computing in numerical weather prediction

Massively parallel I/O in atmospheric models on conformal meshes

MPAS Model for Prediction Across Scales

* The National Center for Atmospheric Research is sponsored by the National Science Foundation.

Preparing for Exascale: Convection-permitting, global atmospheric simulations

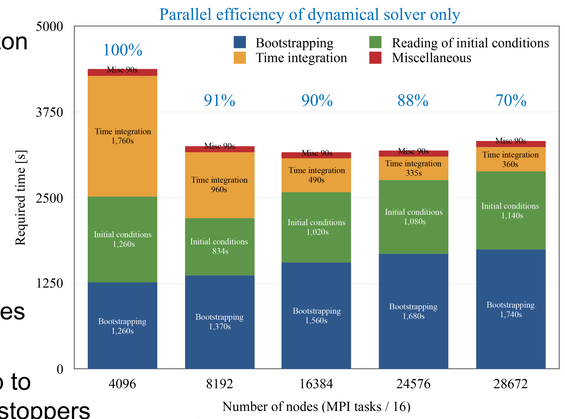
D. Heinzeller[1,2], M.G. Duda[3], H. Kunstmann[1,2], 2016: Geosci. Model Dev., 9, 77-110, http://www.geosci-model-dev.net/9/77/2016

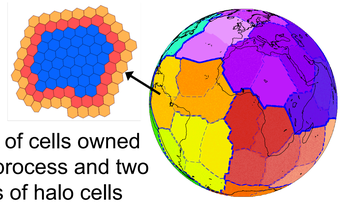
Convection-permitting global model applications are the next grand challenge in NWP and are on the horizon of next-generation, massively parallel HPC systems.

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

- Uniform 3km mesh, 65,536,002 columns
- 41 vertical layers, 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: 6.3x real-time, 1.6 Mio CPUh/24h

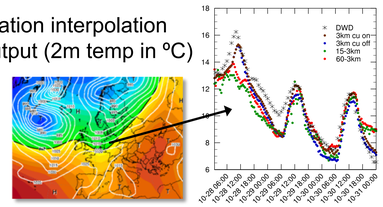
The dynamical solver of MPAS-Atmosphere scales up to 400,000 MPI tasks (160 owned cells per task). Show-stoppers at extreme scale are file I/O and model setup (bootstrapping).





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 start up

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	file input	612	52
2	time integration	1580	1658
2	file output	818	204

Timing results for a uniform 2km mesh with 147,456,002 columns 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

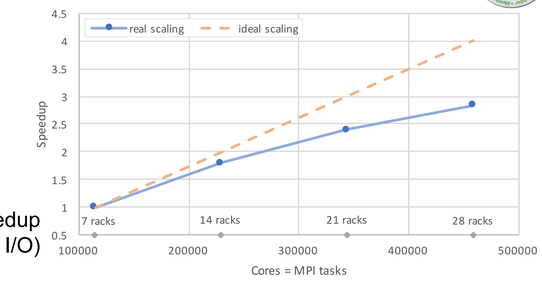
- Hybrid MPI+OpenMP parallelisation to speed up bootstrapping and decrease halo exchange time in dynamical solver of MPAS-A
- Optimisation for latest many-core architectures
- Combine with SIONlib I/O layer improvements for maximum performance at extreme scale
- Threading of one additional routine in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

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

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

- Jablonowski and Williamson (2006) baroclinic wave
- 2km mesh, 147,456,002 columns, single precision
- 26 vertical layers, 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



Simulation speedup (dynamics and file I/O)

Towards new horizons: global 1km model run of MPAS-A on HLRS Hazel Hen (with T. Schwitalla[4])

- Uniform 1km mesh with 589,824,002 columns
- Real-data forecast run, 55 vertical layers: 400TB mem.
- Mesh: 590GB pnetCDF, static data: 1.5TB SIONlib
- Initial conditions: 14TB SIONlib (read/write 135/476s)
- 10-min run (Δt=5s): 3854 nodes, 88642 MPI tasks, 7.5h!

© Boris Lehner for HLRS

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
- Reduce cost with careful application of variable-resolution meshes!

Kramer et al. 2018 (Clim. Dyn., under review): *Numerical Weather Prediction in the grey zone using a global variable resolution mesh and scale-aware convection parameterisation using MPAS.*

[1] Karlsruhe Institute of Technology, Institute of Meteorology and Climate Research, Garmisch-Partenkirchen, Germany
[2] University of Augsburg, Institute of Geography, Augsburg, Germany
[3] National Center for Atmospheric Research, Mesoscale and Microscale Meteorology Laboratory, Boulder, CO, USA
[4] University of Hohenheim, Institute of Physics and Meteorology, Stuttgart, Germany