Performance Intercomparison of Community Atmosphere Model on High-End Computing Platforms

A.A. Mirin – Lawrence Livermore National Lab.
P.H. Worley – Oak Ridge National Lab.
W.B. Sawyer – Swiss Federal Institute Zurich
L. Oliker, M.F. Wehner – Lawrence Berkeley National Lab.
D. Parks – NEC Solutions America

Community Atmosphere Model

- Developed at National Center for Atmospheric Research (NCAR), with contributions from Department of Energy (DOE) and National Aeronautics and Space Administration (NASA) researchers
- Consists of physics and dynamics modules
- Physics evolves along vertical columns and includes radiation, clouds, precipitation, turbulent mixing, boundary layer effects,…
- Dynamics is modeled using choice of methods
  - this study uses finite-volume dynamical core developed by Lin and Rood while at NASA Goddard Space Flight Center (GSFC)

Finite-volume (FV) dynamical core

- Time-integrate dynamical equations within control volumes bounded by material surfaces
  - equations are vertically decoupled except for geopotential calculation, which involves vertical indefinite integrals at each horizontal location
  - conservative semi-Lagrangian method allows violation of longitudinal CFL condition near poles
- Remap Lagrangian surfaces to physical space based on vertical transport
  - remap is horizontally decoupled and conserves mass, momentum and total energy
- Equations solved on latitude/longitude/vertical staggered grid
- Overall method conserves mass

Parallelization of FV dycore

- Multi-two-dimensional domain decomposition methodology
  - latitude/vertical decomposition for main dynamics
  - latitude/longitude decomposition for surface remap and possibly geopotential
  - decompositions connected via transposes (Pilgrim, Mod_comm – NASA/GSFC)
- Assign subdomains to MPI tasks in 1:1 fashion
- Utilize OpenMP for shared memory parallelism
  - with respect to vertical coordinate for main dynamics
  - with respect to latitude for surface remap
- Optionally utilize (threaded) one-sided MPI-2 communications
- Optionally utilize shmem communications on Cray X1E

Vectorization for Cray X1E and Earth Simulator

- **Main dynamics is structured in nested fashion with outer loop over vertical level and inner loops over longitude**
  - ***move latitude loops to lowest level***

- **Finite-volume method is fundamentally one-sided (upwind) and higher order, leading to nested branching throughout**
  - ***perform logical tests in advance and partition latitude loops using indirect indexing***

- **Polar filters use FFTs along complete latitude lines, with vectorization across FFTs**
  - ***at fine domain decomposition there are insufficient latitude lines to efficiently vectorize***

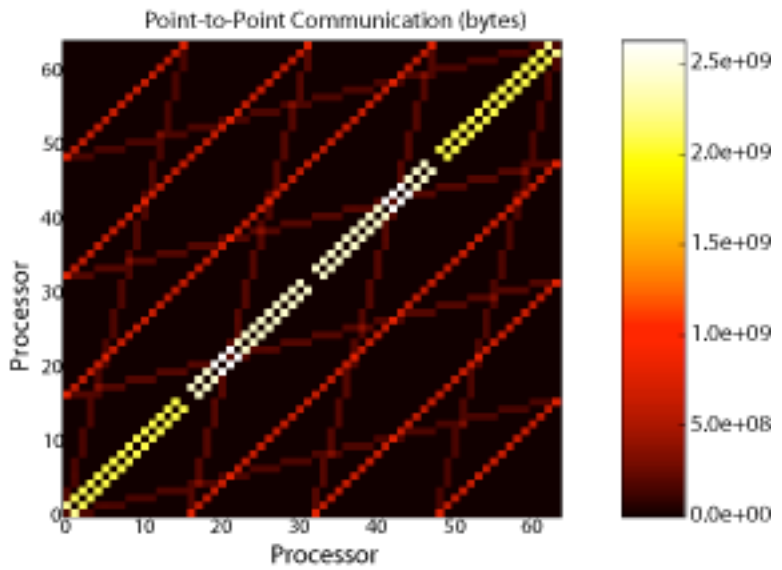
Physical parameterizations

- Uses same computational grid as dynamics, but computations along vertical columns are independent of one-another
- Basic computational structure is the *chunk,* an arbitrary collection of vertical columns
  - vectorization is over the columns in a chunk; MPI and OpenMP parallelism is over chunks
  - number of columns assigned to a chunk can be chosen to optimize for vectorization, cache blocking, and/or MPI or OpenMP parallelism
  - composition of chunk can be chosen to minimize load imbalance, and/or communication overhead when moving between dynamics and physics
- Optimal choice of chunk size and composition is architecture-specific

Architectural intercomparison

Node type	Where	Network	CPU/ Node	Clock GHz	Peak GFlop	Stream BW GB/s/P	Stream / Peak byte/flop	MPI BW GB/s/P	MPI Latency usec	Network Topology
X1E	ORNL	Custom	4	1.13	18.0	9.7	0.54	2.9	5.0	4D-hypercube
ES	ESC	IN	8	1.0	8.0	26.3	3.29	1.5	5.6	Crossbar
Power5	NERSC	Federation	8	1.9	7.6	5.3	0.71	1.1	4.7	Fat-tree
XT3	ORNL	Custom	1	2.4	4.8	?	?	?	?	3D-torus
Itanium2	LLNL	Quadrics	4	1.4	5.6	1.1	0.19	0.25	3.0	Fat-tree
Power4	ORNL	Federation	32	1.3	5.2	?	?	?	?	Fat-tree
Power3	NERSC	Colony	16	0.375	1.5	0.4	0.26	0.13	16.3	Fat-tree

Communication pattern for 16x4 latitude/vertical decomposition



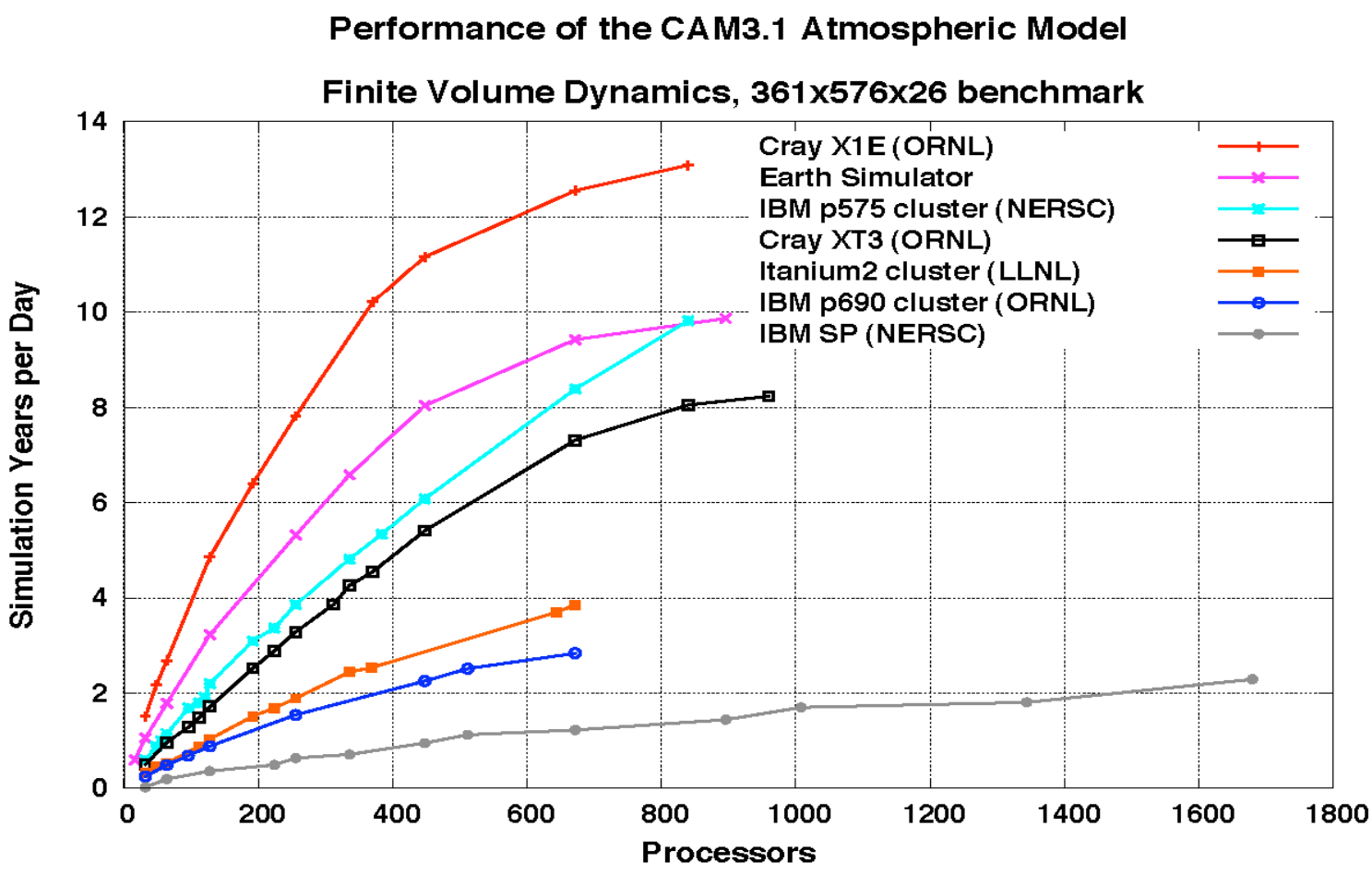
Dots adjacent to main diagonal represent latitudinal boundary communication

Dots parallel to but separated from main diagonal represent vertical geopotential communication

Dots in subdued diamond pattern represent transposes

With 1-D latitudinal decomposition, the only dots present are those adjacent to main diagonal

Architectural performance comparison



Limitations on processor count

- Mesh is 361 (latitude) x 576 (longitude) x 26 (vertical)
- Algorithm requires 3 latitude lines per subdomain
  - —limit is 120 subdomains in latitude
  - —performance often does not improve going from 4 to 3 latitude lines
- Parallel efficiency falls off rapidly with respect to vertical decomposition
  - —practical limit is generally 8 subdomains in vertical
- Maximum number of subdomains (practical limit) = 120 x 8 = 960
  - —this corresponds to 960 processors in absence of OpenMP

Benefits and drawbacks of OpenMP

- OpenMP is of benefit on IBM machines
  - with 8 OpenMP threads on seaborg, one can make effective use of 2944 (92x4x8) processors (vertical decomposition is limited to 4 subdomains)
- OpenMP is of no benefit on the LLNL Itanium2 machine
- OpenMP provides increased throughput on the Cray-X1E at coarse domain decomposition; however at fine domain decomposition, the correspondingly shorter vector lengths offset the gains of additional processor power
- OpenMP is of great benefit on the SGI Origin (Mirin and Sawyer, IJHPCA)

Alternative communications

- One-sided MPI-2 communications, particularly in combination with OpenMP, lead to significant gains on SGI Origin (Mirin and Sawyer, IJHPCA)
- One-sided MPI-2 communications are a detriment on the IBM (separate study)
- Both one-sided MPI-2 communications and Shmem show no benefit on the Cray X1E
  - neither work in conjunction with OpenMP
- One-sided MPI-2 communications with non-contiguous data supported on IBM but not on Cray X1E or SGI Origin

Limitations on scaling (fixed problem size)

- Parallel efficiency falls off with increasing processor count
  - increase in communication to computation ratio
  - increase in load imbalance
    - dynamics work per grid point is relatively uniform except for polar filtering
    - physics work per grid point varies with conditions (sunlight, moisture, etc.)
    - use of separate decomposition for physics improves load balance, at expense of extra communication
  - decrease in vectorization efficiency (shorter vector lengths)

References

- Oliker, L., J. Carter, M. Wehner, A. Canning, S. Ethier, A. Mirin, G. Bala, D. Parks, P. Worley, S. Kitawaki and Y. Tsuda, "Leading Computational Methods on Scalar and Vector HEC Platforms," Supercomputing 2005 Conference, Seattle (2005).
- Mirin, A.A. and W.B. Sawyer, "A Scalable Implementation of a Finite-Volume Dynamical Core in the Community Atmosphere Model," Int'l. Jour. Supercomputing Applications, 19 (3), August 2005.
- Worley, P.H. and J.B. Drake, "Performance Portability in the Physical Parameterizations of the Community Atmospheric Model", Int'l. Jour. Supercomputing Applications, 19 (3), August 2005.
- Worley, P.H., "Benchmarking using the Community Atmospheric Model," Proc. 2006 SPEC Benchmark Workshop, January 2006.

Acknowledgment

- This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098. This research used resources of the 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.
- The work of LLNL authors was performed under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48. LBNL authors were supported by the Office of Advanced Scientific Computing Research in the Department of Energy Office of Science under contract number DE-AC0376SF00098. ORNL authors were supported by the Office of Advanced Scientific Computing Research in the Department of Energy Office of Science under contract number DE-AC05-00OR22725 with UT-Batelle, LLC.

- Accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes.

- This is LLNL Report UCRL-PRES-218927.