Performance Engineering for the Next Generation Community Climate System Model

a Science Application Partnership aligned with Science Application Project *A Scalable and Extensible Earth System Model for Climate Change Science (SEESM)*

Raymond Loy (ANL) Arthur A. Mirin (LLNL) Patrick H. Worley (ORNL)

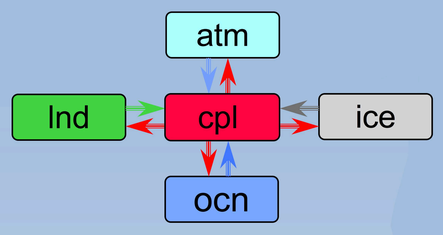
DOE Office of Advanced Scientific Computing Research, Computer Science Principal Investigator Meeting, March 31, 2008, Denver, CO

Background

- The goal of SEESM is to transform the Community Climate System Model (CCSM) into an *earth system model*, fully simulating the coupling between the physical, chemical, and biogeochemical processes in the climate system.
- The goal of this project is to enable this *earth system model* to utilize tens of thousands of processors efficiently while simulating more comprehensive physical processes at much higher resolution than before; our focus is
  - improving processor scalability of components of the climate model
  - optimizing performance at high resolution
  - improving scalability of the full coupled climate model through improved load balance
  - porting to and optimizing on next-generation leadership-class computing systems

CCSM

- Fully-coupled, global climate model that provides state-of-the-art computer simulations of the Earth's past, present, and future climate states
- Developed at the National Center for Atmospheric Research (NCAR), with contributions from external research groups funded by the National Science Foundation, Department of Energy (DOE) and National Aeronautics and Space Administration (NASA)
- Comprised of a coupler and four component models:
  - Atmosphere model, currently the Community Atmosphere Model (CAM3)
  - Ocean model, currently the Parallel Ocean Program (modified version of POP1.4 or POP2.1)
  - Land Model, currently the Community Land Model (CLM3)
  - Sea Ice Model, currently either the Community Sea Ice Model (CSIM5) or the Community Ice CodE (CICE4)
  - CCSM Coupler, currently version 6



(cartoon courtesy of CSEG/NCAR)

CCSM Performance Characteristics

- Each component has its own performance characteristics:
  - *atm*: latitude/longitude/vertical tensor product grid; choice of algorithms for solving evolutionary equations for atmospheric flow (dynamics): finite-volume (FV), spectral Eulerian (EUL), and spectral semi-Lagrangian (SLD).
    - 2D tensor product horizontal and 2D latitude/vertical domain decomposition for finite-volume dynamics; 1D horizontal domain decomposition for spectral dynamics
    - 2D arbitrary horizontal domain decomposition for column physics (subgrid phenomena: e.g., precipitation, radiation)
  - *ocn*: 3D grid using displaced pole or tripole, locally orthogonal horizontal grid; 2D horizontal domain decomposition.
  - *lnd*: 2D horizontal grid (same as atm currently); 2D arbitrary horizontal domain decomposition.
  - *ice*: 2D horizontal grid (same as ocn currently); 2D horizontal decomposition.
- At current resolutions, either the atmosphere or ocean component model, depending on the science being investigated, is the most expensive to run.

Community-Wide Effort

Many people, at many institutions, are currently working to improve the portability, performance, and scalability of the CCSM. Major efforts include:

- Model infrastructure, e.g.
  - multiple modes of operation (concurrent, sequential, and hybrid)
  - (model-) optimized parallel I/O layer
- Component and coupled model porting, performance optimization, and performance scaling analysis on "prototype Petascale systems":
  - IBM BG/L and BG/P
  - Cray XT3 and XT4
- Algorithms, e.g.
  - improved parallel algorithms (for existing numerical algorithms)
  - more scalable numerical algorithms (e.g., new computational grids)

We are involved in all of these efforts, working closely with SEESM-funded researchers T. Craig (NCAR), R. Jacob (ANL), and NSF-funded researchers J. Dennis (NCAR) and M. Vertenstein (NCAR).

New CCSM Architecture

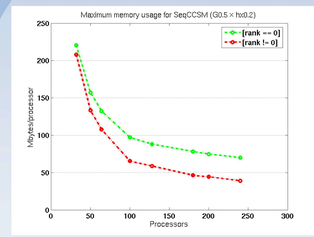
- We are contributing to a new version of the CCSM in which
  - physical components communicate directly with one another using the Model Coupling Toolkit (MCT) instead of going through an outer coupler layer;
  - components can be executed on the same processors (sequentially), on disjoint processor subsets (concurrently), or in a mixed sequential/concurrent mode.
- This new CCSM version is nearing completion and is being used for
  - ultra-high resolution simulations on the Atlas computing system at LLNL
  - porting and optimization on Blue Gene platforms

Parallel I/O

- We are major contributors to the implementation of PIO, a library that allows users to transpose data between a model decomposition and a more I/O-friendly decomposition on a subset of I/O processes. The I/O processes communicate with disk in parallel, using either PNetCDF or MPI-IO for file I/O.
- PIO has been implemented successfully in POP, and implementations in the other CCSM components are nearing completion.
- This capability is important for efficient use of parallel file systems, minimization of memory (e.g., reduction of global storage), and reduction of interprocess communication overhead.

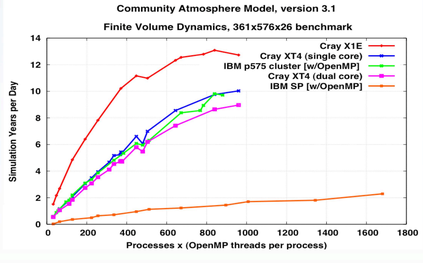
Port to IBM BG/L and BG/P

- We completed the first port of CCSM to a BG/L system, successfully executing at a modest resolution on up to 1024 processors.
- Memory requirements were (and are) a significant issue in porting CCSM to the BG/L and BG/P systems, and to any system when running with large processor counts and/or high resolution grids. Recent work has decreased the memory requirements in the coupler significantly.



Component Optimization: Atmosphere

CAM is the least scalable component. In the last public release no more than 960 MPI processes could be used for a large 0.5 degree horizontal grid; for the production 2 degree grid, the limit was 256 MPI processes.



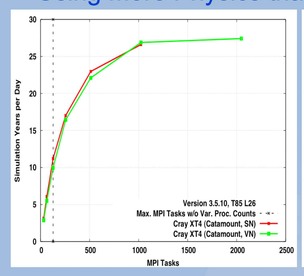
(Original) Limitations to CAM Scalability

- Number of subdomains limited by grid resolution
  - FV dynamics: minimum of 3 grid points in each coordinate direction; with only 26 vertical levels in current production configuration, this severely limited the size of the latitude/vertical decomposition
  - EUL and SLD: decompositions one-dimensional, and number of MPI processes constrained by number of latitudes (CAM allows idle processes in the spectral space)
- All domain decompositions required same number of MPI processes
  - number of MPI processes limited by smallest domain decomposition
- Communication cost of transposes, load imbalance, I/O, global diagnostics, … also affect parallel scalability

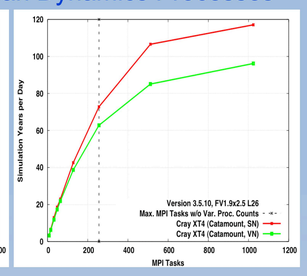
Recent extensions to CAM scalability

- Allowing more physics than dynamics processes
- Allowing inactive processes (to be used as auxiliary processes when needed)
- Allowing larger vertical decomposition (FV)
- Enabling larger longitude/latitude than latitude/vertical decomposition (FV)
- Overlapping tracer advection with main dynamics (FV)
- Using additional processes for tracer advection by decomposing over tracer index (FV)

Using More Physics than Dynamics Processes

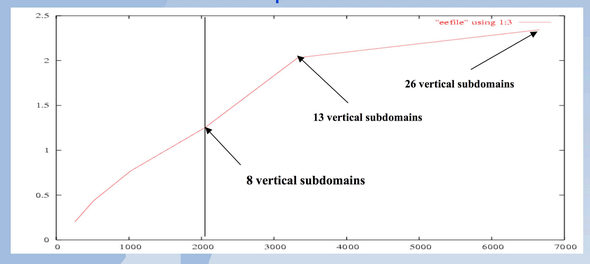


Spectral Eulerian: 128x256x26 grid



Finite Volume: 96x144x26 grid

Increased Decomposition in the Vertical



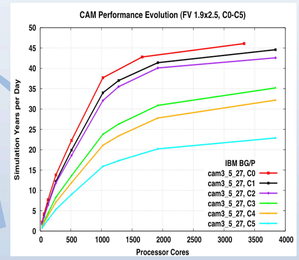
Finite Volume: 768x1152x26 grid on Atlas system at LLNL

Decomposing Over Tracer Index

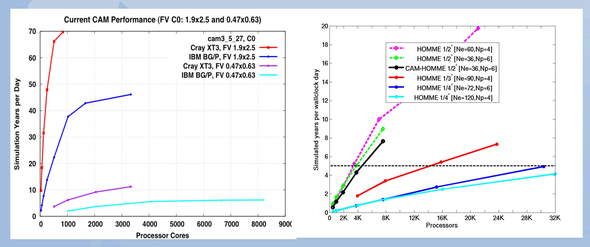
	tracer advection	atm run time
control	3	80
99 extra tracers	78	322
5*832 (=4160) proc	78	270
5-way decomp.	18	224

Finite Volume: 192x288x26 grid on Atlas system at LLNL. Uses 832 processors prior to optimization. Five-way tracer decomposition executes tracer advection 4.3 times faster. Overall throughput increases by 44%. (This includes benefit of extra processors for physics.)

Next Steps: New Science and Higher Resolutions



Finite Volume: 96x144x26 (C0) and 96x144x30 (C1-C5) on IBM BG/P system for a number of new science options being considered for production runs. Cost of new physics increases cost of model by 2.5X. Cost of other proposed changes, e.g., atmospheric chemistry or 48 vertical levels, will increase cost even more.



Finite Volume on Cray XT3 and IBM BG/P (left); HOMME on BG/L (right). Performance scalability work is far from complete, especially for higher resolution grid. Achieving further scalability will likely require moving to a more scalable dycore, such as HOMME or cubed sphere FV.

Future Work (near-term)

- Continue performance optimization of CCSM components on DOE high performance computing platforms for a range of problem configurations, focusing on high resolution simulations
- Complete implementation of parallel I/O and new CCSM architecture
- Optimize full CCSM for current problem resolutions on 1000-10000 processors

Future Work (longer term)

- Evaluate (and optimize where appropriate) alternative CAM dynamical cores that exhibit improved scalability, e.g.
  - NASA/NOAA cubed sphere finite-volume dycore
  - NSF/DOE HOMME dycore
- Optimize new physics and physical parameterizations: for vectorization (SSE/double hummer); for large numbers of advected fields (as with chemistry); for load balancing
- Further improve scalability of the coupler, targeting 10,000 to 100,000 processes, by additional memory minimization, more efficient support for component domain decompositions, and improved support for load balancing

Acknowledgements

- The work described in this presentation was sponsored by the Atmospheric and Climate Research Division and by the Office of Mathematical, Information, and Computational Sciences, both of the Office of Science, U.S. Department of Energy, under Contract No. DE-AC05-00OR22725 with UT-Battelle, LLC, Contract No. DE-AC52-07NA27344 with Lawrence Livermore National Security, LLC, and Contract No. DE-AC02-06CH11357 with UChicago Argonne, LLC.
- These slides have been authored by contractors of the U.S. Government under contracts No. DE-AC05-00OR22725, DE-AC52-07NA27344, and DE-AC02-06CH11357, and are released as LLNL Report LLNL-PRES- . Accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allows others to do so, for U.S. Government purposes.
- This work used resources of the National Center for Computational Sciences at Oak Ridge National Laboratory, which is supported by the Office of Science of the Department of Energy under Contract DE-AC05-00OR22725.