Present and Future Computing Requirements for "Development of Frameworks for Robust Regional Climate Modeling"

L. Ruby Leung

Pacific Northwest National Laboratory

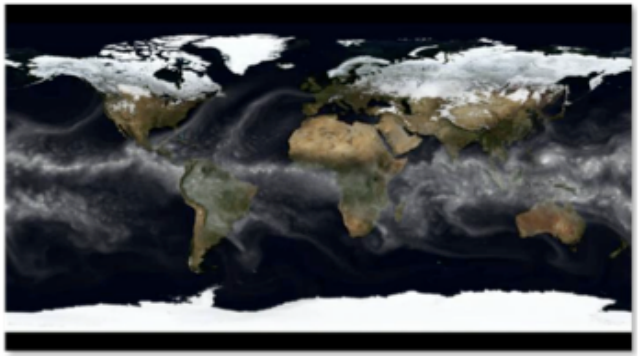
NERSC BER Requirements for 2017
September 11-12, 2012
Rockville, MD

1. Project Description

Ruby Leung, PNNL; Todd Ringler, LANL; Bill Collins, LBNL, and Moet Ashfaq, ORNL; Mark Taylor, SNL

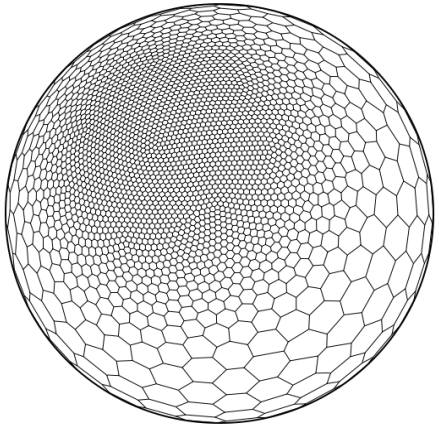
- Objectives:
  - Applies a hierarchical framework to evaluate three dynamical approaches to modeling regional climate through global high resolution models, global variable resolution models, and nested regional climate models, all sharing a common physics package

CAM Spectral Eulerian and HOMME



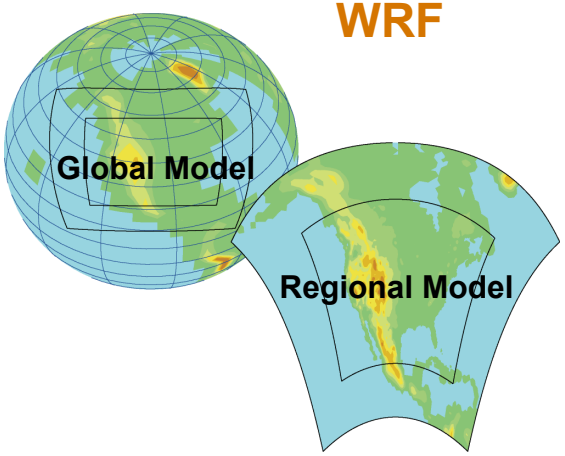
Global high resolution model

CAM MPAS



Global variable resolution model

WRF



Nested regional climate model

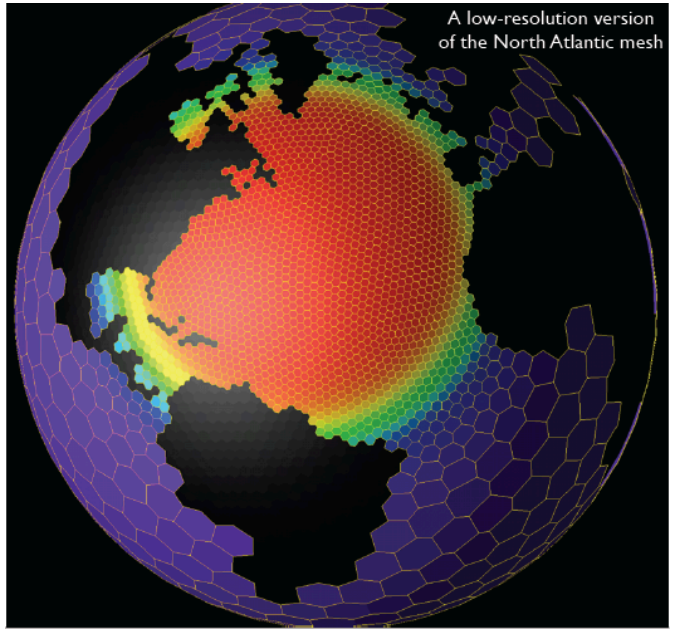
1. Project Description

- Our present focus:
  - Analysis of aquaplanet simulations and AMIP style simulations to assess the impacts of dynamical framework, dynamical core, and model resolution
- By 2017 we expect to:
  - Move towards higher resolution, including cloud resolving simulations
  - More focus on coupled simulations
  - Evaluate interactions among dynamical framework, dynamical core, and model resolution in the context of scale-aware physics parameterizations
  - Applications of models to understand water cycle variability and extremes

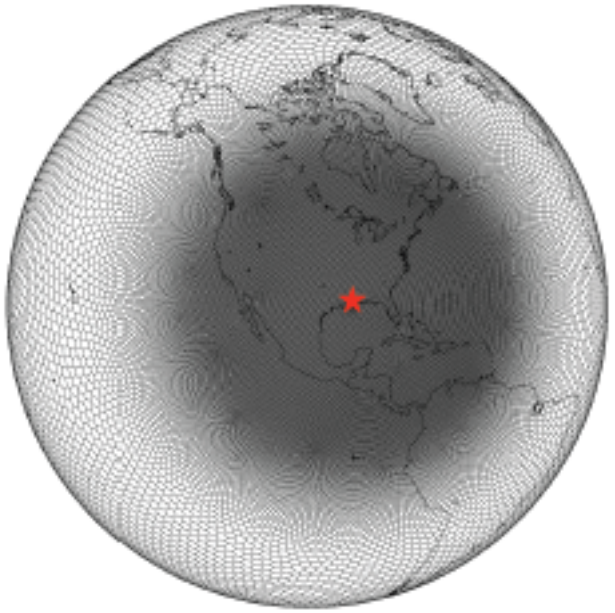
2. Computational Strategies

- We approach this problem computationally at a high level by:
  - Utilizing existing software/hardware to perform idealized and real world simulations with different models at low resolution (1^o), high resolution (0.25^o), and variable resolution ($1^o \rightarrow 0.25^o$)
- The codes we use include offline and coupled atmosphere/ocean models:
  - CAM Spectral Eulerian, POP
  - CAM HOMME, POP
  - CAM MPAS-A, MPAS-O ✔
  - WRF, ROMS ✔

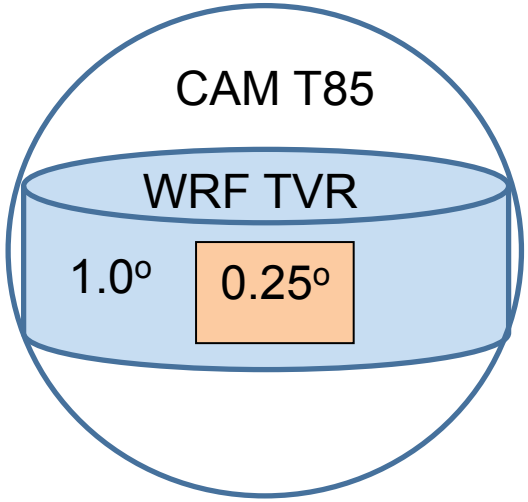
2. Computational Strategies

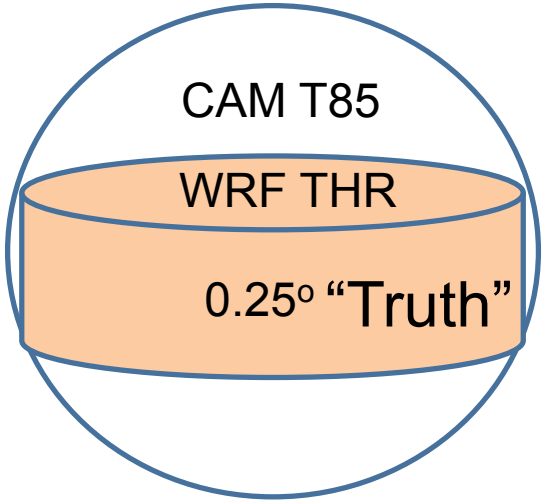


MPAS-O: 10km to 50km



MPAS-A: 30km to 120km





2. Computational Strategies

- MPAS-A and MPAS-O are characterized by these algorithms:
  - The codes are fully explicit - no global reductions or large linear system solvers are used
  - Scaling to large processor counts is dependent upon our ability to transfer “halo” data between processors in a local communication pattern (i.e. 1 proc sending messages to less than 8 procs) on each time step
  - The data model is structured in the vertical, but unstructured in the horizontal – directly address data in the vertical, but require indirect addressing to find neighboring data in the horizontal
  - Data laid out with the vertical index first in order to exploit this structured index - largely mitigates the inefficiencies incurred due to using unstructured addressing in the horizontal

2. Computational Strategies

- WRF is characterized by these algorithms:
  - The ARWRF solver uses a time split finite difference scheme
  - The code has two levels of domain decomposition (patch and tile) designed to run over distributed as well as shared memory
  - Scaling to larger processor counts is limited by IO and communication (little is gained beyond 100 grid points per tile)

2. Computational Strategies

- Our biggest computational challenges are:
  - Completing long integrations require frequent submission of sequential jobs, but long wait time in the queue limits productivity
  - Not getting the amount of resources requested limits what can be accomplished
  - WRF not able to utilize a large number of processors limits efficiency
- Our parallel scaling is limited by I/O for WRF

2. Computational Strategies

- We expect our computational approach and/or codes to change by 2017 in this way:
  - MPAS uses common approaches to high performance computing that exploit large, massively parallel computing systems.
  - Researchers at LANL are exploring mixed parallelism in the form of MPI-OpenMP for MPAS-O and will test parts of the code on accelerators (GPUs) over the next year.

3. Current HPC Usage

- Machines currently used:
  - MPAS: NERSC Hopper, LANL Mustang
  - WRF: NERSC Hopper, NCCS Jaguar
- Hours used in 2012:
  - ~ 3M hours on Hopper, 1.5M hours on Jaguar
- Typical parallel concurrency and run time, number of runs per year
  - MPAS currently uses approximately 4000 cores per run.
  - The maximum number of cores that have been used is 6000.
  - Typically MPAS-O/MPAS-A are not run with multiple jobs concurrently.
  - WRF typically uses 1296 cores per run
  - A global quarter degree tropical channel wrf simulation typically costs about 200 processor hours per a model run day.

3. Current HPC Usage

- Typical parallel concurrency and run time, number of runs per year
  - MPAS: A typical checkpoint is 10GB. Typically about 20% of wall clock time used for I/O. Of this, checkpointing accounts for about 1/4 of I/O time.
  - In WRF about 50% of the time is spent on IO (writing six hourly data).
- Data read/written per run
  - MPAS: Typically write about 100GB per job submission. The NCAR PIO (Parallel I/O) tool is used - about 10X faster than serial I/O for MPAS applications.
  - WRF: The output from a 5 year quarter degree tropical channel WRF run is about 40TB.

3. Current HPC Usage

- Memory used per (node | core | globally)
  - MPAS: The model is designed to disallow any global arrays – this somewhat mitigates thin nodes by spreading the problem over more nodes, even if this does not improve time-to-solution.
- Necessary software, services or infrastructure
  - MPAS: Typically use a large suite of compilers (pgf, ifort, gfortran, xlf, etc) to test robustness of code. Also use various flavors of MPI (mpich and openmpi) and the NCAR PIO tool for parallel output.
  - WRF: Typically use pgf90 and MPI.
- Data resources used (HPSS, NERSC Global File System, etc.) and amount of data stored
  - HPSS: MPAS (50 TB); WRF (100 TB)

4. HPC Requirements for 2017

- Compute hours needed (in units of Hopper hours)
  - 100 M
- Changes to parallel concurrency, run time, number of runs per year
  - MPAS: Typical jobs will use ~25K processors. Small numbers (~5) of jobs might be run concurrently.
  - WRF: Typically jobs will use 1200 – 2400 processors for WRF. Up to 8 such simulations may run concurrently.
- Changes to data read/written
  - MPAS: Checkpoint file size will be approximately 100GB. Expect to use community-supported parallel I/O solutions, such as the NCAR PIO. Typical single job runs will generator 250GB of data. Typical simulations will require 20 to 100 job submissions.
  - WRF: With 12 hour runs, about 30 resubmissions are required to complete a run.

4. HPC Requirements for 2017

- Changes to memory needed per (core | node | globally): MPAS
  - Data-intensive problems that carry on order 100 tracer constituents (for biogeochemistry) with optimal scaling would require ~10 GB per processing unit. For machines with, say, 24 procs per node, this would require approximately 256 GB of memory per node.
- Changes to necessary software, services or infrastructure
  - MPAS: Would benefit from using parallel I/O tools that are used by the broader community.

5. Strategies for New Architectures

- Our strategy for running on new many-core architectures (GPUs or MIC): MPAS
  - Currently not using GPU, but plan to port the code to Titan during this calendar year.
  - Expect to be able to utilize directive-based accelerators in 2013 and beyond.
  - A significant part of current SciDAC project (Multiscale Earth Modeling, PI-Collins) at LANL is directed toward computational efficiency, including the use of accelerators.
- Researchers at LANL are exploring mixed parallelism in the form of MPI-OpenMP and will test parts of the code on accelerators (GPUs) over the next year

5. Summary

- What new science results might be afforded by improvements in NERSC computing hardware, software and services?
  - Improved ability to perform high resolution climate simulations to better predict water cycle changes in the future
  - Established more robust frameworks for high resolution modeling
- Recommendations on NERSC architecture, system configuration and the associated service requirements needed for your science
  - PIO
- NERSC generally refreshes systems to provide on average a 2X performance increase every year. What significant scientific progress could you achieve over the next 5 years with access to 32X your current NERSC allocation?
  - Perform cloud resolving simulations over large regions for evaluating cloud parameterizations and sensitivity to model resolution
- What "expanded HPC resources" are important for your project?
  - Increased memory per node, more nodes, larger storage capacity, shorter wait time in queues
- General discussion