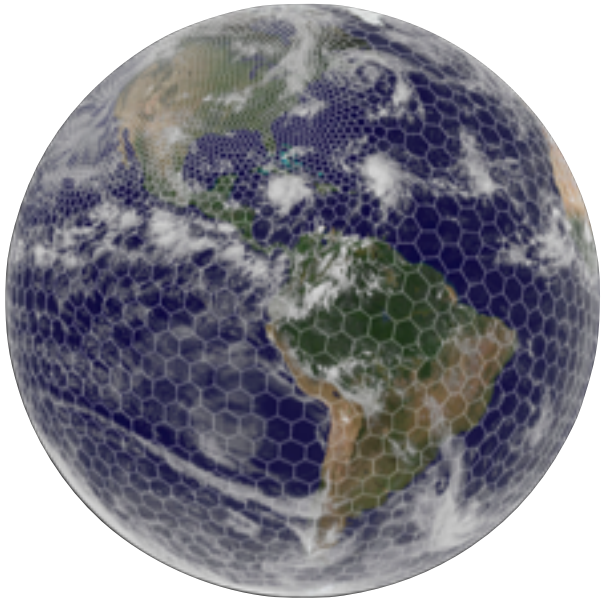
Karlsruhe Institute of Technology



MPAS Extreme Scaling Experiment

Towards convection-resolving, global climate simulations

[1] INSTITUTE OF METEOROLOGY AND CLIMATE RESEARCH, ATMOSPHERIC ENVIRONMENTAL RESEARCH (IMK-IFU)
Regional Climate Systems/Regional Climate and Hydrology

[2] NATIONAL CENTER FOR ATMOSPHERIC RESEARCH, EARTH SYSTEM LABORATORY (NESL)
Mesoscale Prediction

Third JUQUEEN Porting and Tuning Workshop, 2th-6th February 2015

Dominikus Heinzeller[1] and Michael Duda[2]

Karlsruhe Institute of Technology

KIT-Campus Alpin
IMK-IFU: Atmospheric Environmental Research

NCAR
Earth System Laboratory

PRACE

MPAS - Model for Prediction Across Scales

Collaborative project (NCAR/LANL) for developing earth-system simulation components for use in climate, regional climate and weather studies.

MPAS-A
atmospheric component

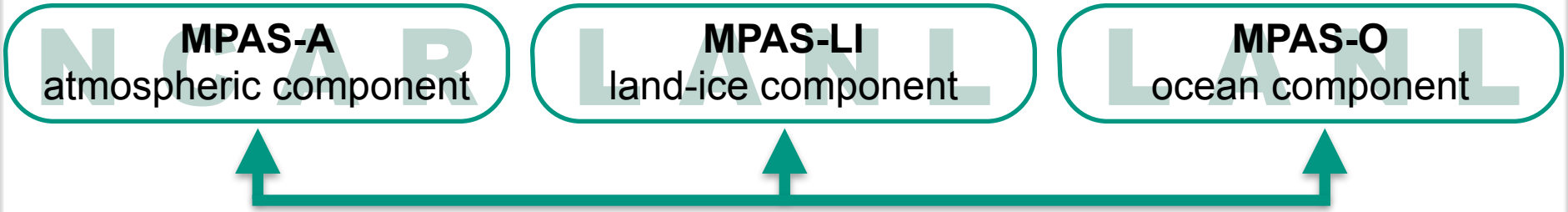
MPAS-LI
land-ice component

MPAS-O
ocean component

At present no coupling between the MPAS components. Work is underway, but a release is not expected in the near future.

MPAS - Model for Prediction Across Scales

Collaborative project (NCAR/LANL) for developing earth-system simulation components for use in climate, regional climate and weather studies.



At present no coupling between the MPAS components. Work is underway, but a release is not expected in the near future.

Directory

MPAS Website: http://mpas-dev.github.io

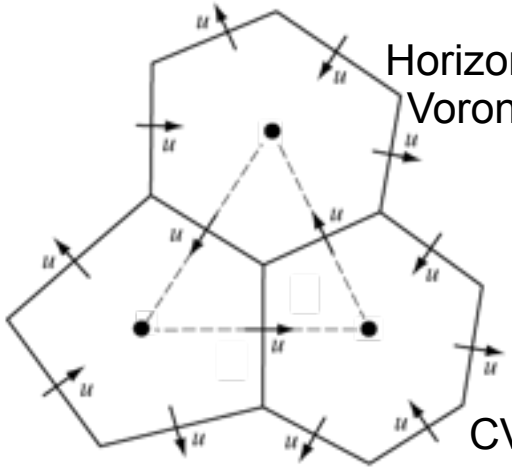
Skamarock et al.: A Multi-scale Nonhydrostatic Atmospheric Model Using Centroidal Voronoi Tesselations and C-Grid Staggering. 2012 MWR, 240, 3090–3105

Ringler et al: Ocean Modelling. 2013 Ocean Modelling, 69(C), 211–232

MPAS in a nutshell

MPAS
Model for Prediction Across Scales

Key features of MPAS



Horizontal, unstructured Voronoi mesh (SCVTs)

C-grid staggering of state variables

Dual Delaunay CVT grid (triangles)

Variable resolution meshes, smooth transitions

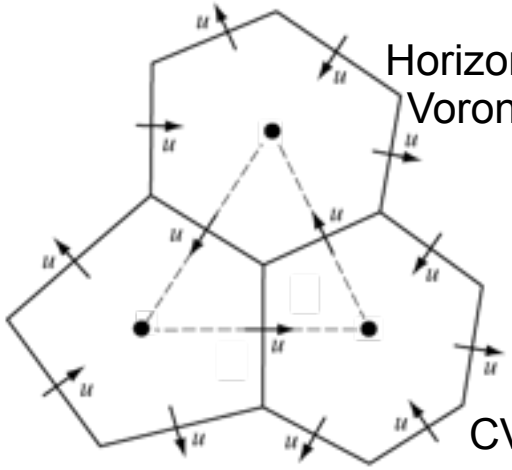
NCAR Earth System Laboratory

KIT Karlsruhe Institute of Technology

MPAS in a nutshell

MPAS
Model for Prediction Across Scales

Key features of MPAS



Horizontal, unstructured Voronoi mesh (SCVTs)

C-grid staggering of state variables

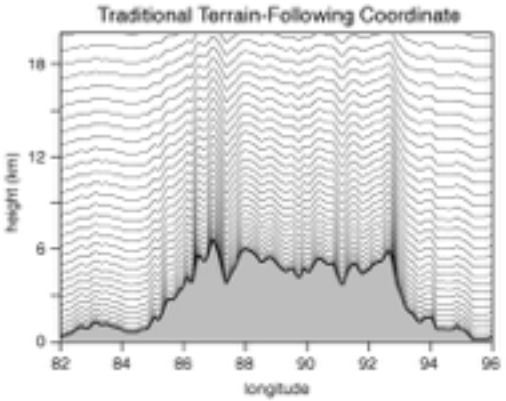
Dual Delaunay CVT grid (triangles)

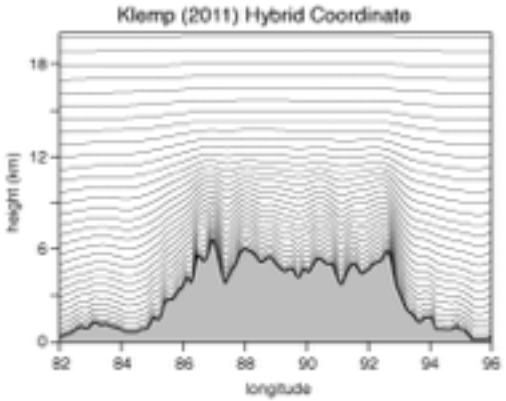
Variable resolution meshes, smooth transitions

Key features of MPAS-A

Split-explicit time integration scheme

Numerical schemes ported from WRF, work underway to make scale-aware





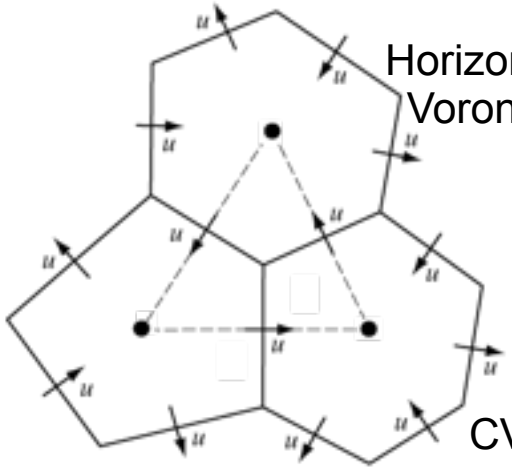
NCAR Earth System Laboratory

KIT Karlsruhe Institute of Technology

MPAS in a nutshell

MPAS
Model for Prediction Across Scales

Key features of MPAS



Horizontal, unstructured Voronoi mesh (SCVTs)

C-grid staggering of state variables

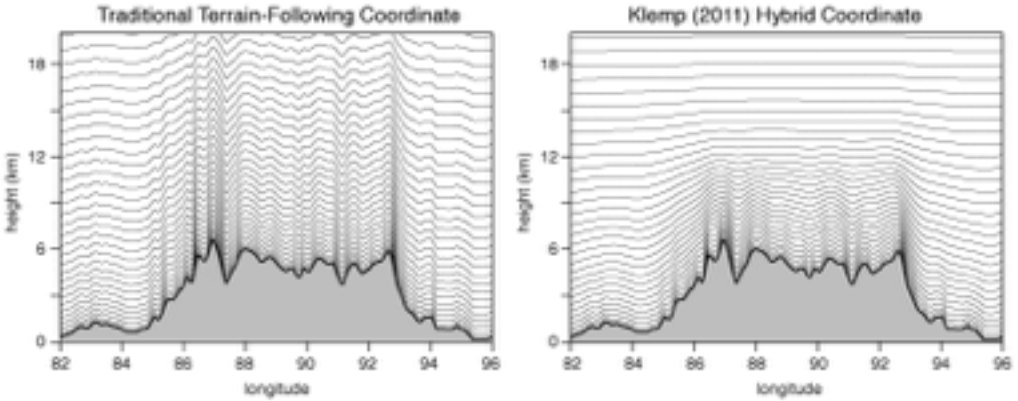
Dual Delaunay CVT grid (triangles)

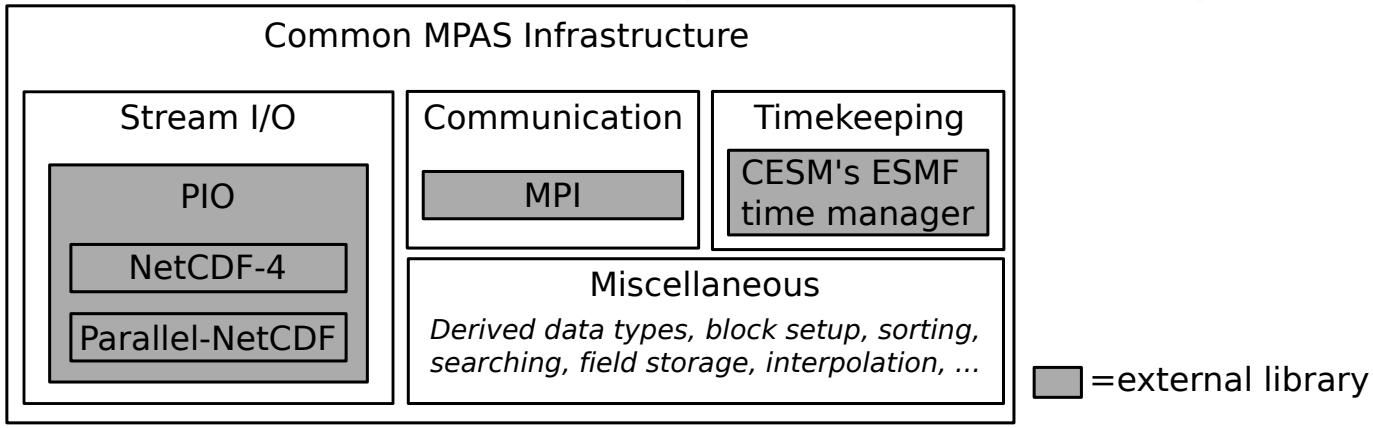
Variable resolution meshes, smooth transitions

Key features of MPAS-A

Split-explicit time integration scheme

Numerical schemes ported from WRF, work underway to make scale-aware



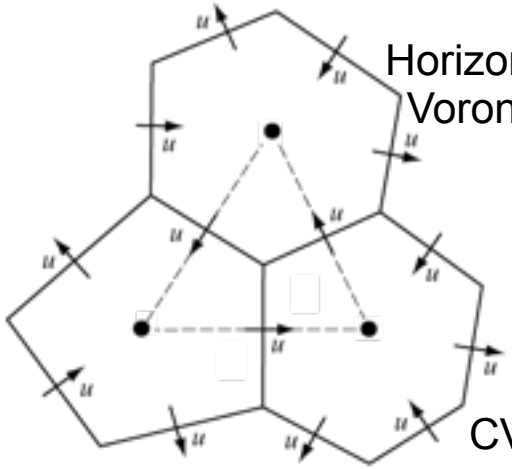


NCAR Earth System Laboratory
KIT Karlsruhe Institute of Technology

MPAS in a nutshell

MPAS
Model for Prediction Across Scales

Key features of MPAS



Horizontal, unstructured Voronoi mesh (SCVTs)

C-grid staggering of state variables

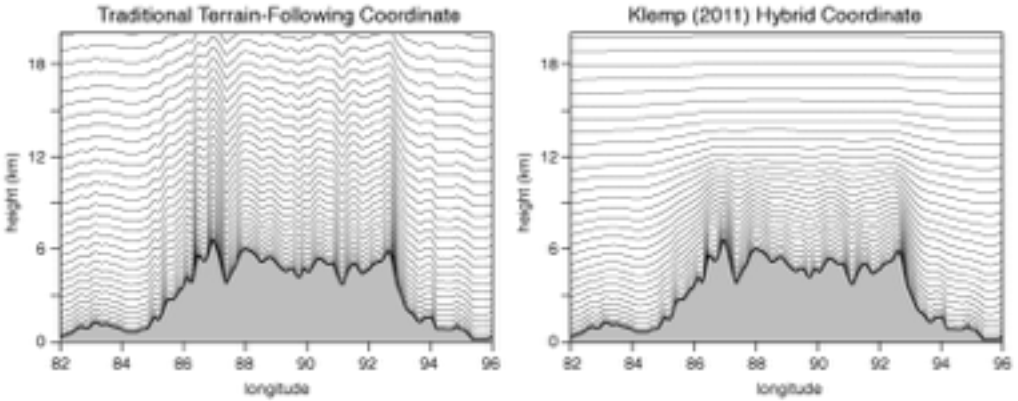
Dual Delaunay CVT grid (triangles)

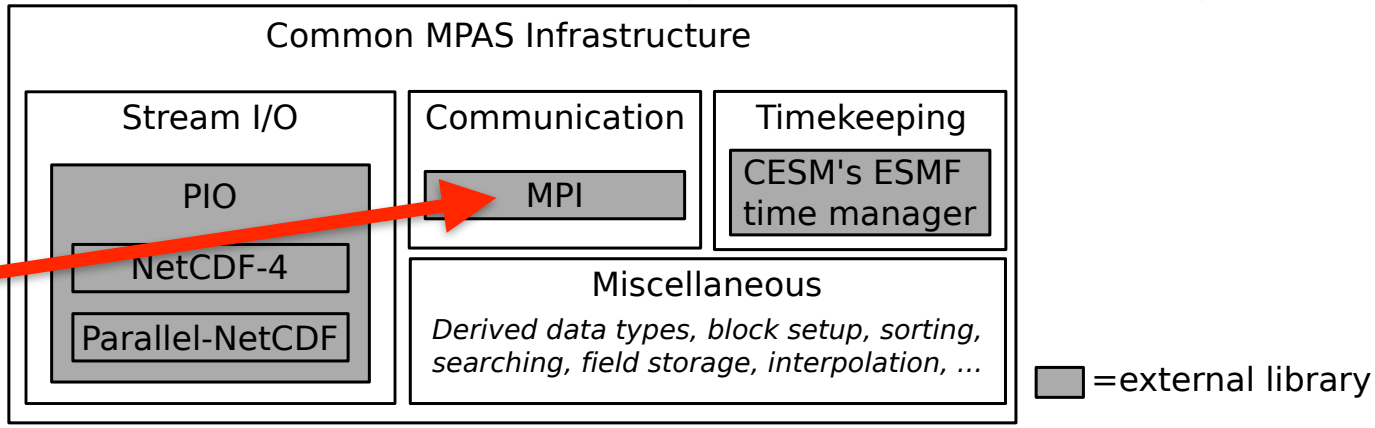
Variable resolution meshes, smooth transitions

Key features of MPAS-A

Split-explicit time integration scheme

Numerical schemes ported from WRF, work underway to make scale-aware





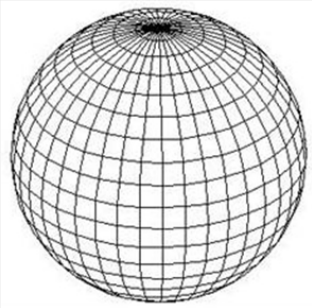
Up to now, pure MPI only (no OpenMP)

NCAR Earth System Laboratory
KIT Karlsruhe Institute of Technology

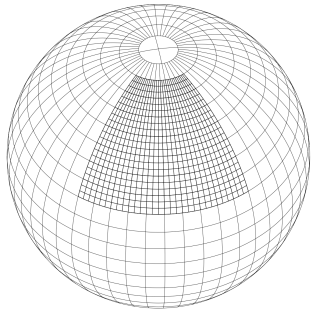
Variable resolution meshes and filtering

MPAS
Model for Prediction Across Scales

WRF Characteristics



- Lat-Lon global grid
  - Anisotropic grid cells
  - Polar filtering required



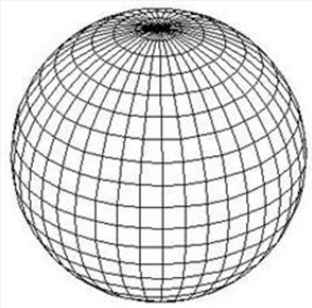
- Grid refinement through domain nesting
  - Flow distortions at nest boundaries

Credits: Bill Skamarock

NCAR Earth System Laboratory
KIT Karlsruhe Institute of Technology

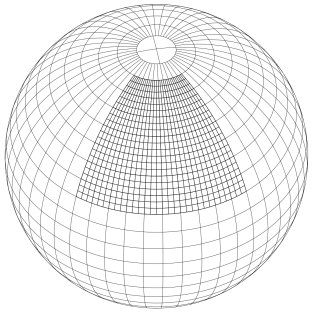
Variable resolution meshes and filtering

MPAS
Model for Prediction Across Scales



WRF Characteristics

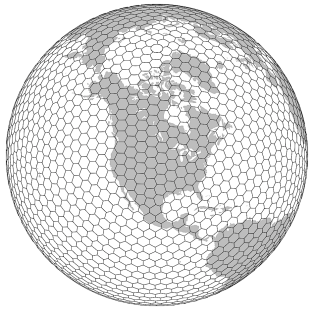
- Lat-Lon global grid
  - Anisotropic grid cells
  - Polar filtering required



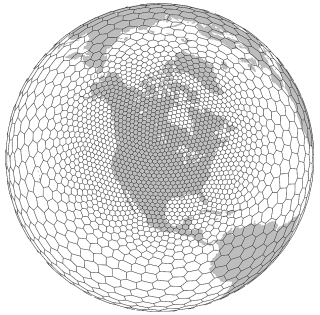
- Grid refinement through domain nesting
  - Flow distortions at nest boundaries

MPAS Characteristics

- Unstructured Voronoi (hexagonal) grid
  - Good scaling on massively parallel computers



- Smooth grid refinement on a conformal mesh
  - Increased accuracy and flexibility in varying resolution

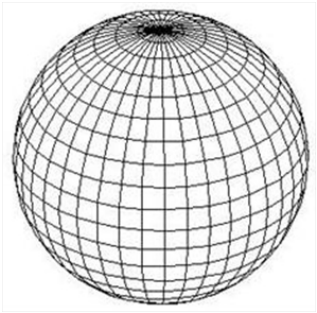


Credits: Bill Skamarock

NCAR Earth System Laboratory
KIT Karlsruhe Institute of Technology

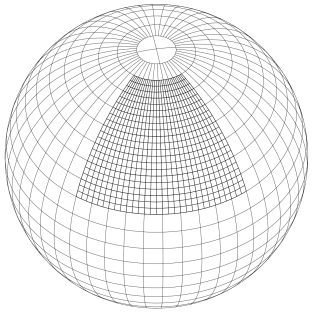
Variable resolution meshes and filtering

MPAS
Model for Prediction Across Scales



WRF Characteristics

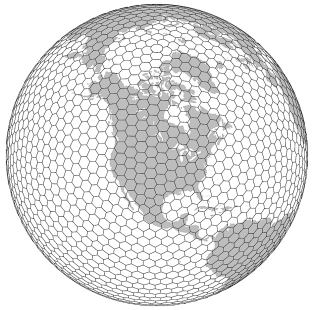
- Lat-Lon global grid
  - Anisotropic grid cells
  - Polar filtering required



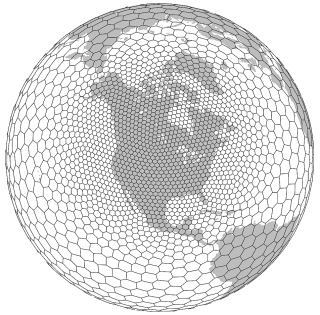
- Grid refinement through domain nesting
  - Flow distortions at nest boundaries

MPAS Characteristics

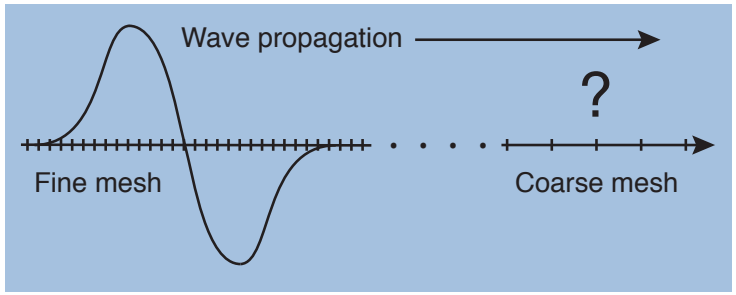
- Unstructured Voronoi (hexagonal) grid
  - Good scaling on massively parallel computers



- Smooth grid refinement on a conformal mesh
  - Increased accuracy and flexibility in varying resolution



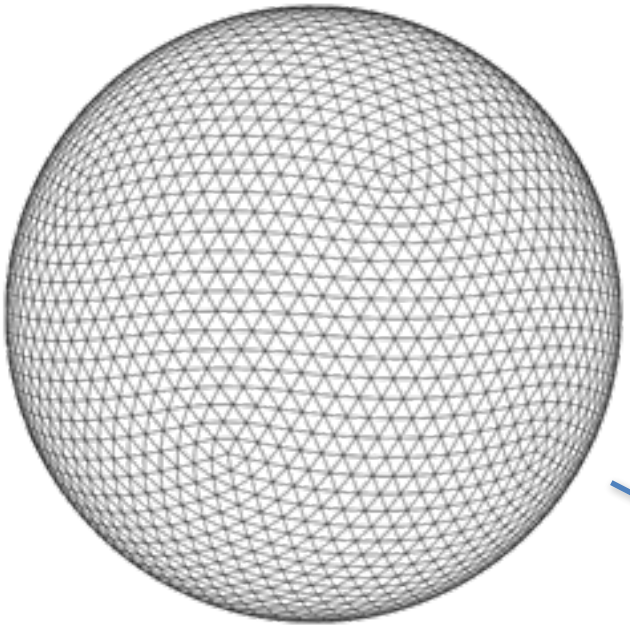
- Short-wavelength modes will be reflected in a fine-coarse mesh transition if not filtered.
- Smooth transitions result in reflection of only the very-shortest wavelength modes (locally), where filters are efficient.
- Abrupt transitions result in reflection of longer waves, where filters are less efficient.
- In MPAS: Smagorinski, viscosity, hyperviscosity filters



Credits: Bill Skamarock

NCAR Earth System Laboratory
KIT Karlsruhe Institute of Technology

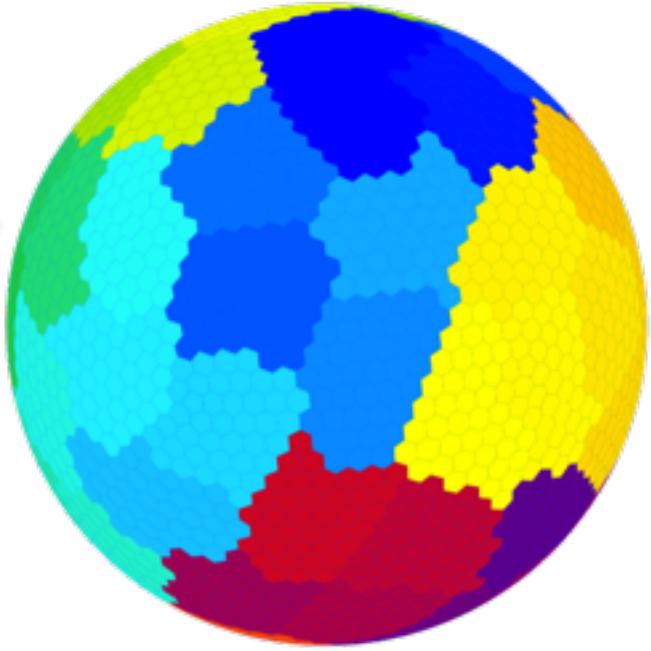
Parallel decomposition



The dual mesh of a Voronoi tessellation is a Delaunay triangulation – the connectivity graph of the cells.

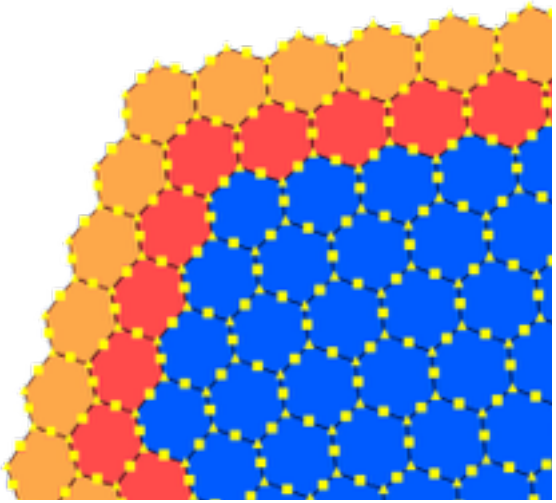
Parallel decomposition of an MPAS mesh becomes a graph partitioning problem: equally distribute nodes among partitions (each process gets equal work) while minimizing the edge cut (minimizing communication).

Graph partitioning



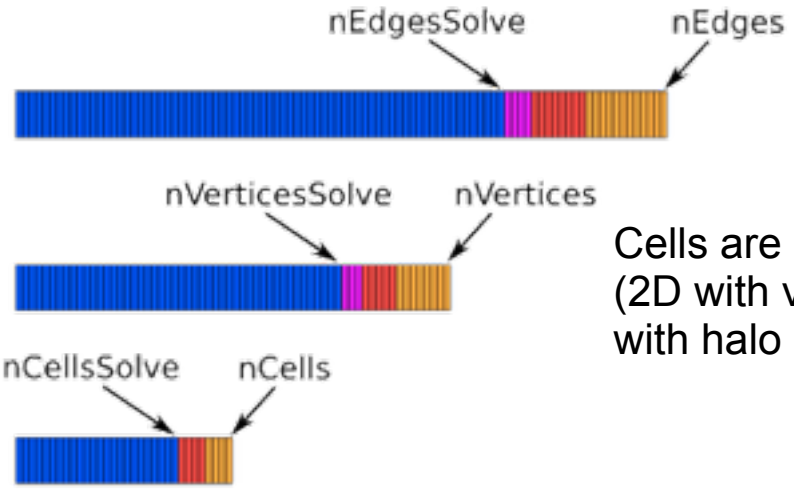
Parallel graph decomposition is done in a pre-processing step using the Metis package (can also handle weighted graph partitioning).

Parallel decomposition (continued)



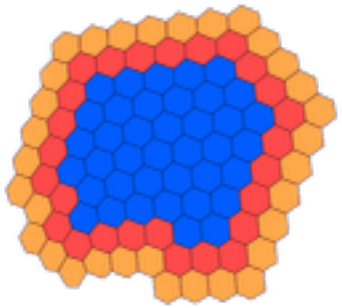
Given an assignment of cells to a process, any number of layers of halo (ghost) cells may be added.

With a list of cells stored in a block, adjacent edge and vertex locations can be found. A simple rule is applied to determine the ownership of edges and vertices adjacent to real cells in different blocks.



Cells are stored in a 1D array (2D with vertical dimension…), with halo cells at the end.

Block of cells owned by a process

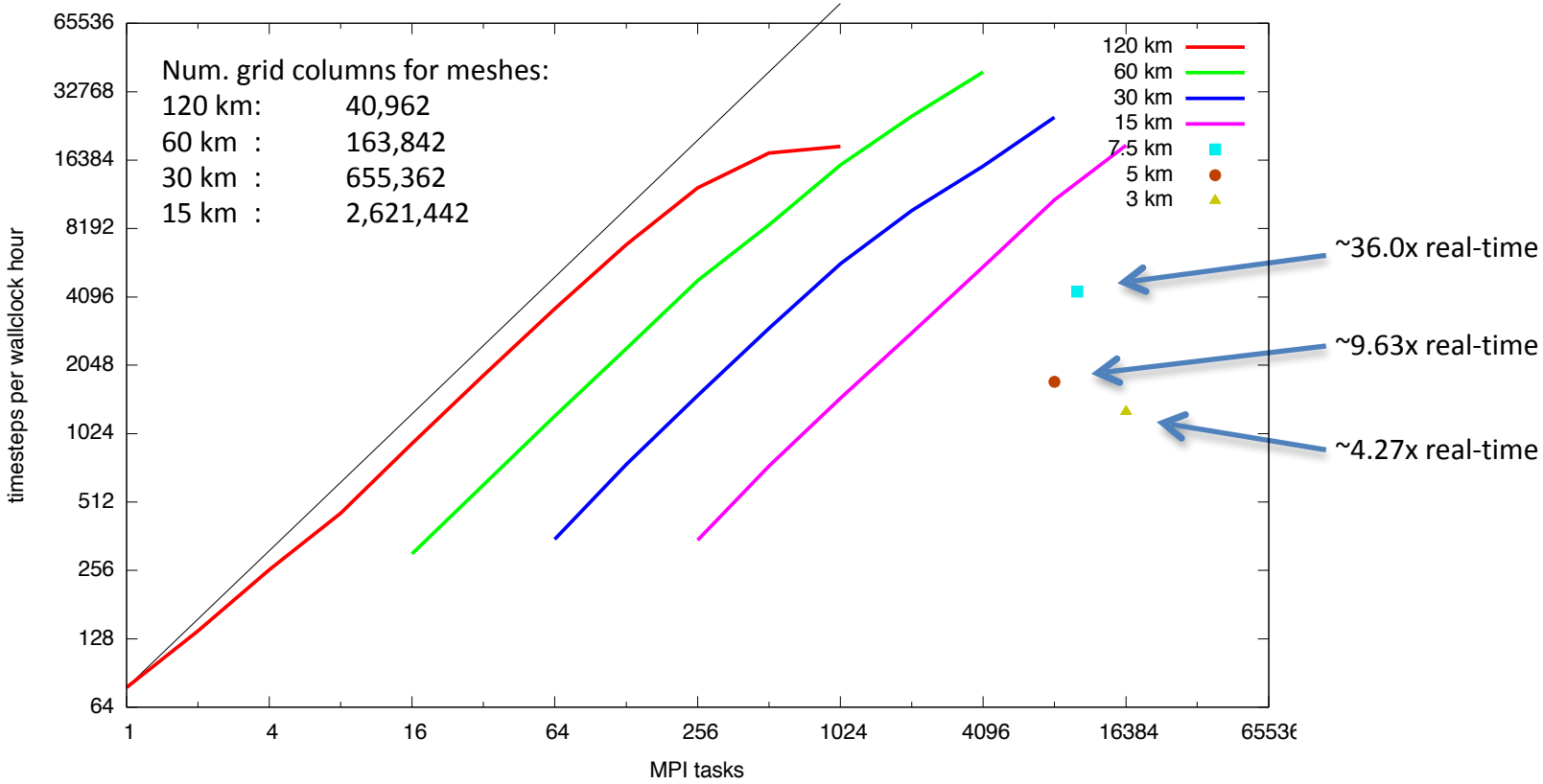


Block plus two layers of halo/ghost cells

6

Overall MPAS performance

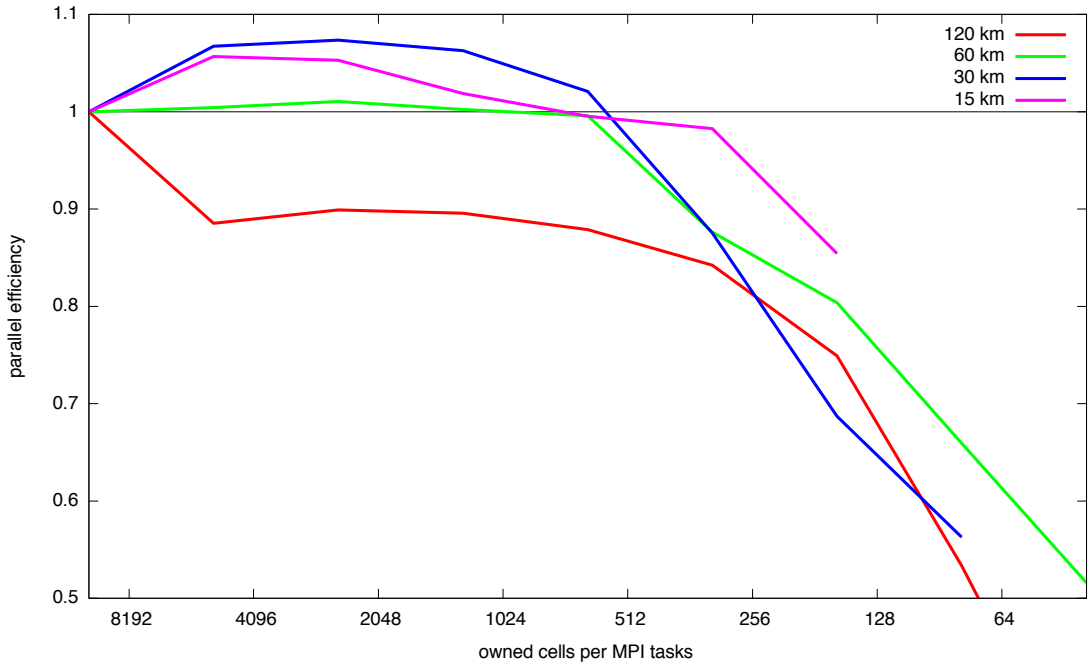
Consider performance in terms of number of model time steps completed in one wallclock hour (allows for easier comparison of performance and scaling).



Simulation rate of double-precision MPAS-A on NCAR's "Yellowstone" computer; 45 vertical layers (41 layers for 7.5, 5, and 3 km), full suite of physics, no I/O.

Overall MPAS-A parallel efficiency

The full MPAS-A solver (30-km, double-precision, physics+dynamics, no I/O) achieves 68.69% efficiency down to about 160 owned cells per MPI task.



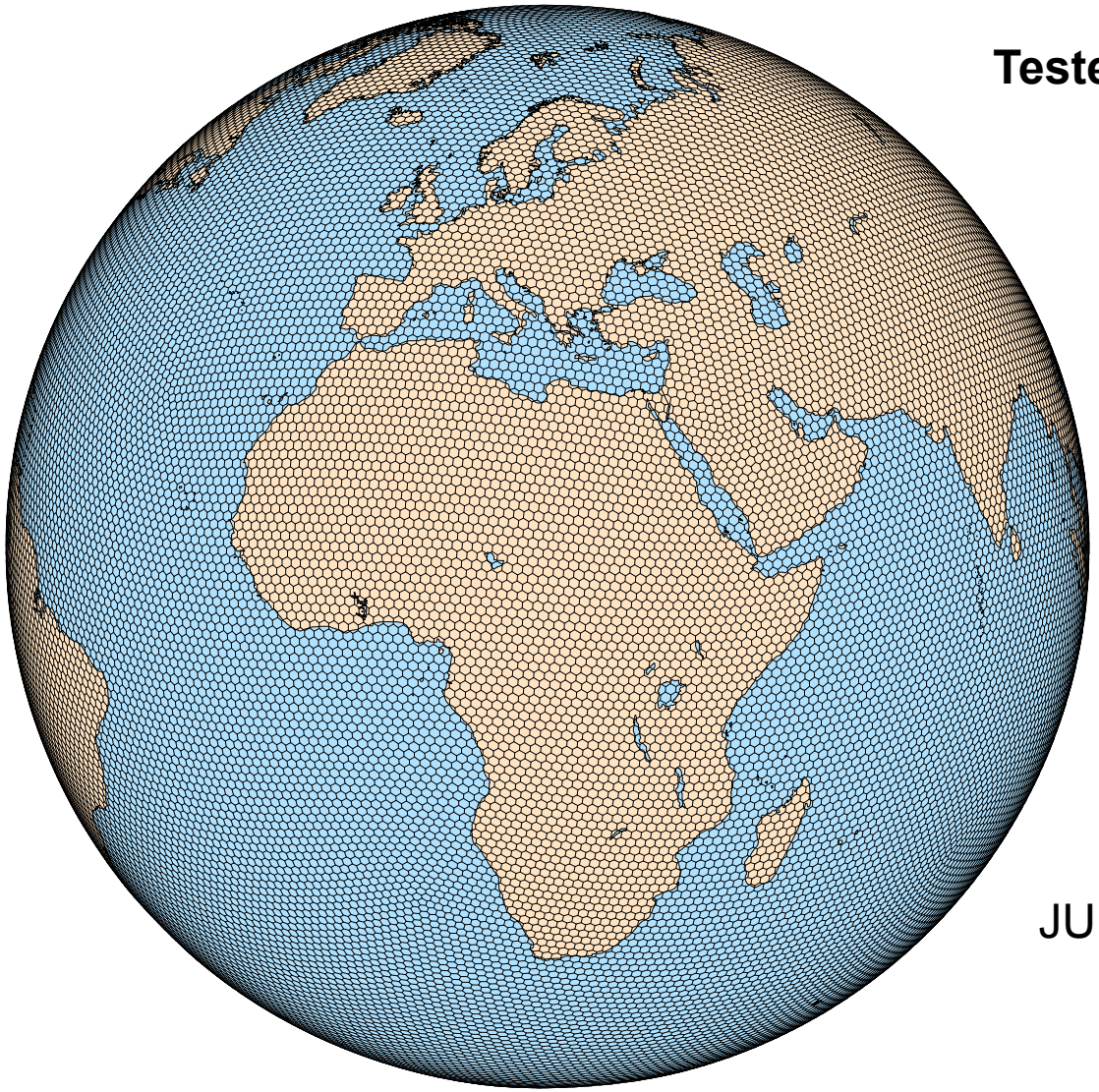
MPAS-A parallel efficiency relative to 10240 owned cells per MPI task, Yellowstone

MPI tasks	Cells per task	Speedup	Efficiency
64	*10240*	*1.00*	*100.0%*
128	*5120*	*2.13*	*106.7%*
256	*2560*	*4.29*	*107.3%*
512	*1280*	*8.50*	*106.2%*
1024	*640*	*16.33*	*102.1%*
2048	*320*	*28.00*	*87.52%*
4096	*160*	*43.96*	*68.69%*
8192	*80*	*72.03*	*56.27%*

MPAS-A scaling – 30-km mesh, Yellowstone (ncell = 655362)

Test case 1: regular 120km mesh x1.40962

MPAS
Model for Prediction Across Scales



Tested systems

TGCC
CURIE
(PRACE)

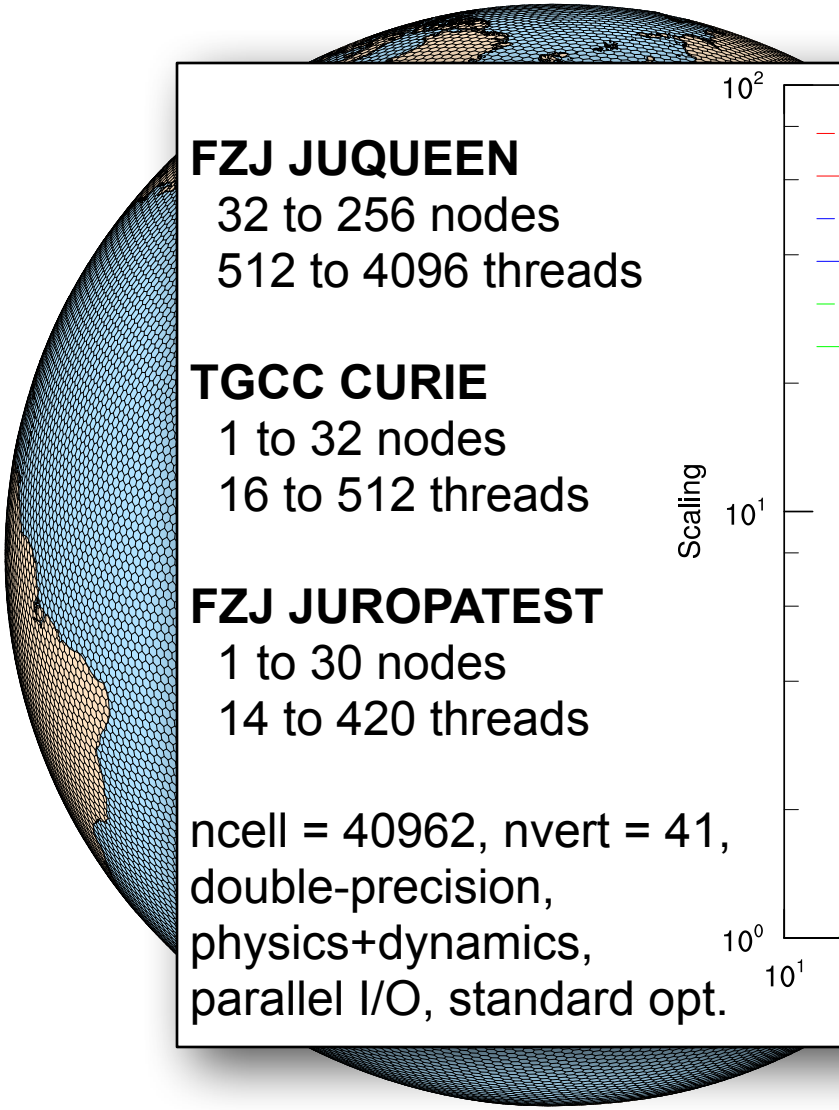
FZJ
JUQUEEN
(PRACE)

FZJ
JUROPATEST
(JUROPA)

NCAR Earth System Laboratory
KIT Karlsruhe Institute of Technology

Test case 1: regular 120km mesh x1.40962

MPAS
Model for Prediction Across Scales

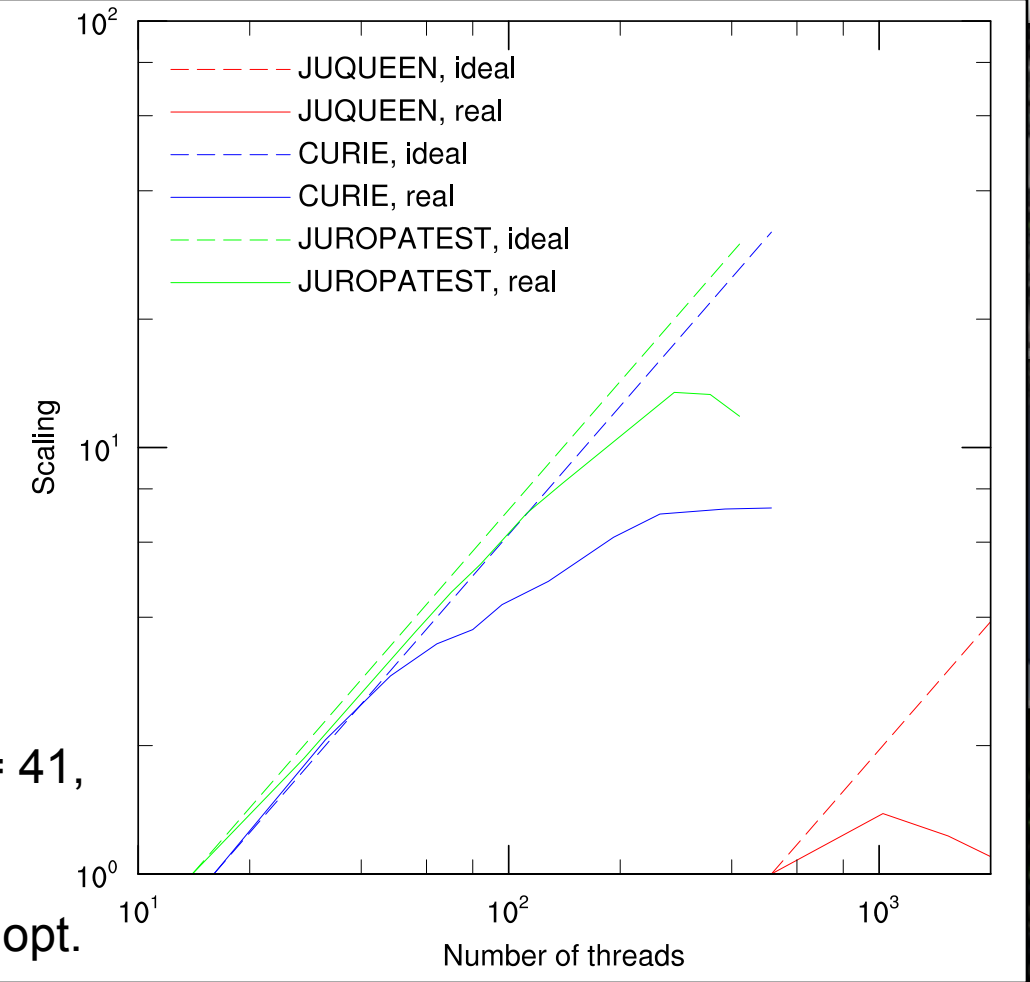


FZJ JUQUEEN
32 to 256 nodes
512 to 4096 threads

TGCC CURIE
1 to 32 nodes
16 to 512 threads

FZJ JUROPATEST
1 to 30 nodes
14 to 420 threads

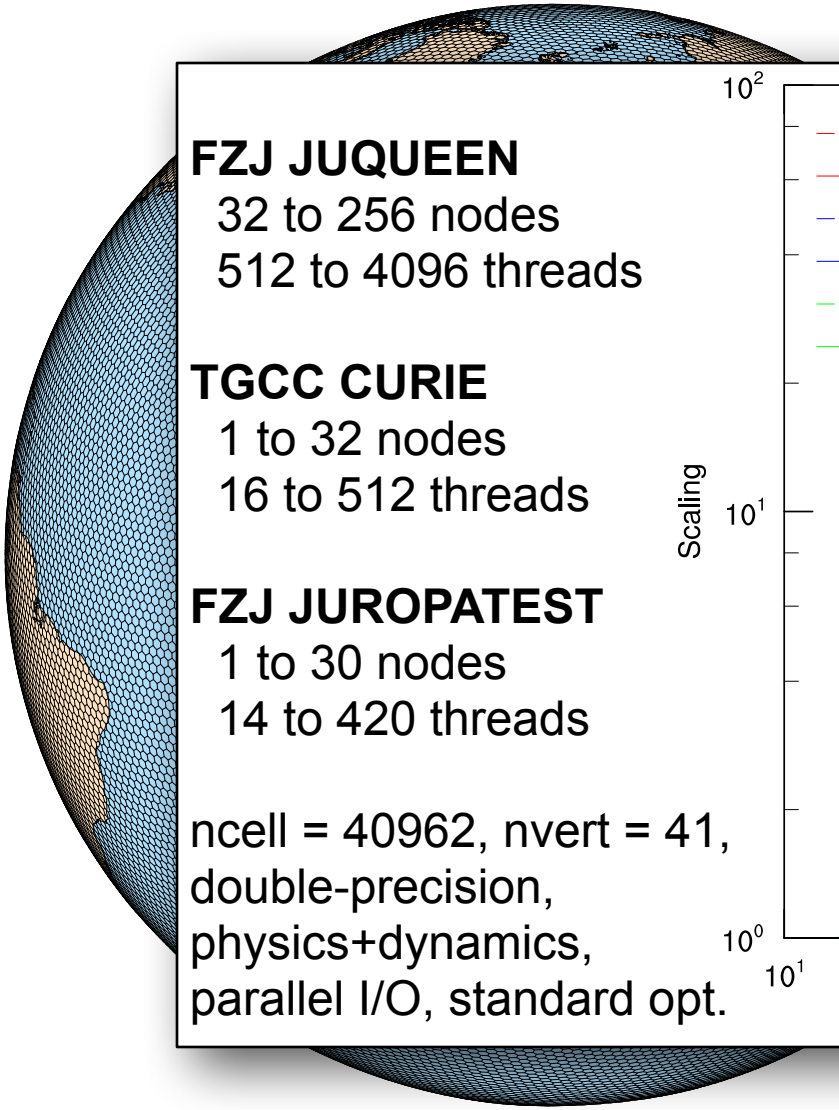
ncell = 40962, nvert = 41,
double-precision,
physics+dynamics,
parallel I/O, standard opt.



NCAR Earth System Laboratory
KIT Karlsruhe Institute of Technology

Test case 1: regular 120km mesh x1.40962

MPAS
Model for Prediction Across Scales

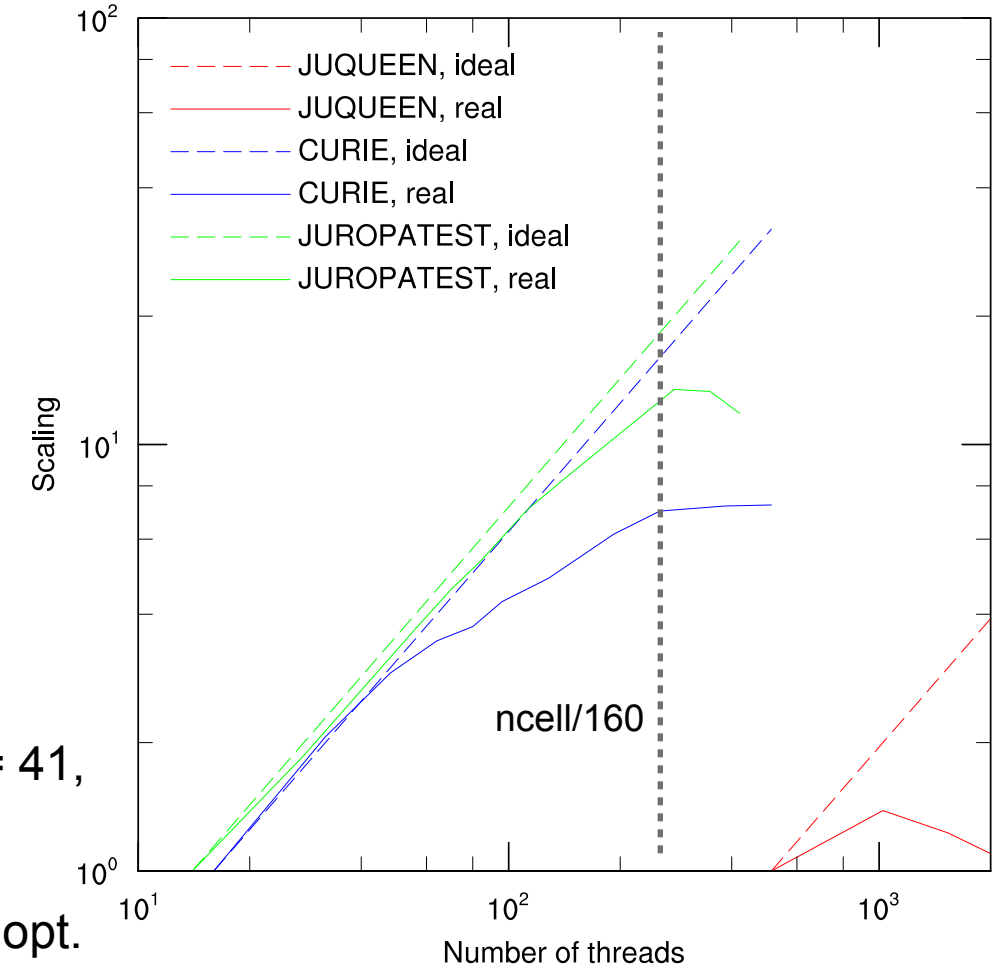


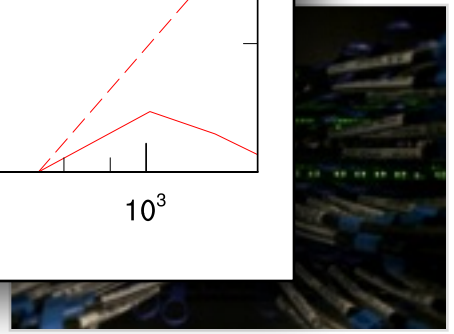
FZJ JUQUEEN
32 to 256 nodes
512 to 4096 threads

TGCC CURIE
1 to 32 nodes
16 to 512 threads

FZJ JUROPATEST
1 to 30 nodes
14 to 420 threads

ncell = 40962, nvert = 41,
double-precision,
physics+dynamics,
parallel I/O, standard opt.

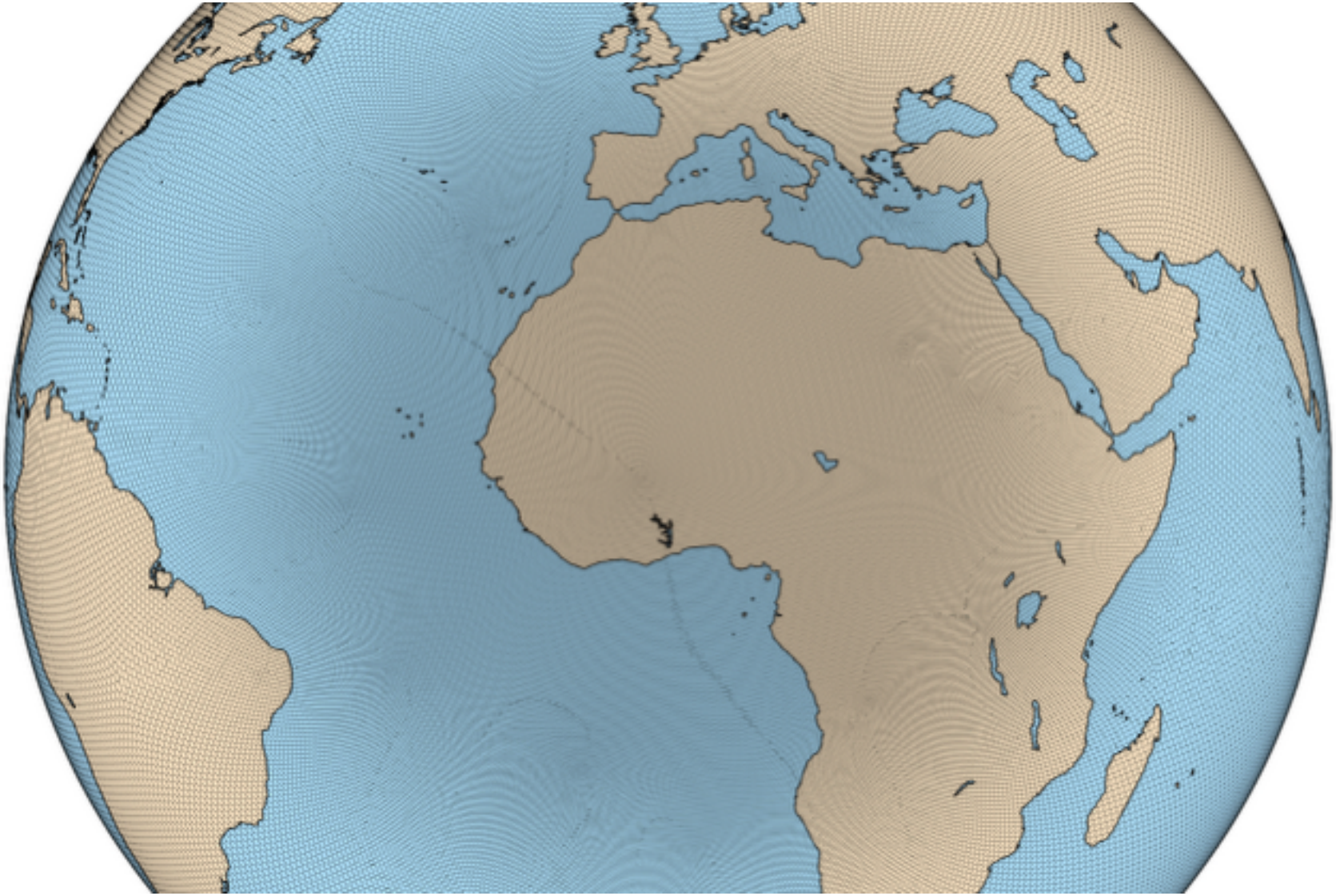




NCAR Earth System Laboratory
KIT Karlsruhe Institute of Technology

Test case 2: variable 120/25km mesh x4.163842

MPAS
Model for Prediction Across Scales

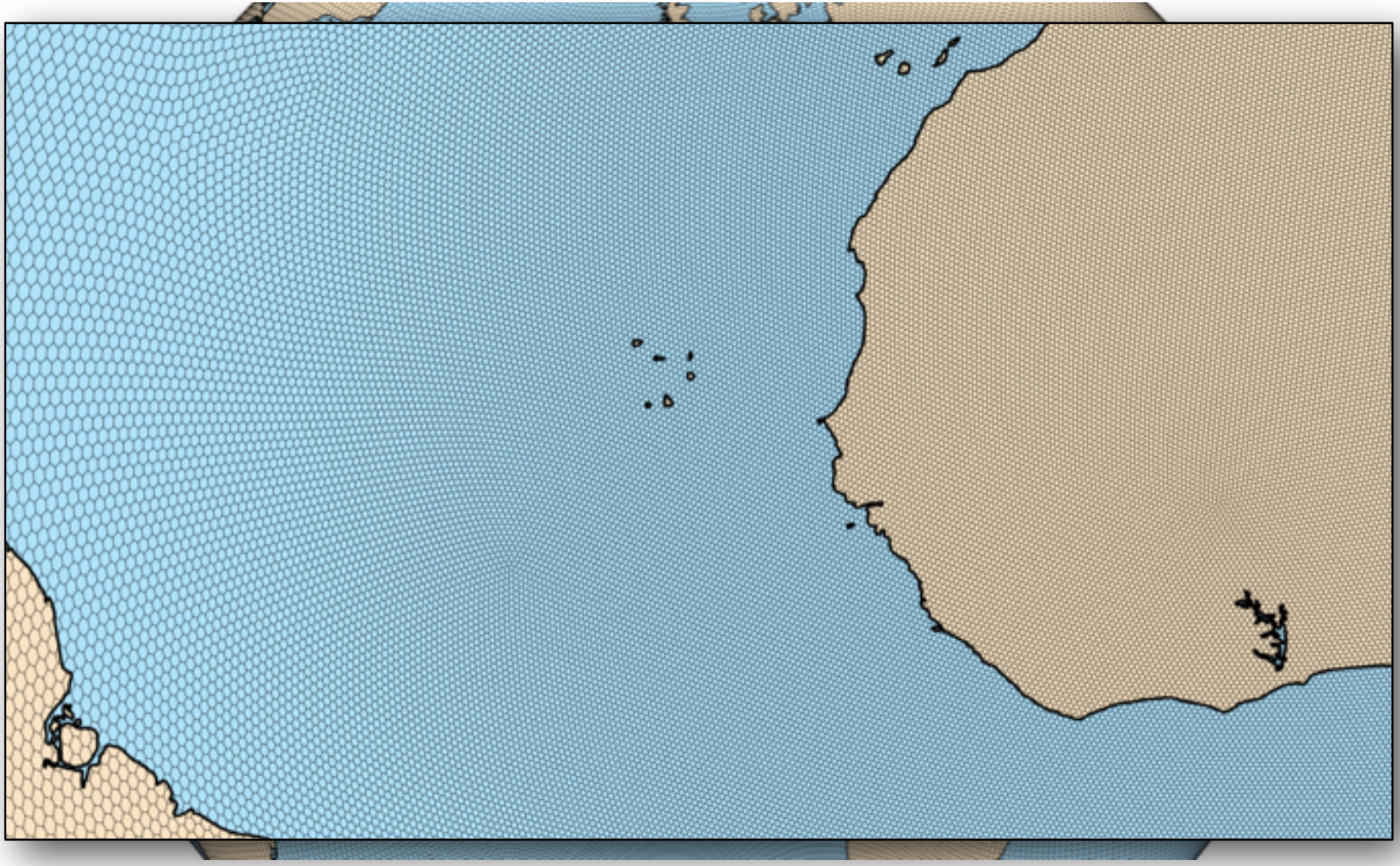


NCAR Earth System Laboratory

KIT Karlsruhe Institute of Technology

Test case 2: variable 120/25km mesh x4.163842

MPAS
Model for Prediction Across Scales



NCAR Earth System Laboratory

KIT Karlsruhe Institute of Technology

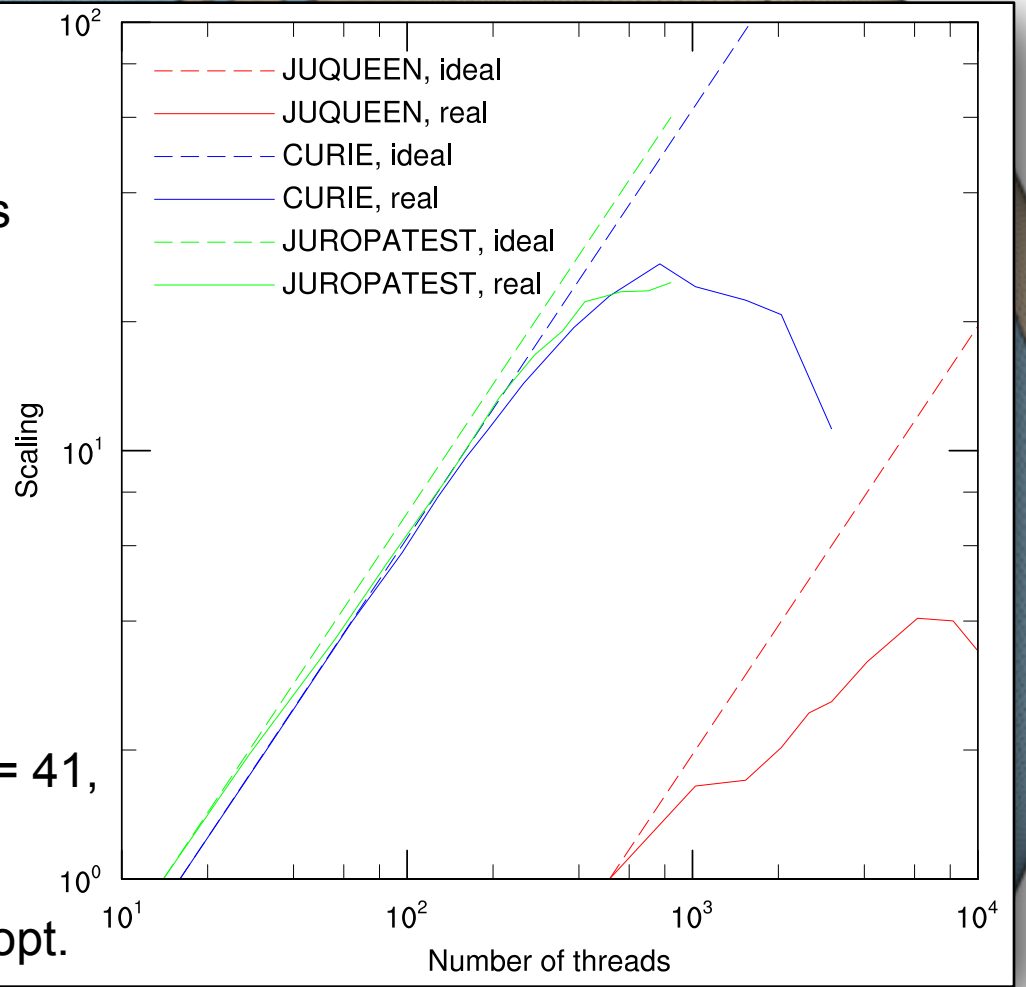
Test case 2: variable 120/25km mesh x4.163842

FZJ JUQUEEN
32 to 1024 nodes
512 to 16384 threads

TGCC CURIE
1 to 192 nodes
16 to 3072 threads

FZJ JUROPATEST
1 to 60 nodes (max)
14 to 840 threads

ncell = 163842, nvert = 41,
double-precision,
physics+dynamics,
parallel I/O, standard opt.



Test case 2: variable 120/25km mesh x4.163842

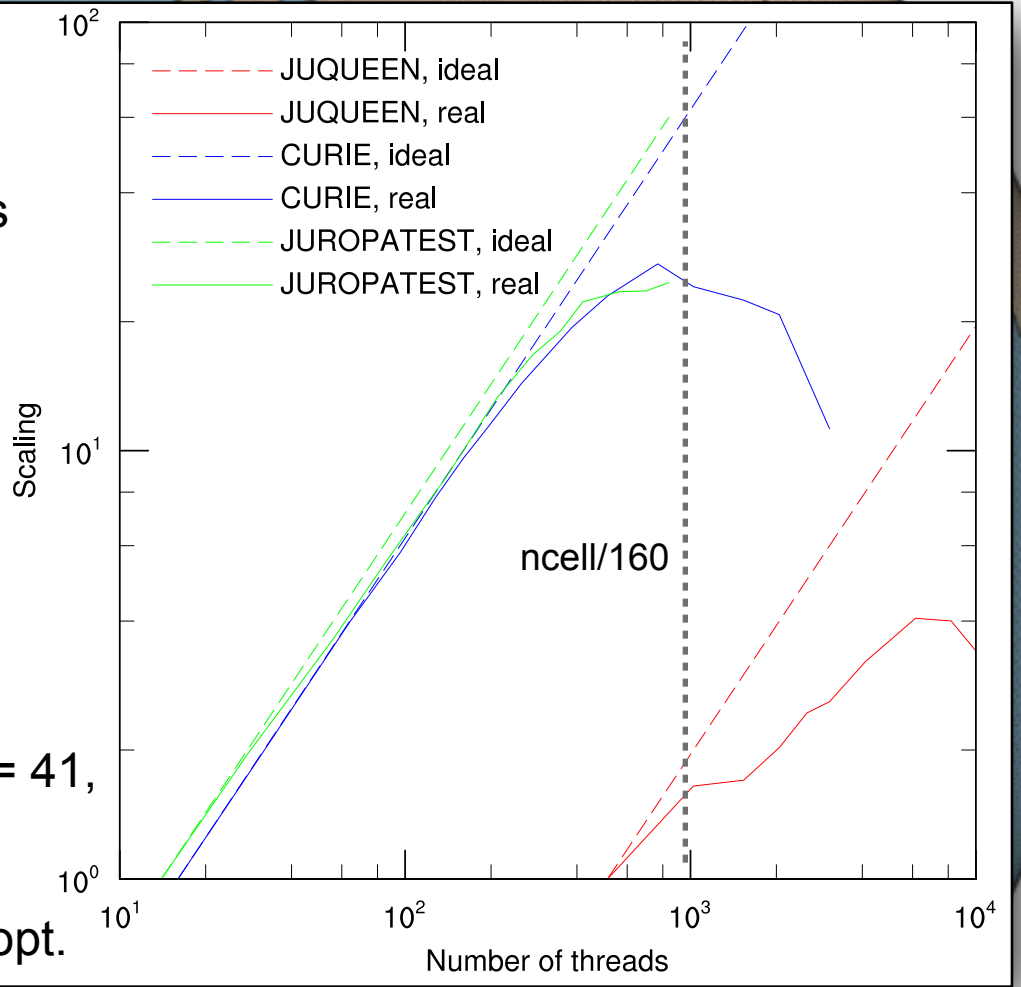
MPAS
Model for Prediction Across Scales

FZJ JUQUEEN
32 to 1024 nodes
512 to 16384 threads

TGCC CURIE
1 to 192 nodes
16 to 3072 threads

FZJ JUROPATEST
1 to 60 nodes (max)
14 to 840 threads

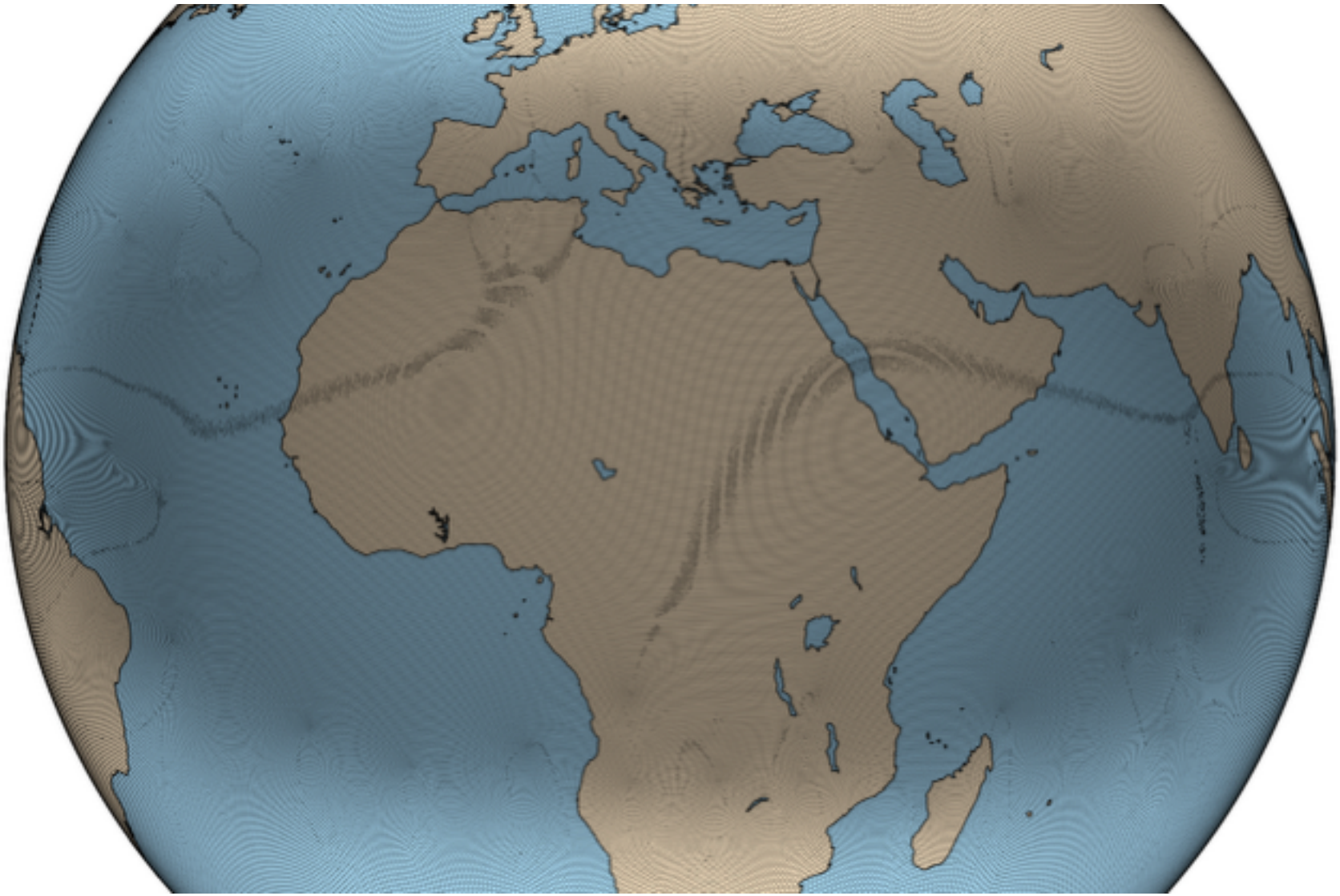
ncell = 163842, nvert = 41,
double-precision,
physics+dynamics,
parallel I/O, standard opt.



NCAR Earth System Laboratory
KIT Karlsruhe Institute of Technology

Test case 3: variable 90/12km mesh x4.535554

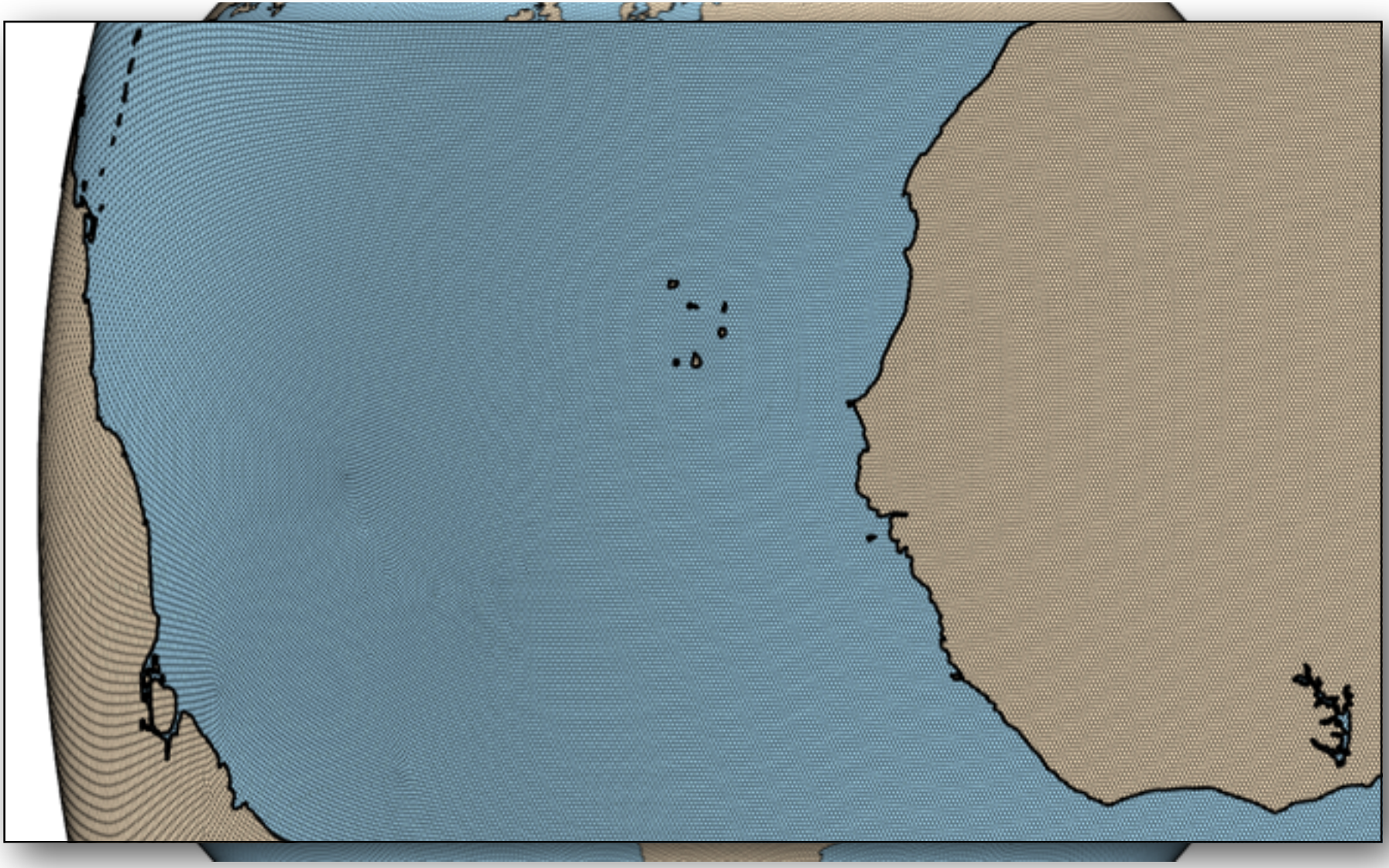
MPAS
Model for Prediction Across Scales



NCAR Earth System Laboratory
KIT Karlsruhe Institute of Technology

Test case 3: variable 90/12km mesh x4.535554

MPAS
Model for Prediction Across Scales



NCAR Earth System Laboratory

KIT Karlsruhe Institute of Technology

Test case 3: variable 90/12km mesh x4.535554

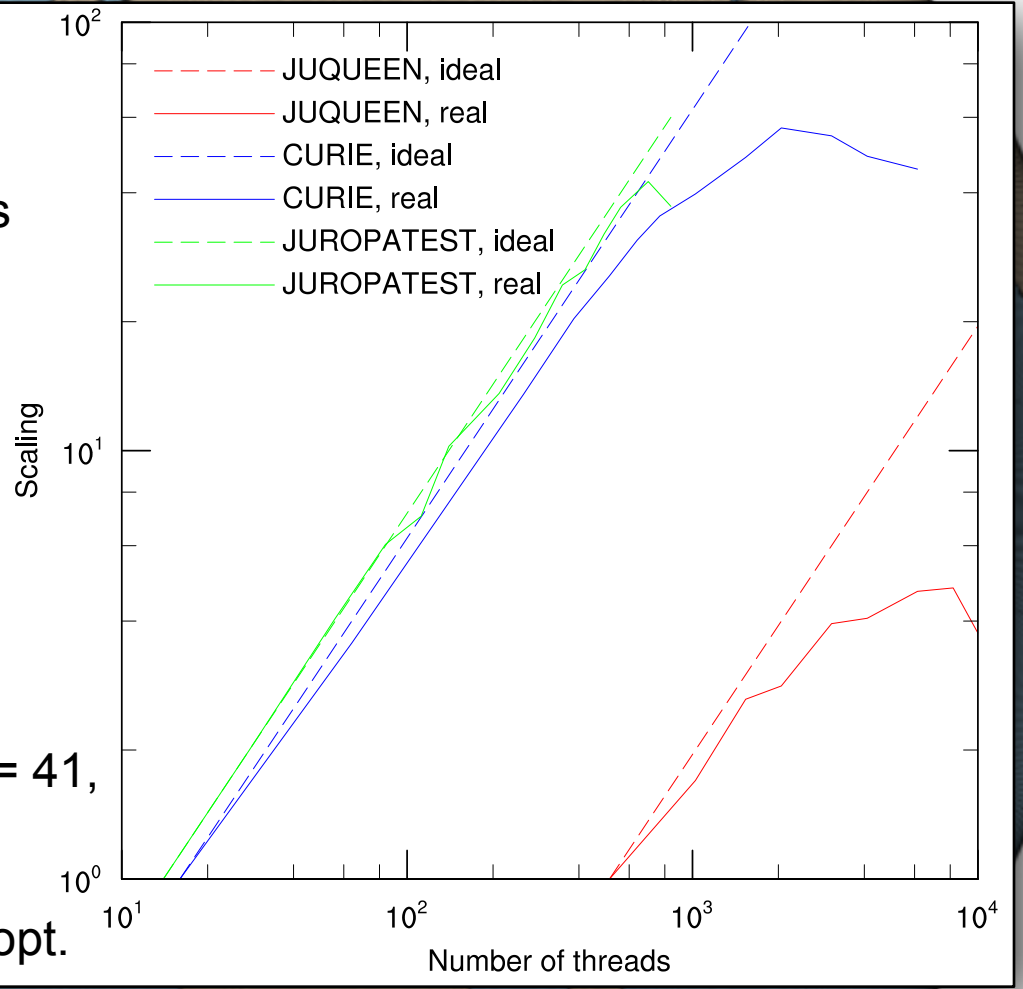
MPAS
Model for Prediction Across Scales

FZJ JUQUEEN
32 to 1024 nodes
512 to 16384 threads

TGCC CURIE
1 to 384 nodes
16 to 6144 threads

FZJ JUROPATEST
1 to 60 nodes (max)
14 to 840 threads

ncell = 535554, nvert = 41,
double-precision,
physics+dynamics,
parallel I/O, standard opt.



NCAR Earth System Laboratory
KIT Karlsruhe Institute of Technology

Test case 3: variable 90/12km mesh x4.535554

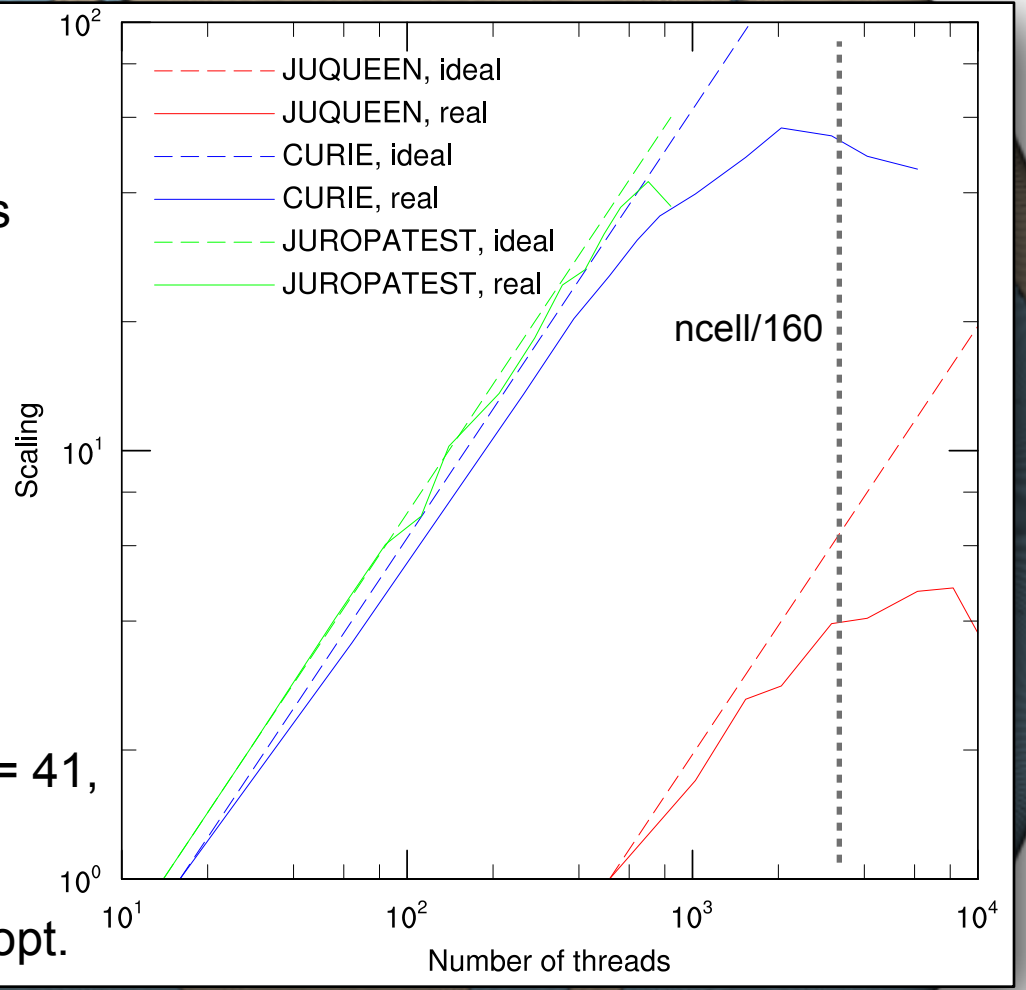
MPAS
Model for Prediction Across Scales

FZJ JUQUEEN
32 to 1024 nodes
512 to 16384 threads

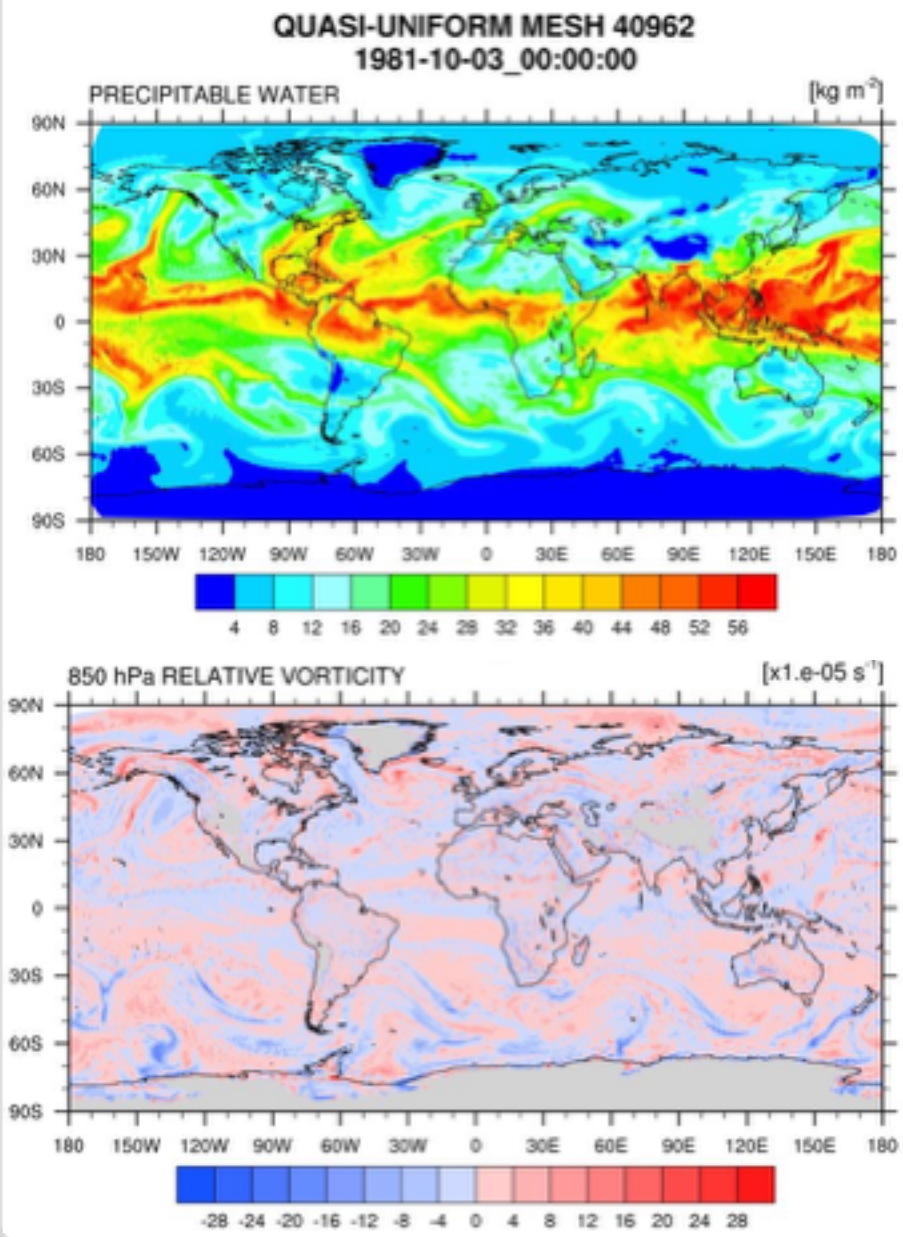
TGCC CURIE
1 to 384 nodes
16 to 6144 threads

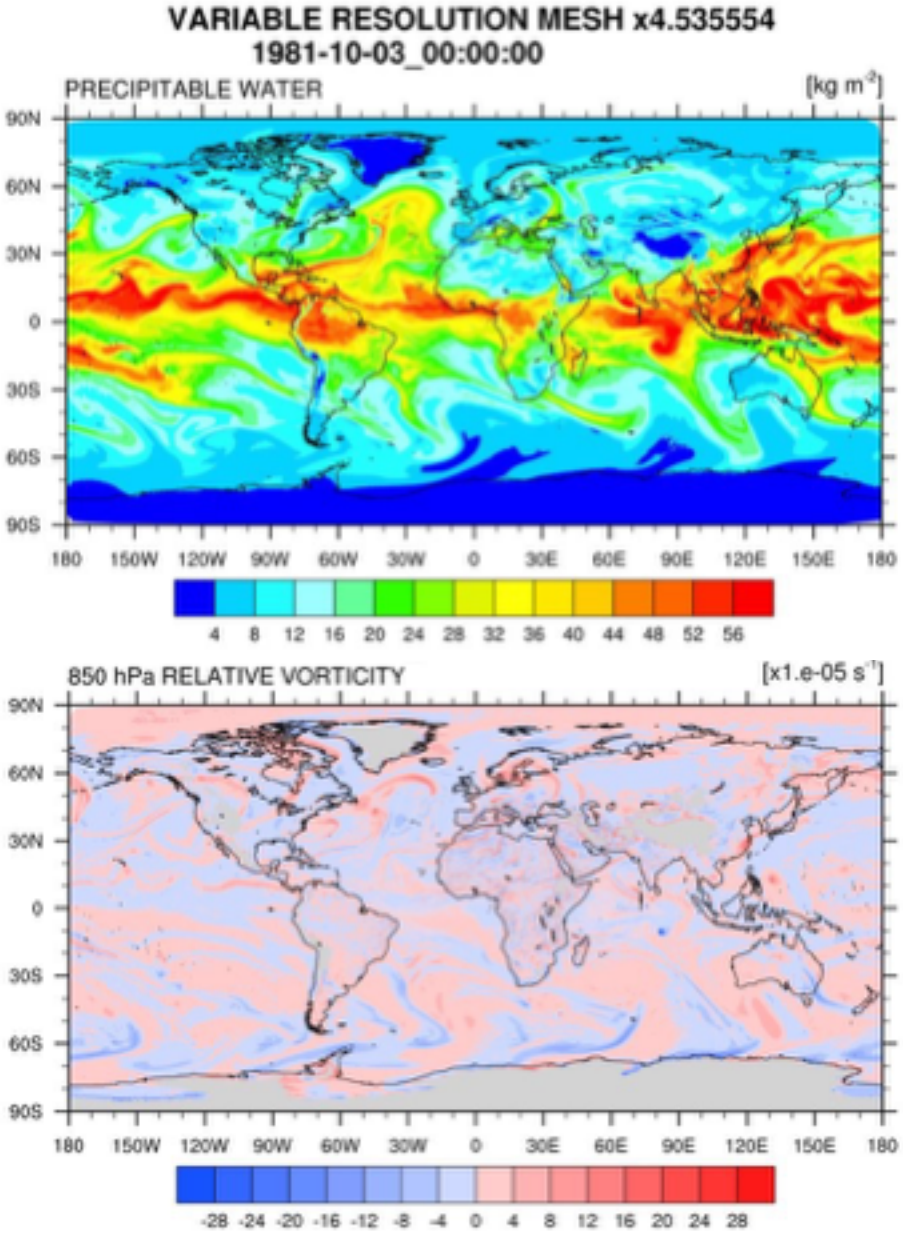
FZJ JUROPATEST
1 to 60 nodes (max)
14 to 840 threads

ncell = 535554, nvert = 41,
double-precision,
physics+dynamics,
parallel I/O, standard opt.



NCAR Earth System Laboratory
KIT Karlsruhe Institute of Technology





NCAR Earth System Laboratory

KIT Karlsruhe Institute of Technology

Extreme scaling: regular 3km mesh x1.65536002

MPAS Model for Prediction Across Scales

Key data

ncell = 65536002, nvert = 41	expect ≥70% parallel efficiency up to 400000 threads (ncell/160)
with parallel I/O and without I/O	test scaling of dynamical solver and communication separately from I/O
simulation rate about 2x real-time (rt) on Yellowstone (16384 cores)	expect 1/2.5x rt on JUQUEEN (np=16384) to ≥6.5x rt (np≥393216, i.e., 24576 nodes)
double precision, physics+dynamics: memory requirement 0.175Mb per cell	11.5Tb of memory required for 3km mesh, minimum number of nodes is approx. 750
I/O sizes: initial condition file 1.1Tb, restart file 2.1Tb, history file 250Gb	

Notes

Pure MPI parallelization in MPAS → Trial OpenMP hybrid mode during tuning and porting workshop (OpenMP test code in place)

Expect breakdown of parallel efficiency at about 400000 threads → Exploit limiting factors in the model when the scaling breaks for extreme numbers of threads

Extreme scaling: regular 3km mesh x1.65536002

MPAS Model for Prediction Across Scales

Key data

ncell = 65536002, nvert = 41	expect ≥70% parallel efficiency up to 400000 threads (ncell/160)
with parallel I/O and without I/O	test scaling of dynamical solver and communication separately from I/O
simulation rate about 2x real-time (rt) on Yellowstone (16384 cores)	expect 1/2.5x rt on JUQUEEN (np=16384) to ≥6.5x rt (np≥393216, i.e., 24576 nodes)
double precision, physics+dynamics: memory requirement 0.175Mb per cell	11.5Tb of memory required for 3km mesh, minimum number of nodes is approx. 750
I/O sizes: initial condition file 1.1Tb, restart file 2.1Tb, history file 250Gb	

Notes

Pure MPI parallelization in MPAS → Trial OpenMP hybrid mode d… porting workshop (OpenMP test code in place)

Probably not!

Expect breakdown of parallel efficiency at about 400000 threads → Exploit limiting factors in the model when the scaling breaks for extreme numbers of threads

NCAR Earth System Laboratory KIT Karlsruhe Institute of Technology