Exascale issues in climate model coupling

Robert Jacob

1st Workshop on Dynamical Cores for Climate Models
December 14th - 16th, 2011.

U.S. Global Change Research Program

- Coordinates and integrates U.S. federal research on changes in the global environment and their implications for society
- 13 U.S. departments and agencies involved. Only a few develop climate models.

NCAR

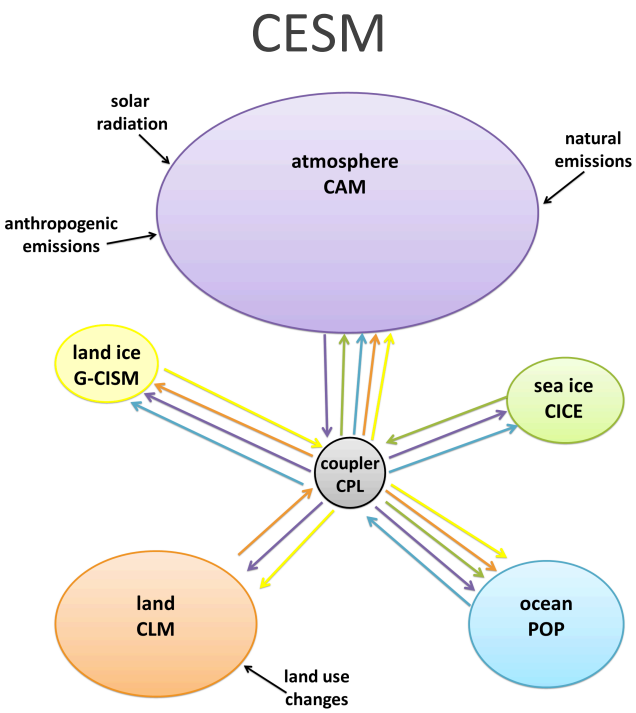
ANL, LANL, ORNL, LBNL, LLNL, PNNL, SNL

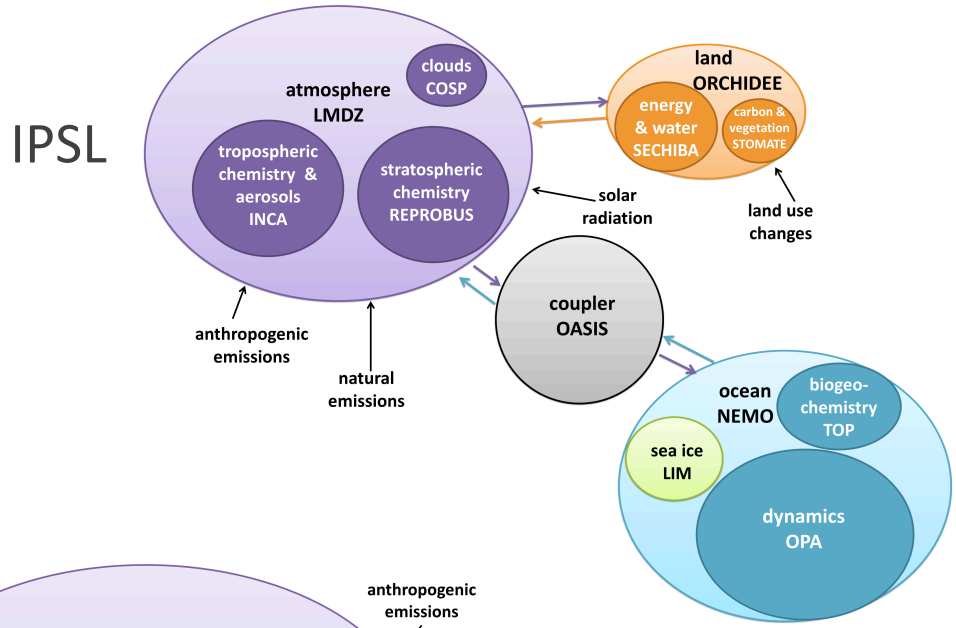
Develop Community Earth System Model (CESM; Formerly known as CCSM)

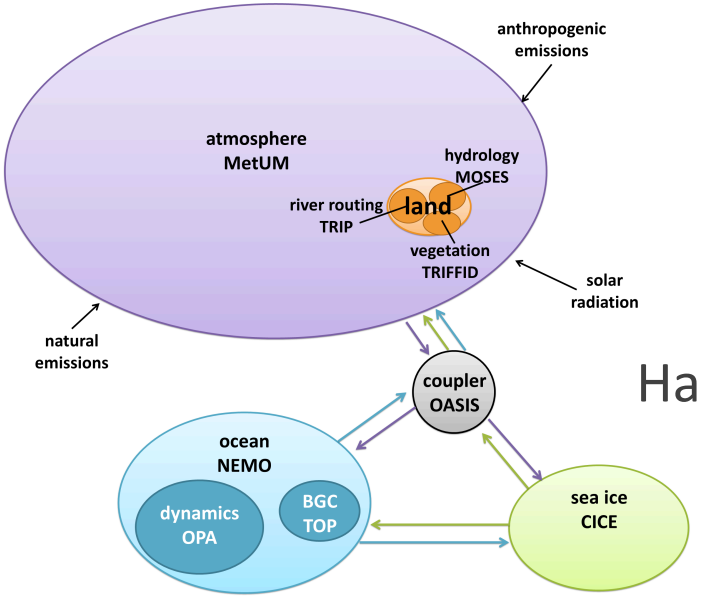
NOAA-GFDL

Develops CM3, ESM2M, HIRAM

Coupler is an important component of all climate models.

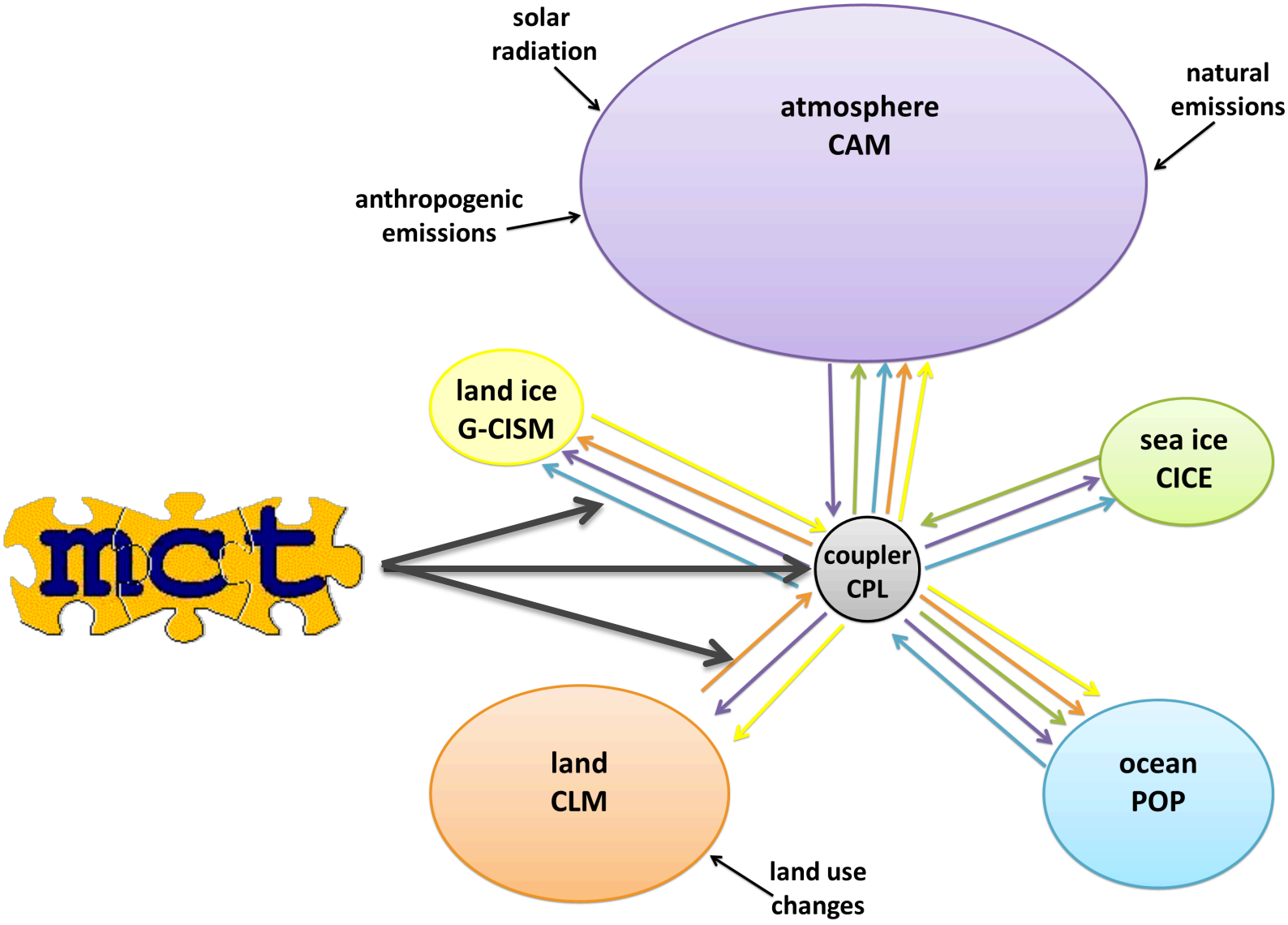






Figures from Kaitlin Alexander and Steve Easterbrook. Ovals proportional to code size.

Model Coupling Toolkit (MCT) in CESM: Handles all communication and interpolation



MCT Philosophy: Separating science concerns in building a coupled system from coupling "mechanics"

- Scientific *answers* of a coupled model determined partly by:
  - Number of components coupled
  - Number of fields exchanged
  - Interpolation method
  - Intercomponent flux calculation methods.
  - Time stepping of models and frequency of exchange

- MCT provides methods for
  - Component model registry
  - Parallel data movement
  - Domain decomposition description
  - Flexible, indexible field data storage

- MCT also provides science neutral methods for other coupling problems:
  - Averaging/Integration in time
  - Averaging/Integration in space
  - Merging

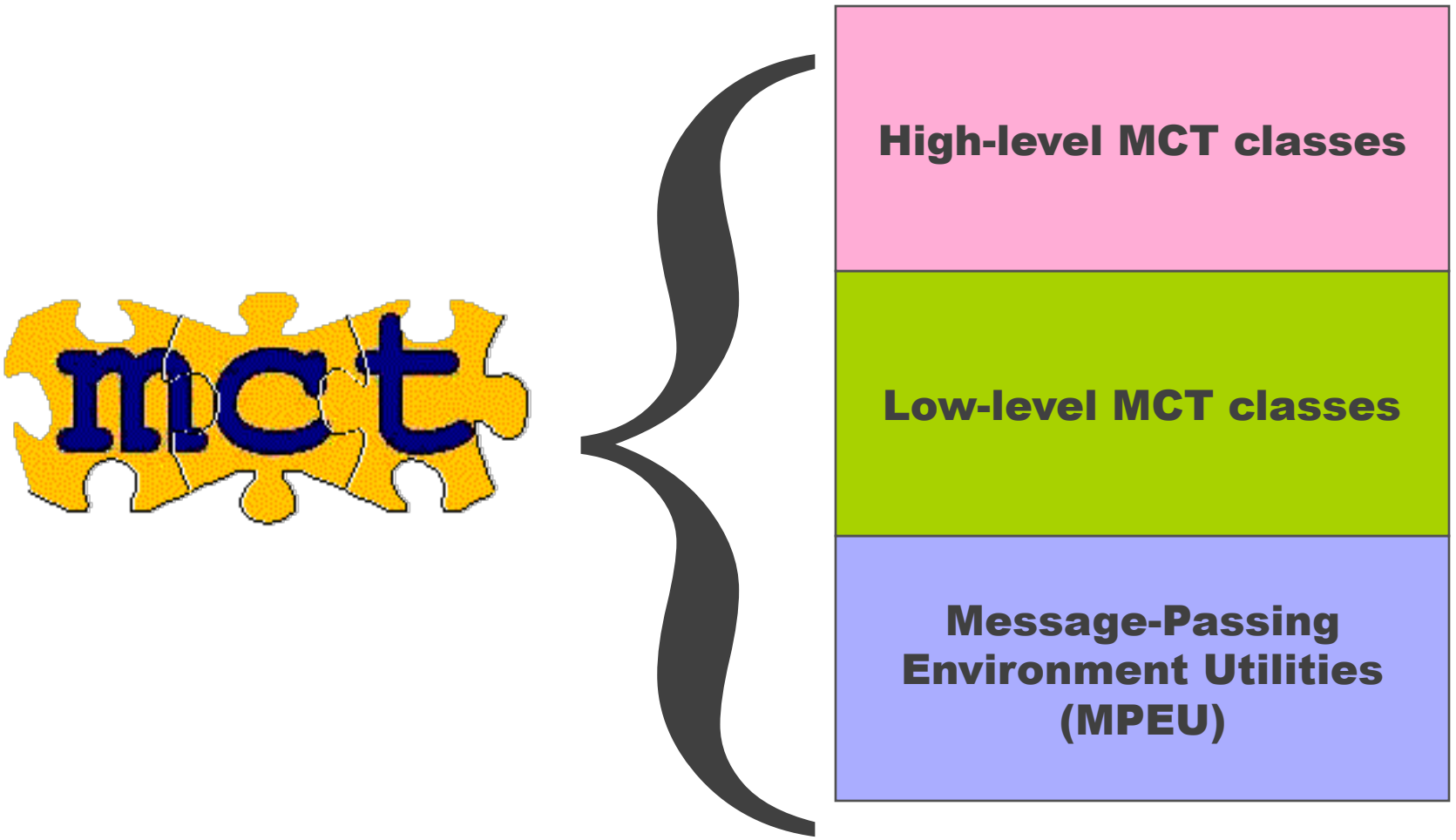
Tight and loose model coupling

- The various physics within a system like the atmosphere are "tightly coupled" - operate on entire state every time step.
  - Physics-dynamics coupling in atmosphere model is tight. Custom coupler often employed.

- Programming language serves as the framework in most cases.
  - "Tight integration between science and coding" -Steve Easterbrook

- *Between* models, translation (between datatypes, between grids, possibly between languages) is necessary.
  - *Excellent place for a general approach* (MCT).

MCT Architecture



MCT Low-level Classes

- Coupled model registry (describe how many models are coupled (*no limit*))
  - `MctWorld`
- Multi-field data storage (hold data being transferred (*any amount*))
  - `AttrVect`
- Domain decomposition (*any grid, any decomposition*)
  - `GlobalSegMap`
- Intercomponent parallel data transfer scheduler (*between two GSMaps*)
  - `Router`
- Intercomponent parallel data transfer (*For a Router and an Av*)
  - `Transfer`
- Intracomponent parallel data redistribution (*for an AV and a GSMap of the same grid*)
  - `Rearranger`

MCT High-Level Classes and Modules

- Interpolation (sparse) matrix object
  - `SparseMatrix`
- Sparse matrix – Attribute Vector multiply (for interpolation)
  - `MatAttrVectMult`
- Physical Grid Description
  - `GeneralGrid`
- Time averaging and accumulation support
  - `Accumulator`
- Masked/unmasked spatial integrals and averages
  - `SpatialIntegral`
- Combining sources from two or more models
  - `Merge`
- Communication methods for MCT datatypes
  - `AccumulatorComms`
  - `AttrVectComms`
  - `GeneralGridComms`
  - `GlobalSegMapComms`
  - `SparseMatrixComms`

Typical MCT Use:

ATM (M nodes)	CPL (N nodes)	OCN (P nodes)
Call MCT World	Call MCT World	Call MCT World
Define GlobalSegMap Define AttrVect Define Router	Define GlobalSegMaps Define AttrVects Define Routers Define Accumulators Read Matrix elements	Define GlobalSegMap Define AttrVect Define Router

Initialization

Read Atmosphere Data DO WORK		Read Ocean Data DO WORK
MCT_Send(AtrVect, Router) →	MCT_Recv(AAtrVect, ARouter) MCT_Recv(OAtrVectin, ORouter) ←	MCT_Send(AtrVect, Router)
	(Interpolation) MCT_AvMatVectMult(AAtrVect, SparseMatrix, OAtrVectout)	
	Compute Fluxes	
MCT_Recv(AtrVect, Router) ←	MCT_Send(AAtrVect, ARouter)	
	MCT_Send(OAtrVect, ORouter) →	MCT_Recv(AtrVect, Router)

MCT Users

- The NSF/DOE Community Earth System Model (CESM)
  - MCT is the default coupling method in CCSM4 and CESM1 (released July, 2010).
  - MCT datatypes always used in top level coupler driver.
  - MCT methods/datatypes are default for driver-component communication.
  - **All AR5 simulations by CCSM4/CESM1 and AR3 with CCSM3 use MCT.**
- U.S. University coupled systems:
  - COAMPS/ROMS - Hurricanes
  - ROMS/Swan - coastal oceanography
  - WRF/ROMS - Hurricanes
- Soon: Users of OASIS3?

Petascale to Exascale

IBM BlueGene series

- **BG/L (5.7 TF/rack, 210 MF/W) – 130nm ASIC (2004 GA)**
  - Scales >128 racks, 0.734 PF/s, dual-core system-on-chip,
  - 0.5/1 GB / Node

- **BG/P (13.9 TF/rack, 357 MF/W) – 90nm ASIC (2007 GA)**
  - Scales >256 racks, 3.5 PF/s, quad core SOC, DMA
  - 2/4 GB / Node
  - SMP support, OpenMP, MPI
  - **Intrepid at Argonne: 164K cores, .5PF**

- **BG/Q (209 TF/rack, 2000 MF/W) – 45nm ASIC (Early 2012 GA)**
  - Scales >256 racks, 53.6 PF/s, 16 core/64 thread SOC
  - 16 GB / Node
  - Speculative execution, sophisticated L1 prefetch, transactional memory, fast thread handoff, compute + IO systems

Mira - Argonne's BlueGene/Q

- 48 racks
- 1024 nodes per rack
- 1.6 Ghz 16-2ay core processor and 16 GB RAM per node
- 348 I/O nodes
- 240 GB/s 35PB Storage
- 768K cores
- 768 TB Ram
- 10PF peak

DOE National System Architecture Targets

System attributes	2010	"2015"		"2018"	
System peak	2 Peta	200 Petaflop/sec		1 Exaflop/sec	
Power	6 MW	15 MW		20 MW	
System memory	0.3 PB	5 PB		32-64 PB	
Node performance	125 GF	0.5 TF	7 TF	1 TF	10 TF
Node memory BW	25 GB/s	0.1 TB/sec	1 TB/sec	0.4 TB/sec	4 TB/sec
Node concurrency	12	O(100)	O(1,000)	O(1,000)	O(10,000)
System size (nodes)	18,700	50,000	5,000	1,000,000	100,000
Total Node Interconnect BW	1.5 GB/s	150 GB/sec	1 TB/sec	250 GB/sec	2 TB/sec
MTTI	day	O(1 day)		O(1 day)	

Challenges for Exascale

- Processor architecture is still unknown
- System power is the primary constraint
  - Scaling up today's petaflop computer; an exaflop would need 200MW. Target must be 20-40MW in 2020 for 1 exaflop
- Memory bandwidth and capacity are not keeping pace with the increase in flops; technology trends against a constant or increasing memory per core.
- Clock frequencies will decrease to conserve power. So processing units will increase. Billion-way concurrency expected.
- Cost of data movement, both in energy consumed and in performance, not expected to improve. Algorithms must minimize data movement, not flops.
- New Programming model necessary: heroic compilers will not be able to hide the level of concurrency from applications.
- I/O at all levels (chip to memory, memory to I/O node, I/O node to disk) will be harder to handle and slow
- Reliability and resiliency will be crucial. "silent errors" more likely.

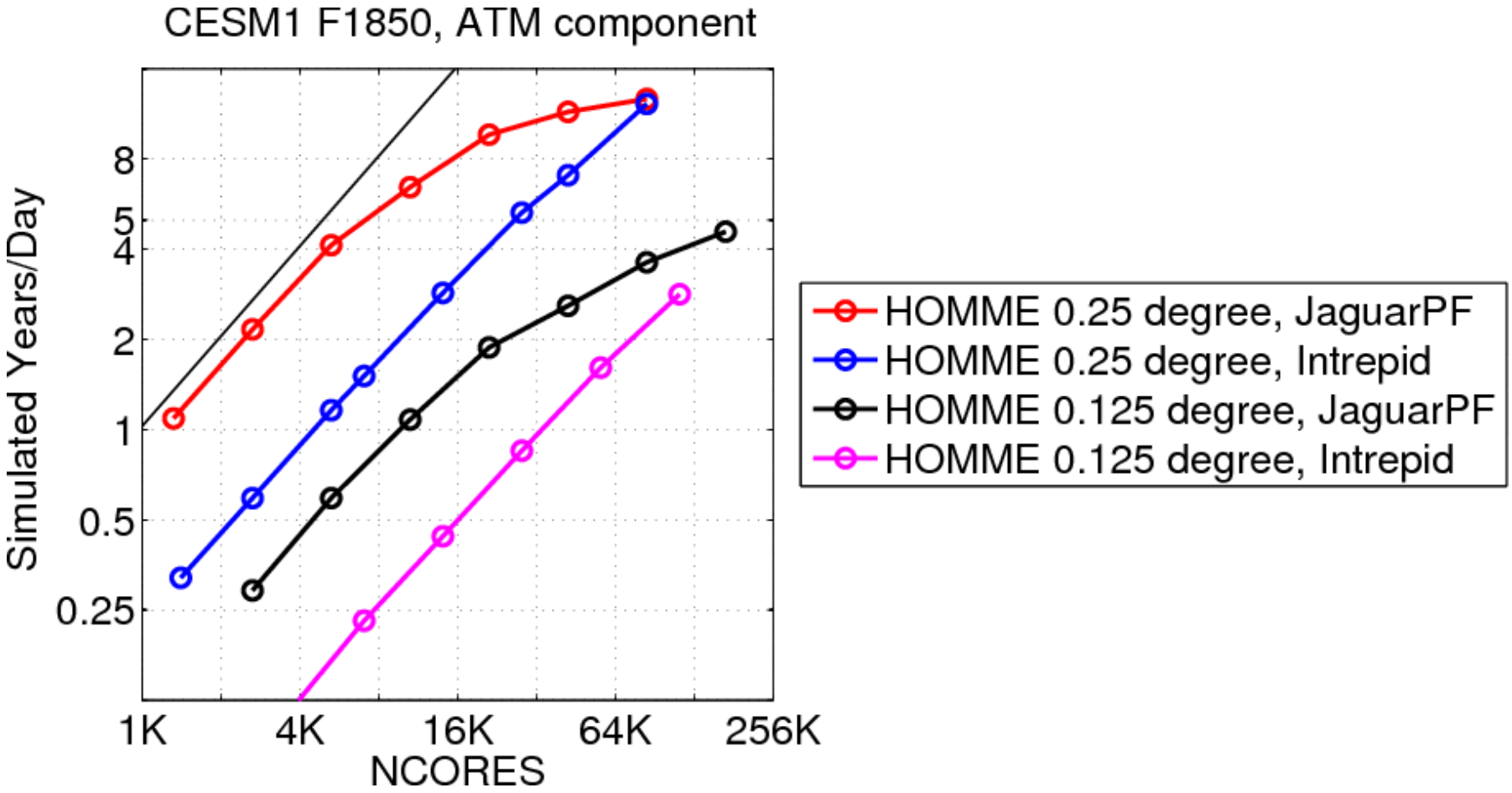
DOE Exascale Co-design Centers

"DOE labs, working with academic and industrial collaborators, will research challenges to creating exascale-capable applications in specific disciplines."

- Combustion Exascale Co-Design Center (PI Jacqueline Chen, Sandia Nat. Lab)
- CESAR, the Center for Exascale Simulation of Advanced Reactors (PI Robert Rosner, U. Chicago)
- Exascale Co-Design Center for Materials in Extreme Environments (ExMatEx) (PI Timothy Germann, Los Alamos Nat. Lab)

No climate co-design center yet….

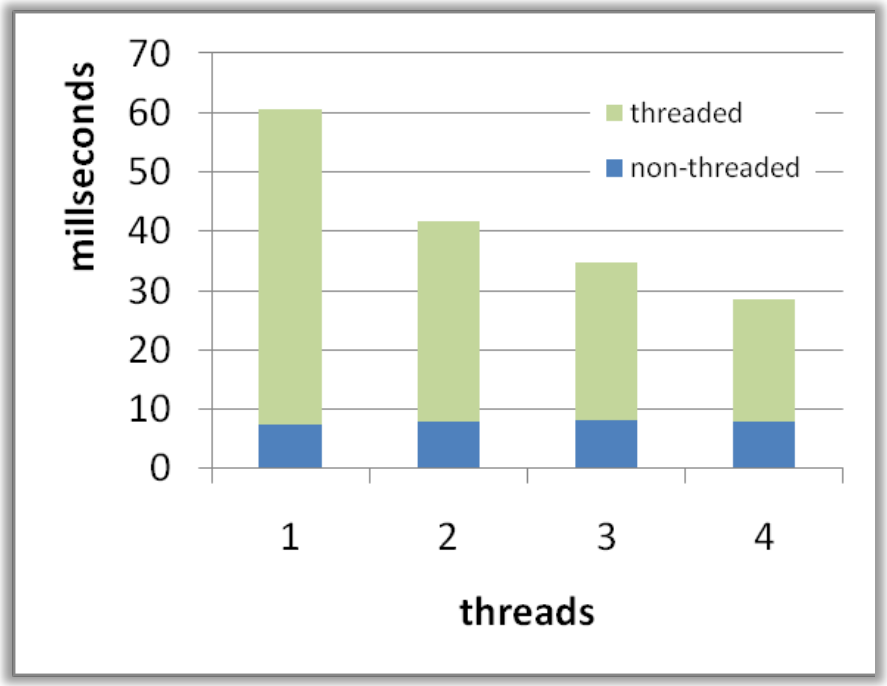
We can run climate model components on 100K cores. How to we get to 1 Billion?



From Mark Taylor, SNL

Atmosphere Dynamical Cores: exploiting parallelism in the vertical dimension

- Problems with vertical parallelism
  - Dynamics generally dependence-free except for implicit solver for W-winds
  - Physics has many vertical dependencies
    - Reductions
    - Recurrences
    - Searches
    - All-to-all dependencies
  - Hopeless?
- Cloud microphysics
  - Computationally expensive
  - Principal source of load imbalance
  - Dependencies
    - Calculating internal time step is a reduction
    - Computing fall speeds is recurrent
  - *The rest is dependency-free in vertical*
  - Initial attempt shows 2.3x on 4 threads



Threading vertical dimension of WSM5 cloud microphysics
(S. Y. Hong et al. Yongsei University)
Dual quad-core Intel 5570 2.93 GHz
Various KMP_AFFINITY granularity settings

Slide from J. Michalakes, NREL

Coupler vs. Exascale

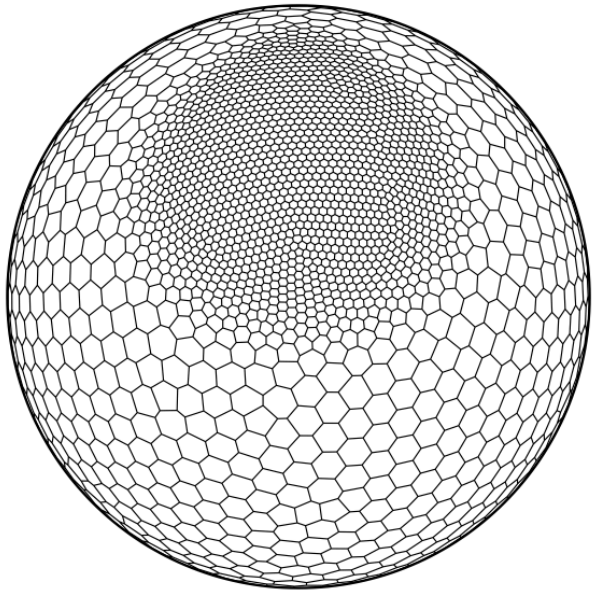
- Coupler is almost entirely 2D
  - Limited amount of parallelism
  - But not a huge number of flops compared to full model

- Coupler does lots of memory movement (which is expensive)
  - Moving data between model's native data type and coupler data type.
  - Moving data from one model's processors to another's.

Coupler vs. Exascale: solutions

- More parallelism through more components executing concurrently
  - Ensembles
  - Different models
- Reduce memory movement
  - One data type?
  - Co-located decompositions.

At exascale, coupler will have to deal with new grids.

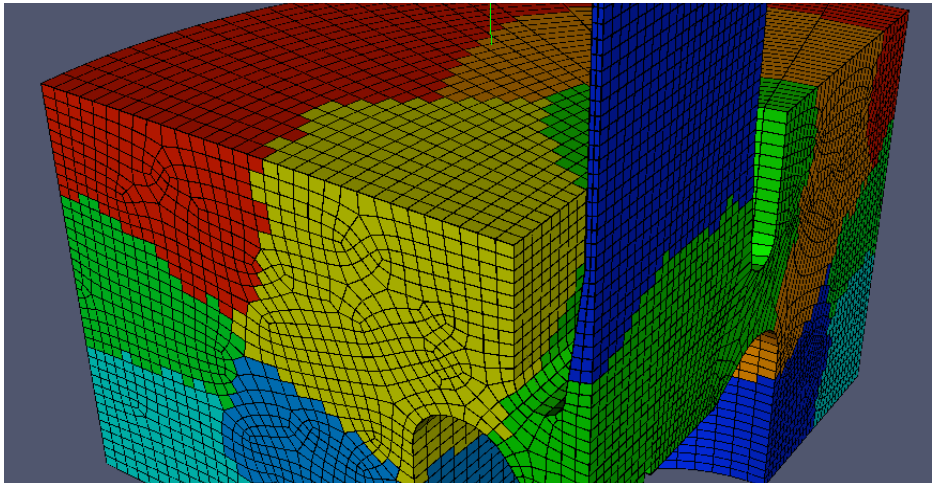


- Very high resolution; unstructured or semi-structured grids.
- Possibly adaptive grids will require redefining grid decomposition; recomputing mapping weights.

Solution: Re-Implement MCT data model with MOAB

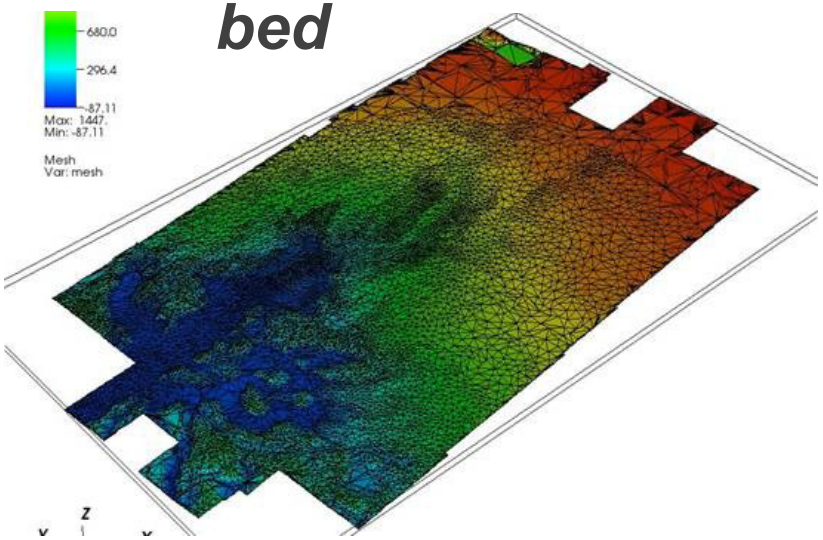
- MOAB = Mesh Oriented dAtaBase
  - A database for mesh (structured and unstructured) and field data associated with mesh
  - *Tuned for memory efficiency first*, speed a close second
  - Serial, parallel look very similar, parallel data constructs imbedded in MOAB interface
  - http://trac.mcs.anl.gov/projects/ITAPS/wiki/MOAB
  - Developed under DOE SciDAC program
  - Includes parallel I/O and visualization capabilities.

MOAB is already used in other projects, notably DOE-funded cryosphere modeling and nuclear reactor simulation.
Like MCT, it is “battle tested”.



Klystron Mesh

Ice Sheet bed



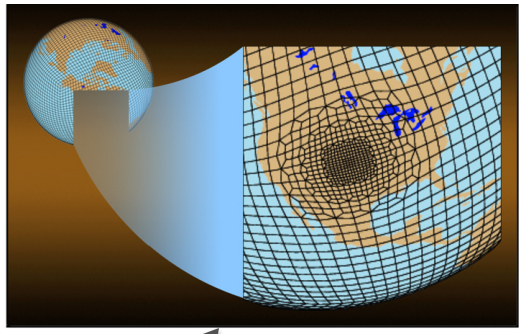
Climate Science for a Sustainable Energy Future

Cross-cutting Themes and Labs to advance CESM to address high priority DOE climate research

3 Science Themes:
- Numerics
- Testbeds
- Uncertainty Quantification

3 Components:
- Atmosphere
- Land
- Ocean and Sea-Ice

3 Research Directions:
- Hydrologic simulation improvement
- Variable-resolution numerical methods
- Carbon cycle uncertainty reduction



9 Labs:
ANL
BNL
LANL
LBNL
LLNL
ORNL
PNNL
SNL
NCAR

CSSEF 2016 Goal

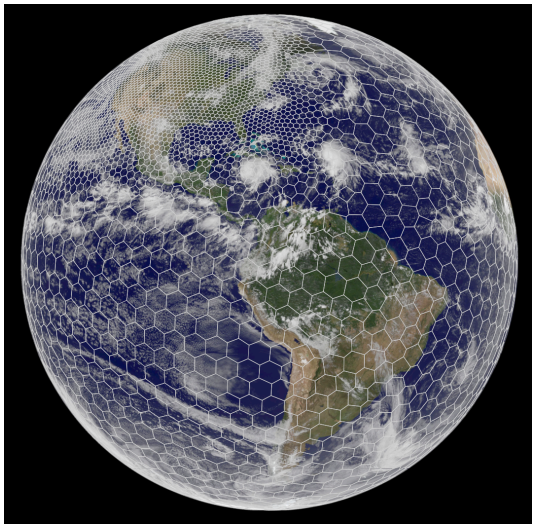
- Develop predictive capability to address key climate change science drivers in the 2016-2020 time period, i.e. to answer questions posed in the period after the publication of the IPCC 5th Assessment Report
- Directed toward the development of Community Earth System Model Version 3(CESM3) – two generations from current model
- Time frame coincident with the advent of Exascale computing and deployment of new climate observational data streams

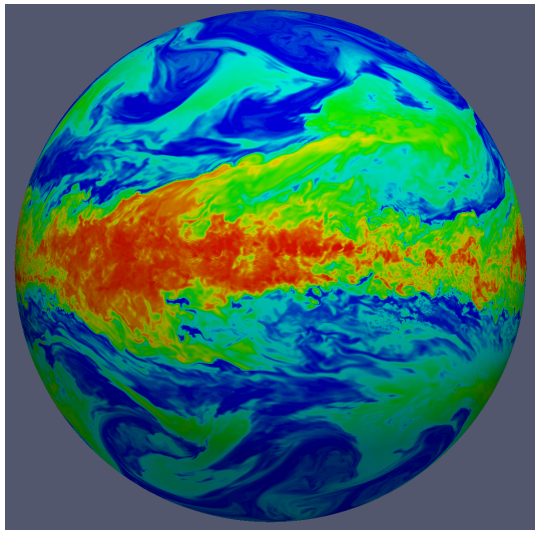
Tightly Integrated and Coordinated Approach

- Integrated within project and CESM and complementary to CESD projects.
- Accelerate incorporation of new knowledge, including process data and observations, into climate models.
- Develop new methods for rapid evaluation of improved models.
- Develop novel approaches to exploit computing at the level of tens to hundreds of petaflops in climate models.

Status: waiting on FY12 DOE budget

Nonhydrostatic Global Modeling with MPAS





MPAS
Model for Prediction Across Scales

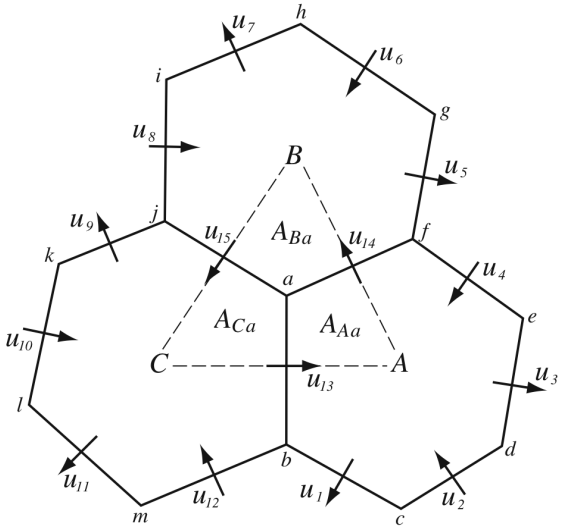
Based on unstructured centroidal Voronoi (hexagonal) meshes using C-grid staggering and selective grid refinement.

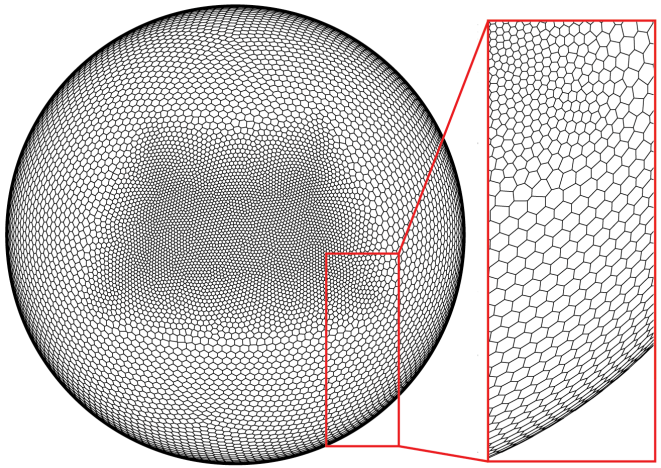
Jointly developed, primarily by NCAR/NSF and LANL/ DOE, for weather, regional climate, and climate applications

MPAS infrastructure - NCAR, LANL, others.
MPAS - Atmosphere (NCAR)
MPAS - Ocean (LANL)
MPAS - Ice, etc

Bill Skamarock, Joe Klemp, Michael Duda,
Sang-Hun Park and Laura Fowler NCAR
Todd Ringler Los Alamos National Lab (LANL)
John Thuburn Exeter University
Max Gunzburger Florida State University
Lili Ju University of South Carolina

Global Atmospheric Modeling Using Voronoi Meshes: The MPAS Model





Applications

- *Primarily NWP and Regional Climate*

Equations

- *Fully compressible nonhydrostatic equations, vector invariant form.*

Solver Technology

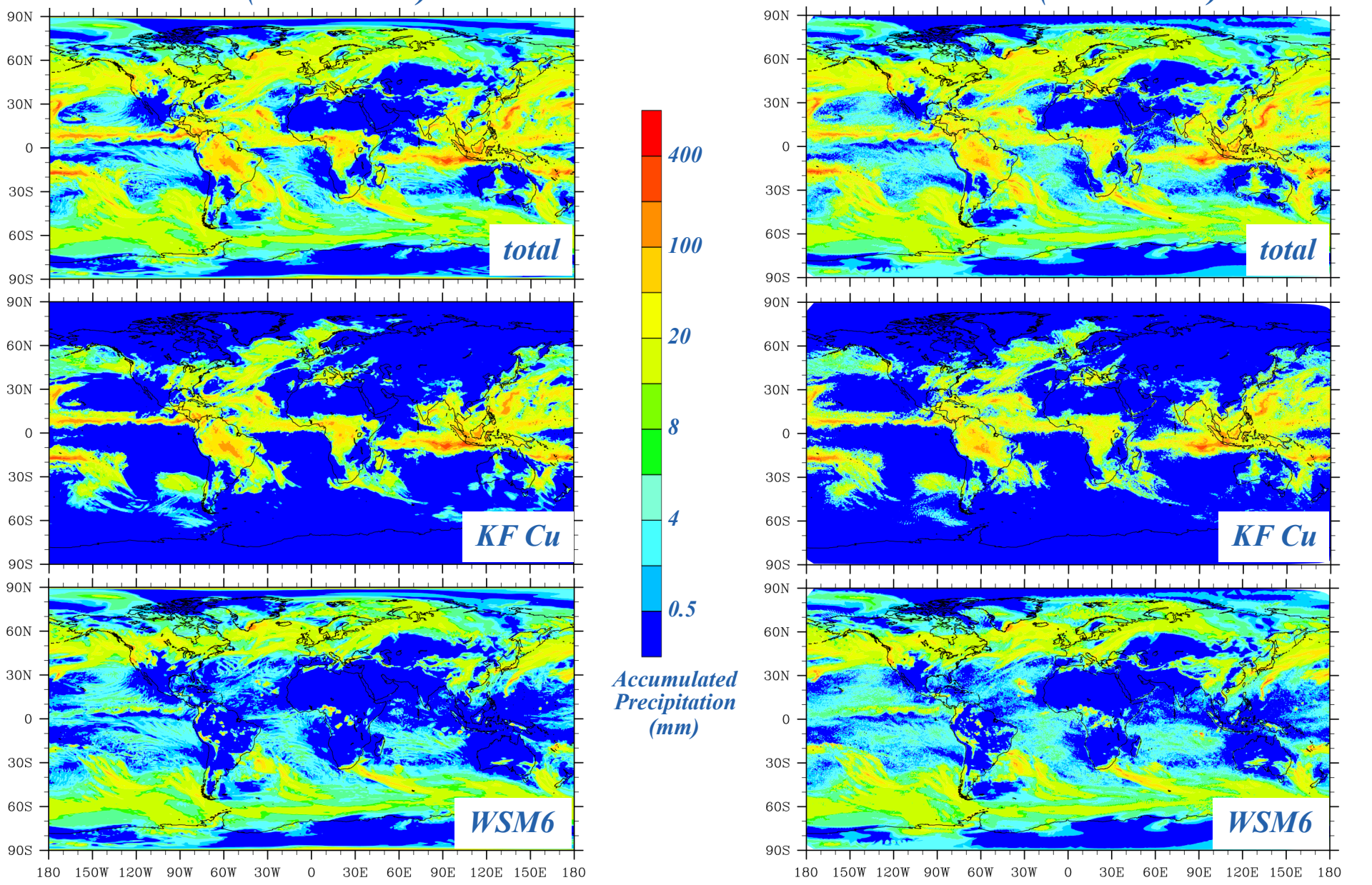
- *Most of the techniques for integrating the nonhydrostatic equations come from WRF.*

C-grid centroidal Voronoi mesh

- *Emphasis on accuracy for divergent modes.*
- *We have developed accurate and efficient transport schemes (consistent, conservative, PD and monotonic).*

5 Day Accumulated Precipitation Forecasts

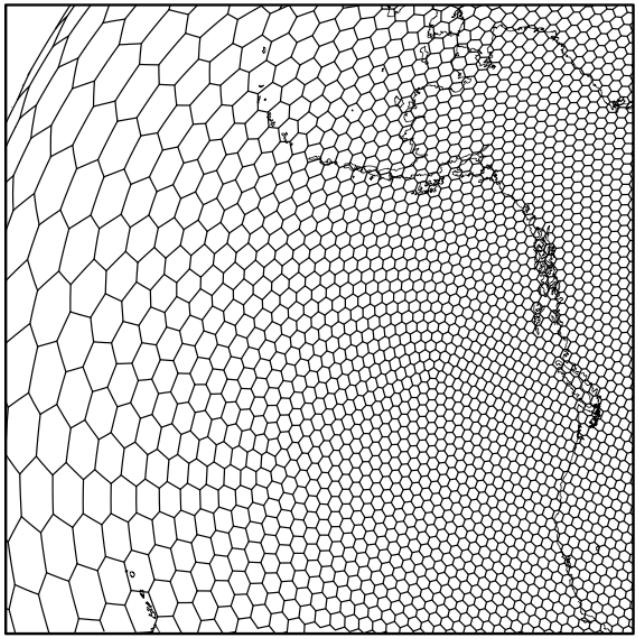
valid 0Z 28 October 2010

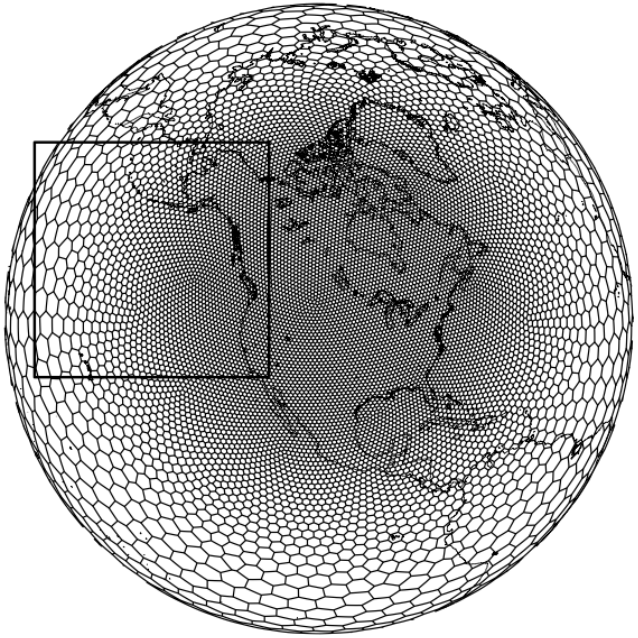


Nonhydrostatic Global Modeling with MPAS

Physics:

WSM6 cloud microphysics
Kain_Fritsch convection
Monin-Obukhov surface layer
YSU pbl
Noah land-surface
RRTMG lw and sw radiation.





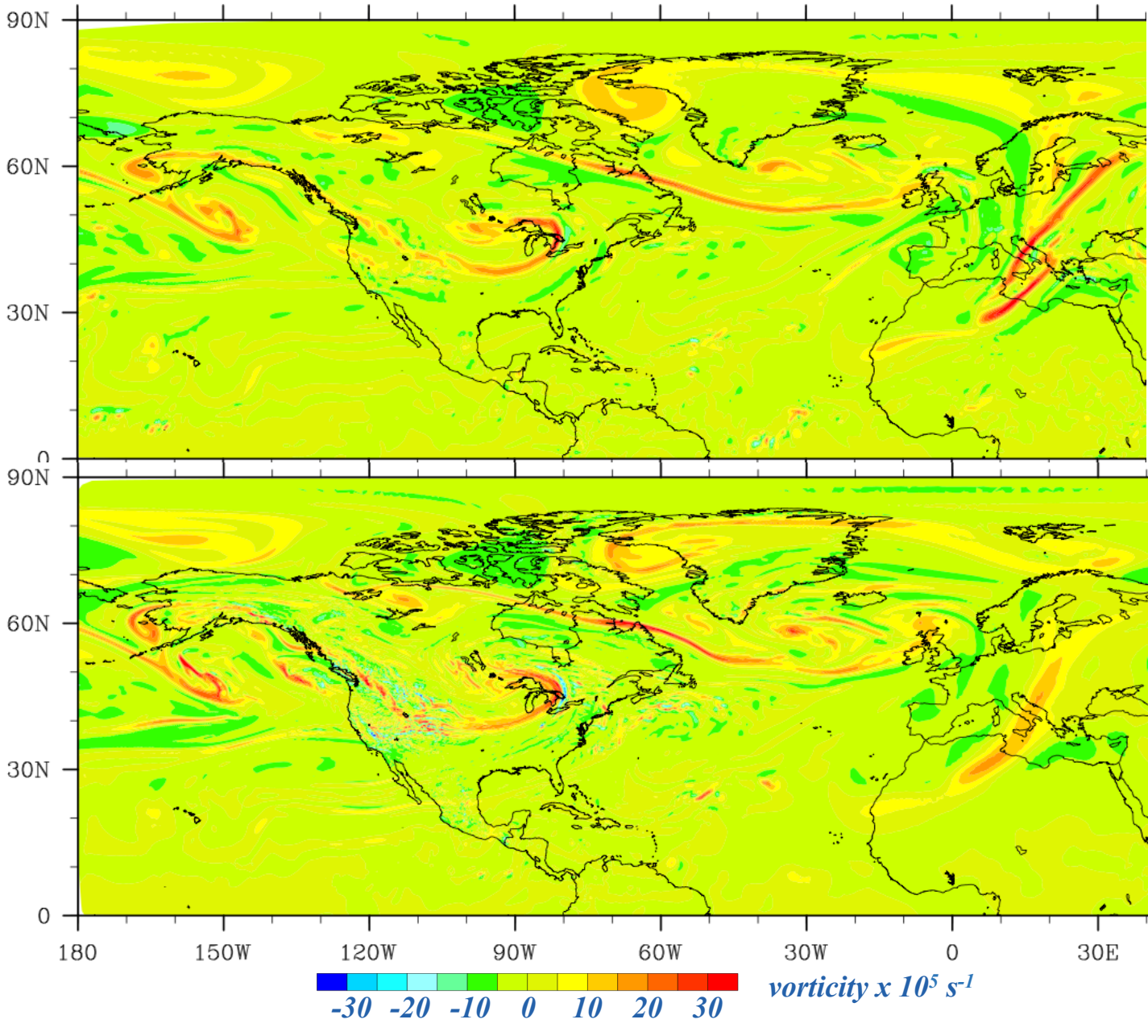
MPAS 4-day forecast, valid 27 October 2010

Day 4
500 hPa
relative vorticity

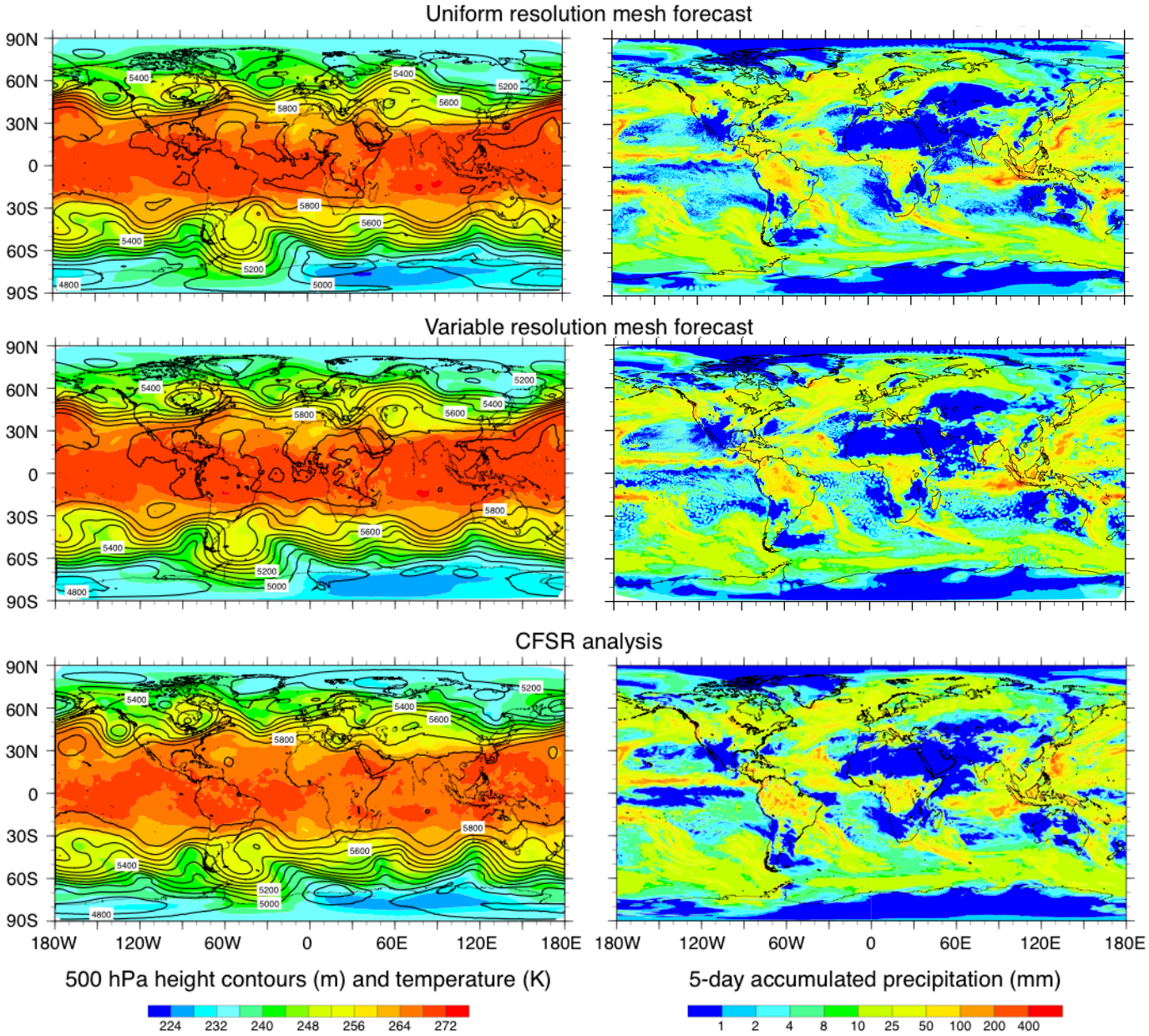
Uniform resolution
~ 60 km cell spacing

500 hPa
relative vorticity

8x variable res.
~ 21-162 km
cell spacing



5-Day Forecasts and Analysis Valid 28 October 2010



MPAS Status

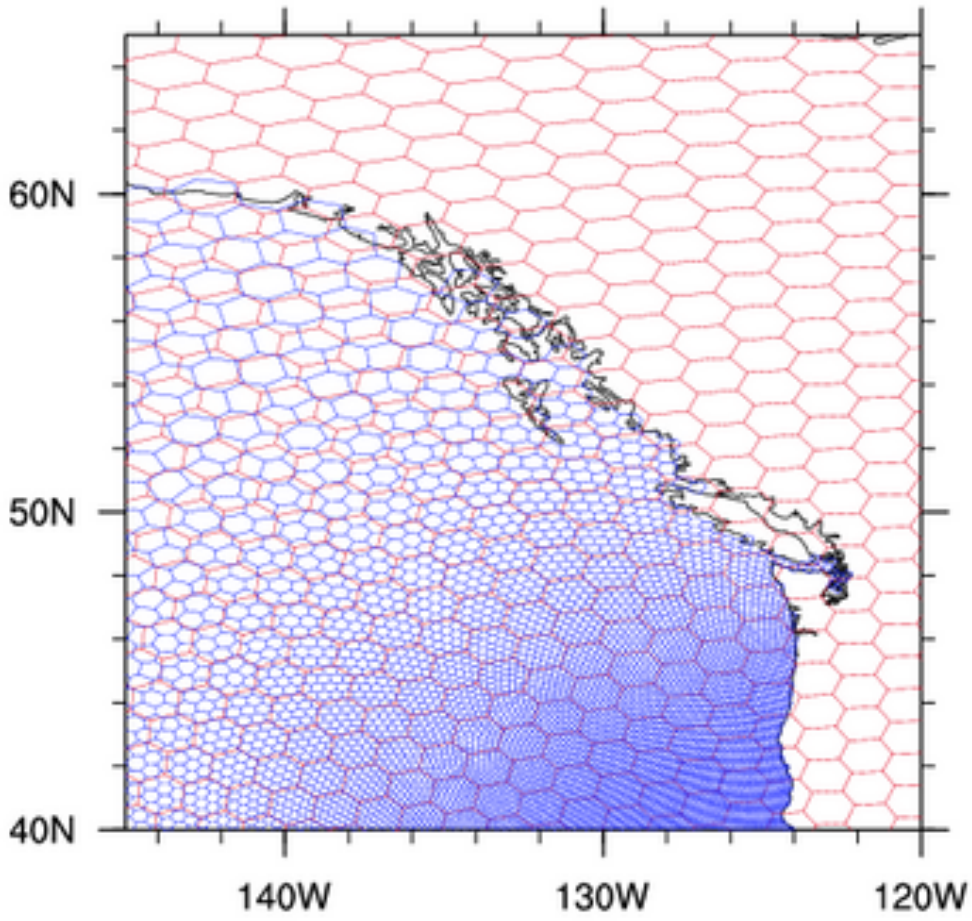
- Hydrostatic MPAS dynamical core ported to NCAR Community Atmosphere Model with CAM4 physics. Simple aqua-planet runs completed.

- Non-hydrostatic MPAS currently being ported.

- Big question: scale sensitive atmospheric physics

- "Coupled, High Resolution, Multi-Scale Climate Simulations on Unstructured Meshes" - DOE SciDAC proposal submitted last week.
  - PI Ian Foster, Argonne
    - Argonne: Mihai Anitescu, Ramesh Balakrishnan, Emil M. Constantinescu, Yan Feng, Paul Hovland, Robert Jacob, Rao Kotamarthi, Lois Curfman McInnes, Sheri Mickelson, Rob Ross, Tim Tautges
    - NCAR: Cindy Bruyere, Michael Duda, Greg Holland, Bill Skamarock
    - LBNL: David Bailey
    - NREL: John Michalakes
    - UT-Austin: Robert Moser

Coupling on unstructured, refined meshes?



Fini