SciDAC
Scientific Discovery through Advanced Computing

Office of Science
U.S. DEPARTMENT OF ENERGY

Performance Engineering in the Community Atmosphere Model

Patrick H. Worley
Oak Ridge National Laboratory

Arthur A. Mirin
Lawrence Livermore National Laboratory

John B. Drake
Oak Ridge National Laboratory

William Sawyer
Swiss Federal Institute of Technology

SciDAC 2006

June 26-29, 2006

Hyatt Regency Denver at Colorado Convention Center

Denver, CO

Office of Science | NCAR | Lawrence Livermore National Laboratory | Los Alamos | ornl | Pacific Northwest National Laboratory | Data Assimilation Office | SciDAC

Background

- Community Climate System Model (CCSM)
  - Fully-coupled, global climate model that provides state-of-the-art computer simulations of the Earth's past, present, and future climate states.
  - Comprised of a coupler and four component models: atmosphere, ocean, land, and sea ice.
  - 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).
- SciDAC project *Collaborative Design and Development of the Community Climate System Model for Terascale Computers* contributed to CCSM development over the past 5 years, focusing on
  - introduction of physical processes, e.g. biogeochemistry, important to DOE,
  - support for problem specifications, e.g. IPCC climate change scenarios, important to DOE, and
  - porting to and performance optimization on DOE computer systems.
- SciDAC project *Performance Evaluation Research Center* collaborated on the performance engineering activities.

Community Atmosphere Model (CAM)

- Atmospheric global circulation model
- Primary consumer of computer resources in typical CCSM simulations.
- Timestepping code with two primary phases per timestep
  - *Dynamics*: advances evolution equations for atmospheric flow
  - *Physics*: approximates subgrid phenomena, such as precipitation, clouds, radiation, turbulent mixing, …
- Multiple options for dynamics:
  - Spectral Eulerian (EUL) dynamical core (*dycore*)
  - Spectral semi-Lagrangian (SLD) dycore
  - Finite-Volume semi-Lagrangian (FV) dycore

  all using tensor product *latitude* x *longitude* x *vertical level* grid over the sphere, but not same grid, same placement of variables on grid, or same domain decomposition in parallel implementation
- Separate data structures for dynamics and physics and explicit data movement between them each timestep

CAM Performance Engineering Goals

1) Maximize single processor performance, e.g.
   a) Optimize memory access patterns
   b) Maximize vectorization or other fine-grain parallelism
2) Minimize parallel overhead, e.g.
   a) Minimize communication costs
   b) Minimize load imbalance
   c) Minimize redundant computation

for

- a range of target systems,
- a range of problem specifications (grid size, physical processes, …)
- a range of processor counts

while preserving maintainability and extensibility.

No optimal solution for all desired (platform,problem,processor count) specifications. Approach: compile-time and runtime optimization options.

CAM Performance Optimization Options

1. Physics data structures
   - Index range, dimension declaration
2. Physics load balance
   - Variety of load balancing options, with different communication overheads
   - SMP-aware load balancing options
3. Communication options
   - MPI protocols (two-sided and one-sided)
   - Co-Array Fortran
   - SHMEM protocols

   and choice of pt-2-pt implementations or collective communication operators
4. OpenMP parallelism
   - Instead of some MPI parallelism
   - In addition to MPI parallelism
5. Aspect ratio of dynamics 2D domain decomposition (FV-only)
   - 1D is latitude-decomposed only
   - 2D is latitude/longitude-decomposed in one part of dynamics, latitude/vertical-decomposed in another part, with remaps to/from the two decompositions during each timestep.

CAM Performance Experiments

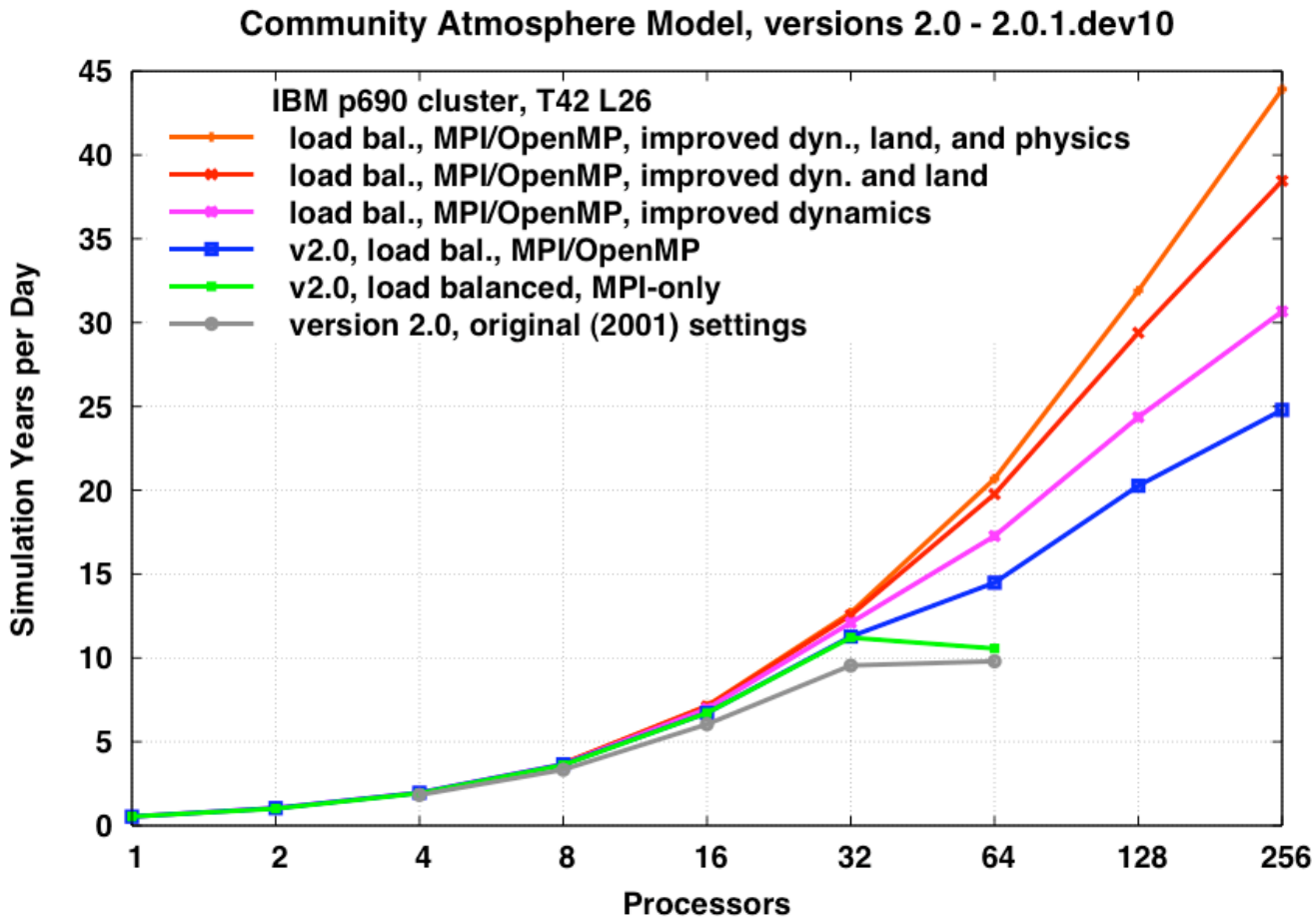
1. Spectral Eulerian dycore running on T42 L26 computational grid
   - 64x128x26 (latitude by longitude by vertical) grid
   - Production dynamical core and grid resolution in CCSM in 2002.
2. Spectral Eulerian dycore running on T85 L26 computational grid
   - 128x256x26 (latitude by longitude by vertical) grid
   - Current production dynamical core and grid resolution in CCSM
3. Finite Volume dycore running on 1.9x2.5 L26 computational grid
   - 96x144x26 (latitude by longitude by vertical) grid
   - Finite volume dycore is the preferred dycore for atmospheric chemistry due to its conservation properties. 1.9x2.5 L26 is the initial production grid size when using the FV dycore in CCSM.
4. Finite Volume dycore running on 1x1.25 L26 computational grid
   - 181x288x26 (latitude by longitude by vertical) grid
5. Finite Volume dycore running on 0.5x0.625 L26 computational grid
   - 361x576x26 (latitude by longitude by vertical) grid

Unless otherwise indicated, performance result for a given dycore, problem size, platform, and processor count is the optimal observed over all compile and runtime optimization options.

Experimental Platforms

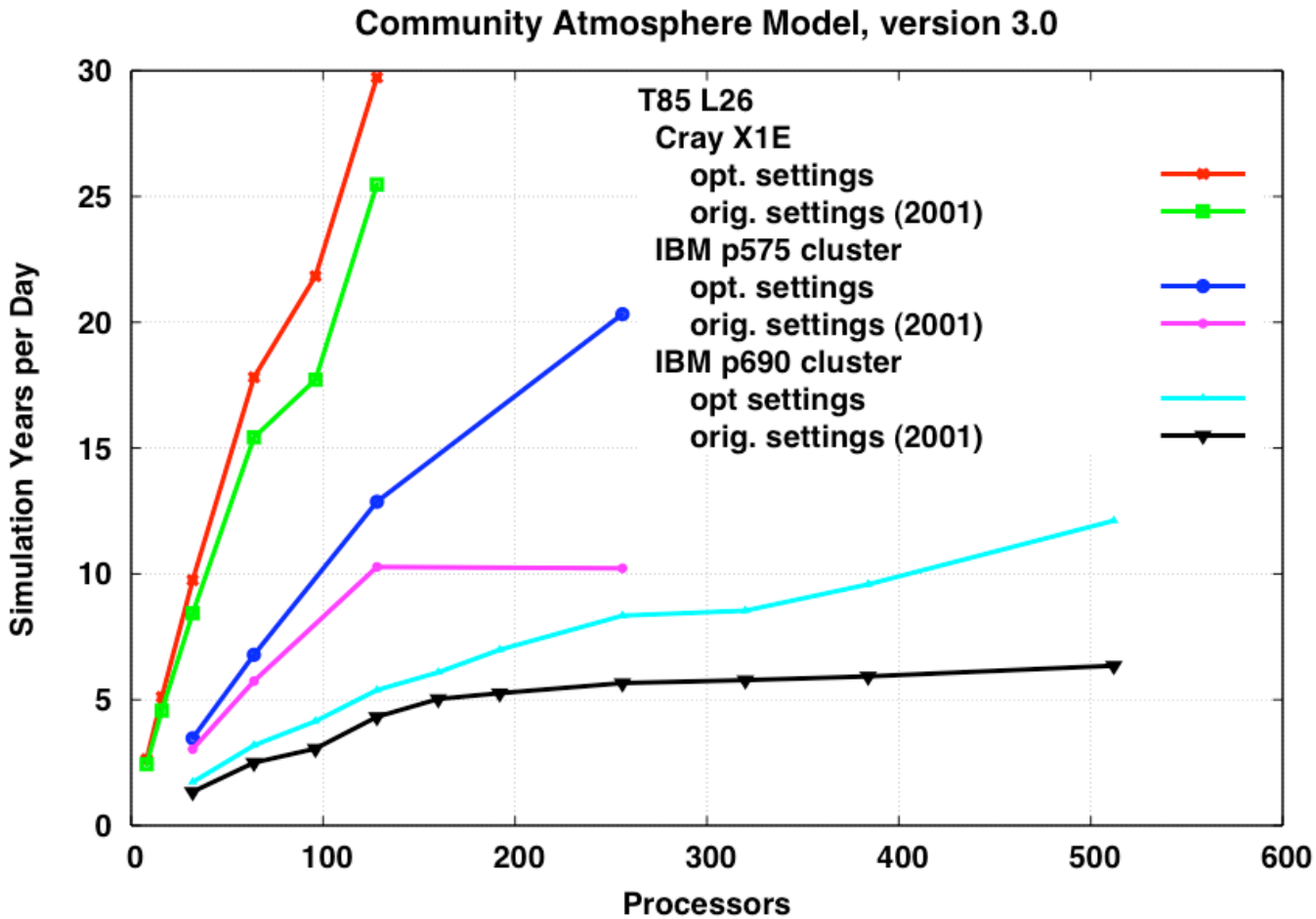
- **Cray X1** at Oak Ridge National Laboratory (ORNL): 128 4-way vector SMP nodes. Each processor has 8 64-bit floating point vector units running at 800 MHz. Nodes are fully connected within 4-node subsets, and are connected via 2-D torus between subsets.
- **Cray X1E** at ORNL: 256 4-way vector SMP nodes. Each processor has 8 64-bit floating point vector units running at 1.13 GHz. Nodes are fully connected within 8-node subsets, and are connected via 2-D torus between subsets.
- **Cray XT3** at ORNL: 5294 single processor nodes (2.4 GHz AMD Opteron) and a 3-D torus interconnect.
- **Earth Simulator**: 640 8-way vector SMP nodes and a 640x640 single-stage crossbar interconnect. Each processor has 8 64-bit floating point vector units running at 500 MHz.
- **IBM p575 cluster** at the National Energy Research Scientific Computing Center (NERSC): 122 8-way p575 SMP nodes (1.9 GHz POWER5) and an HPS interconnect with 1 two-link network adapter per node.
- **IBM p690 cluster** at ORNL: 27 32-way p690 SMP nodes (1.3 GHz POWER4) and an HPS interconnect with 2 two-link network adapters per node.
- **IBM SP** at NERSC: 184 Nighthawk II 16-way SMP nodes (375MHz POWER3-II) and an SP Switch2 with two network adapters per node.
- **Itanium2 cluster** at Lawrence Livermore National Laboratory (LLNL): 1024 4-way Tiger4 nodes (1.4 GHz Intel Itanium 2) and a Quadrics QsNetII Elan4 interconnect.
- **SGI Origin 3800** at NASA Ames Research Center: 1024 processors and a NUMAlink interconnect that supports cccNUMA global shared memory. Each processor is a 600 MHz MIPS R14000.

CAM EUL Performance History: to Nov. 2002



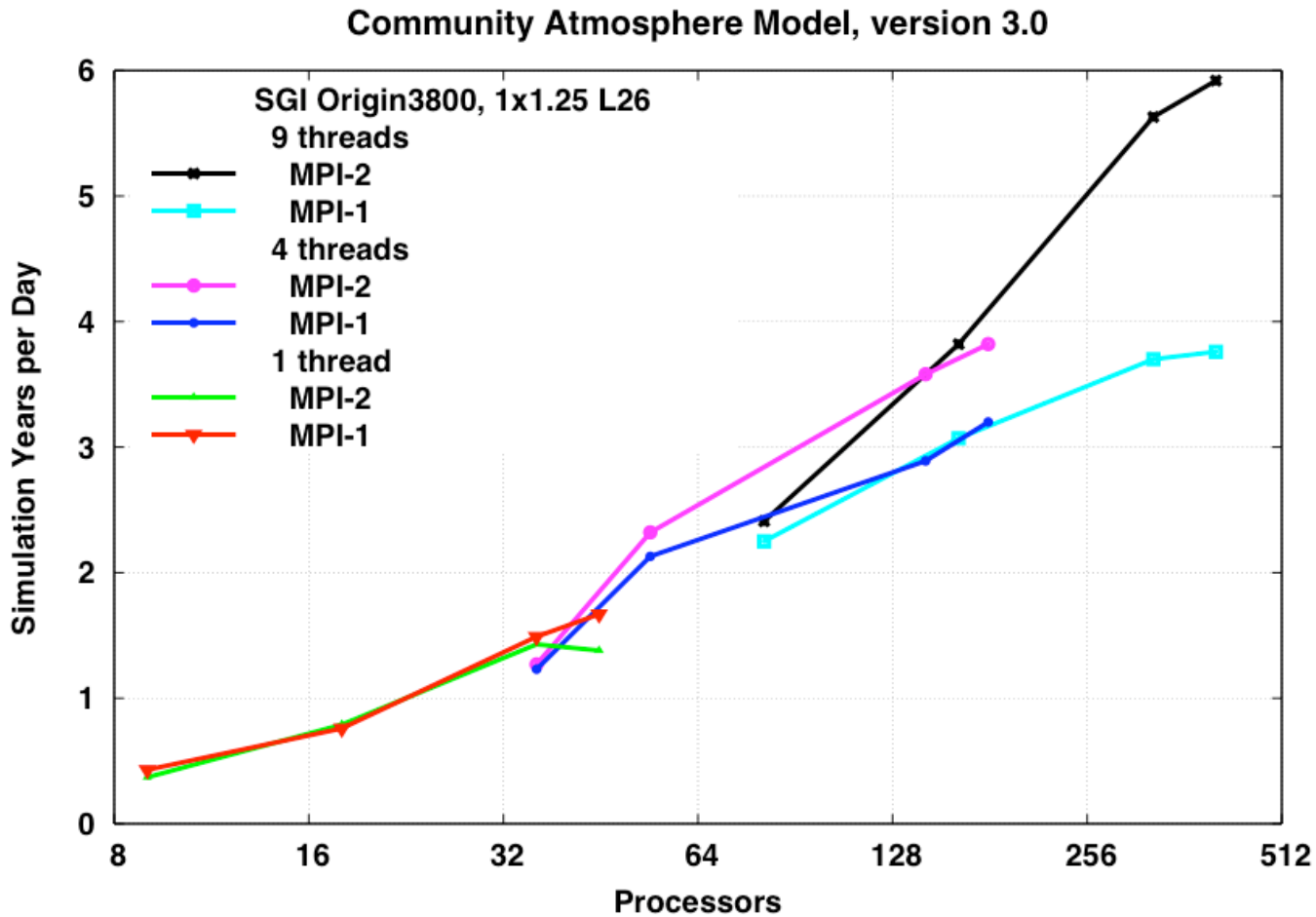
- Performance impact of SciDAC contributions to CAM on the IBM p690 cluster when using the spectral Eulerian dycore from June 2001 to November 2002.
- “original settings” curve is performance when running version 2.0 with settings corresponding to how version 1.0 was run in 2001.
- p690 cluster interconnect was the SP Switch2 during this time period.

CAM EUL Performance History: to May 2004



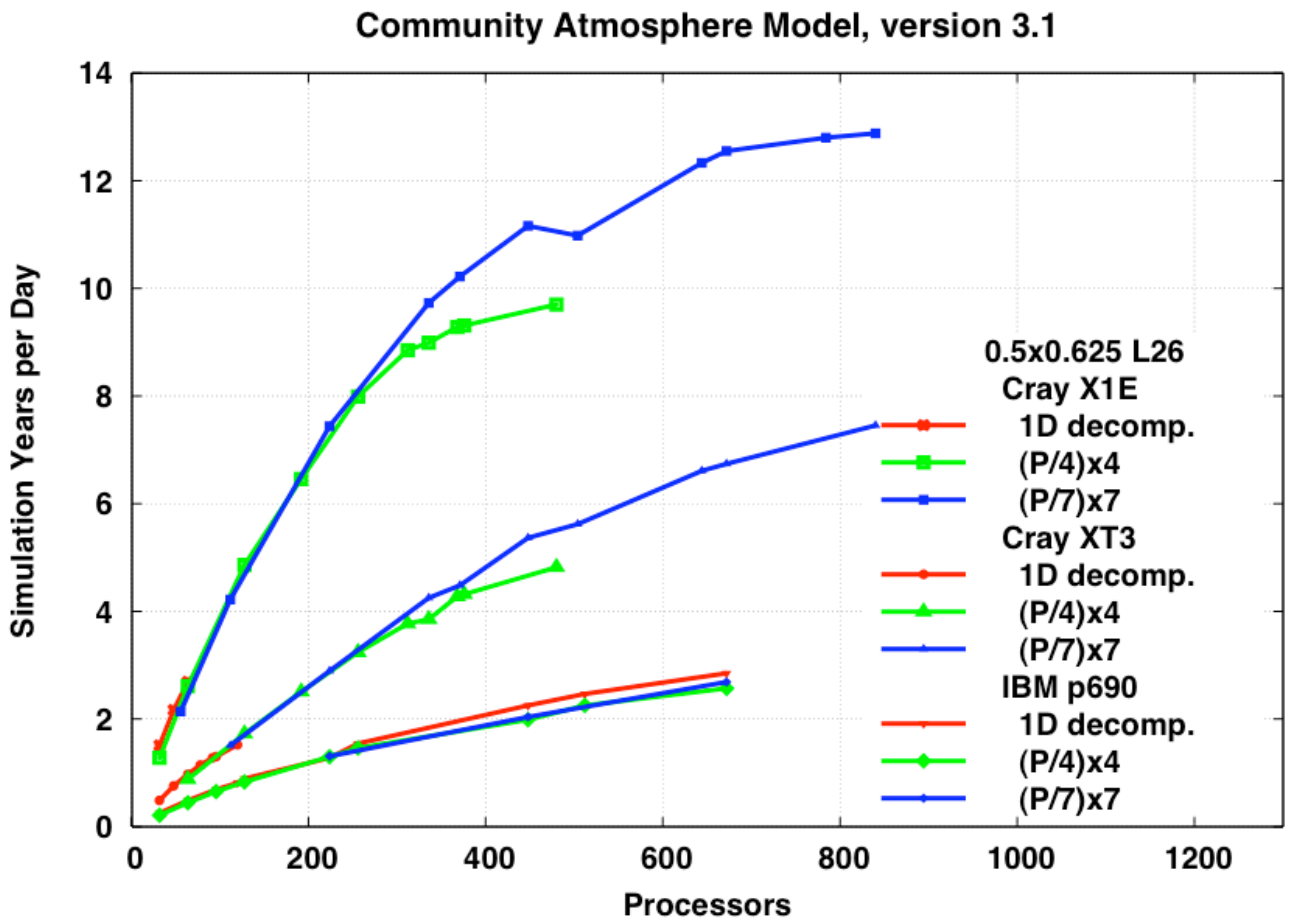
- Performance impact of SciDAC-introduced CAM tuning options (up to May 2004). Comparing performance of optimal settings with performance of settings corresponding to how version 1.0 was run in 2001.
- p690 cluster interconnect was HPS for these experiments.

CAM FV Performance: MPI-1 vs. MPI-2



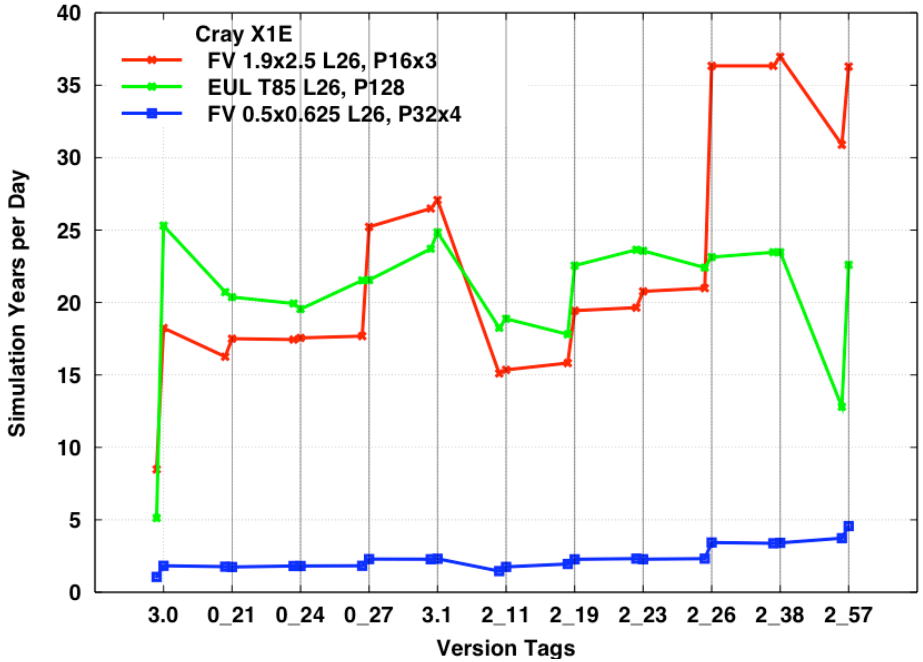
- Performance impact of MPI-2 one-sided communications and threading on the Origin3800.
- One-sided communication and threading do not improve performance on all platforms, and it is important that the communication protocols are options in CAM.

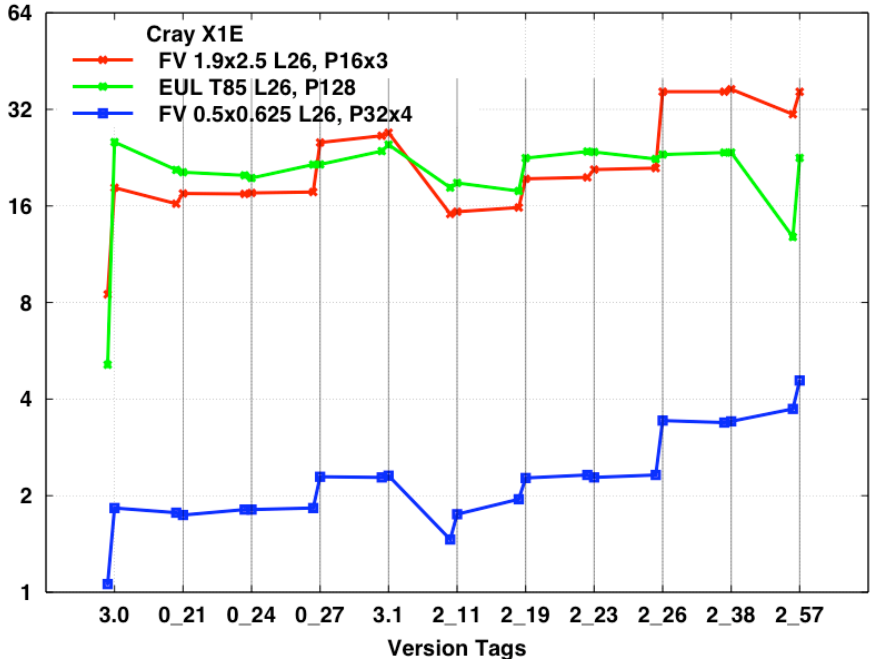
CAM FV Performance: 1D vs. 2D decomp.



- Performance impact of different domain decomposition strategies.
- Performance degrades when too many processors applied to latitude dimension. 2D decompositions improve performance, and scalability, for large processor counts.
- OpenMP parallelism is more efficient than 2D decomposition on p690 cluster.

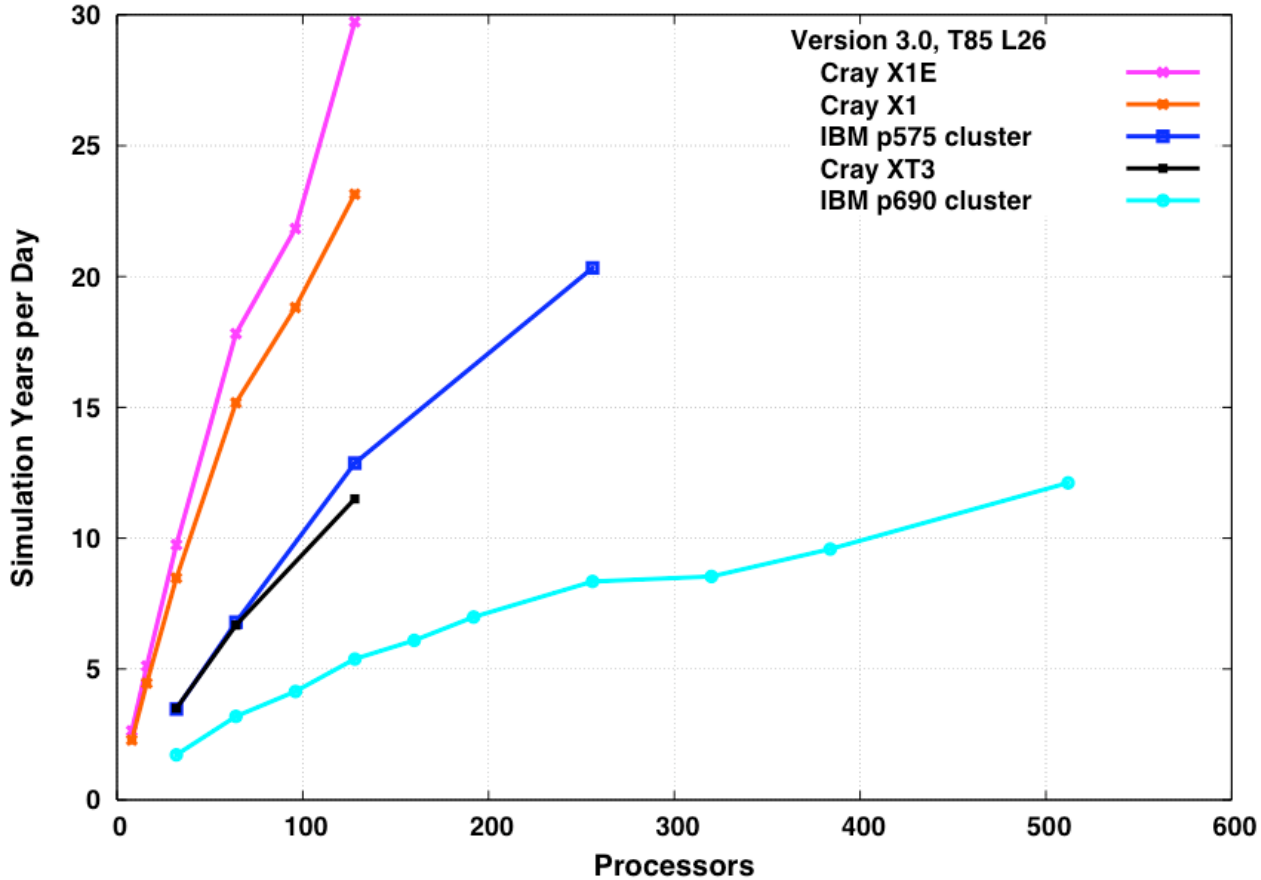
CAM Performance History: Vectorization





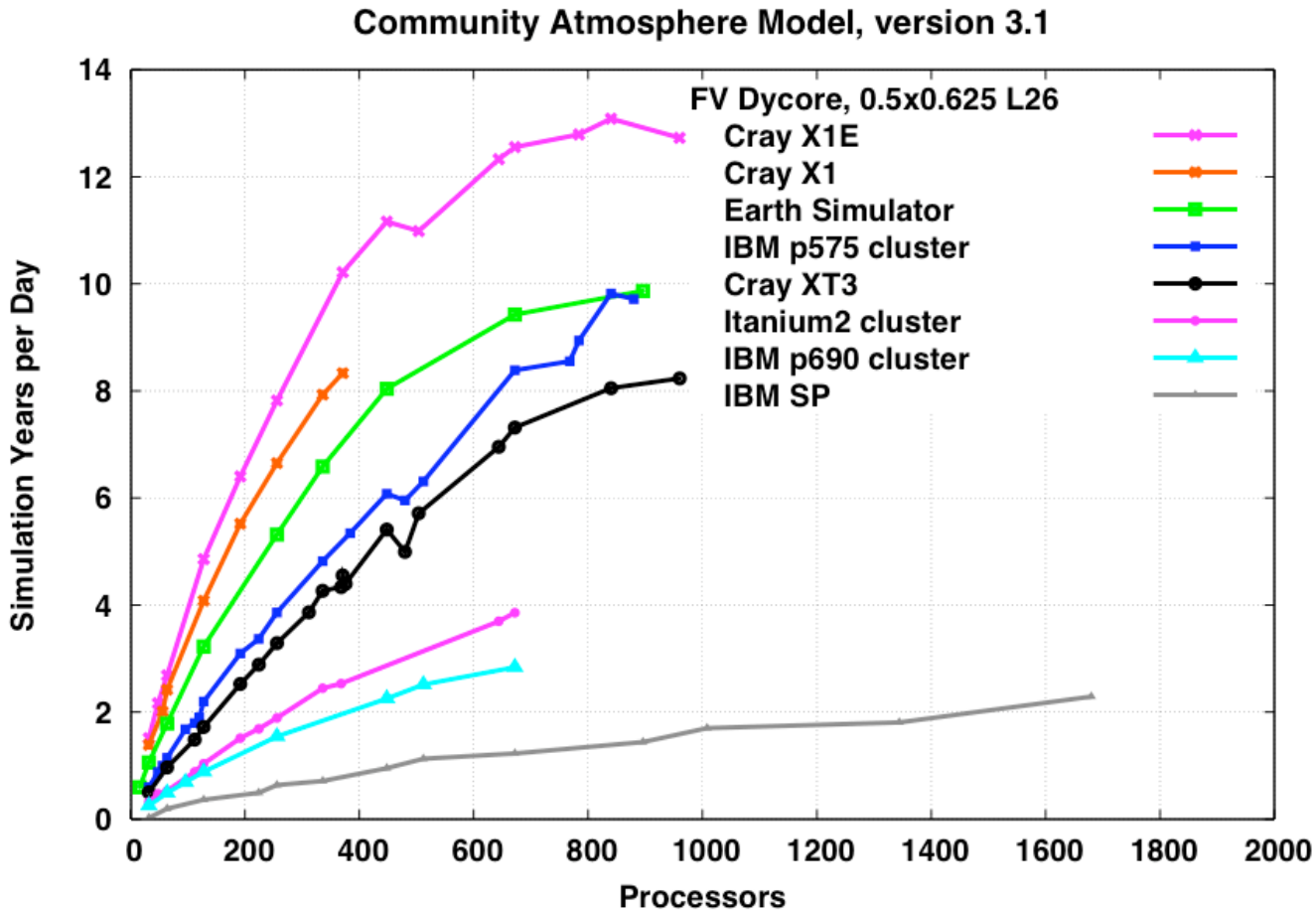
- Performance impact of SciDAC check-ins from March 2004 to April 2006 on the Cray X1E, plotting performance for both named version tag and for immediately preceding version.
- Not all check-ins improved performance, nor were expected to - some improved portability, added new performance tuning options, or fixed bugs.
- Maintaining vector performance as CAM evolves is (and will be) as important as further performance improvements.

Current CAM EUL Performance



- Maximum number of MPI processes is 128 for T85 L26. IBM systems use OpenMP to increase scalability.
- Recent performance optimizations backported into CAM 3.0 for these experiments.

Current CAM FV Performance

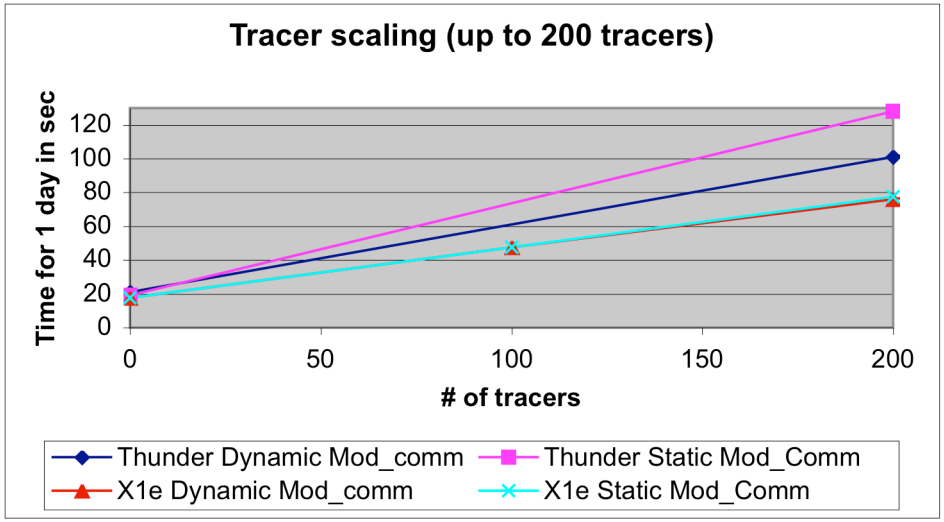


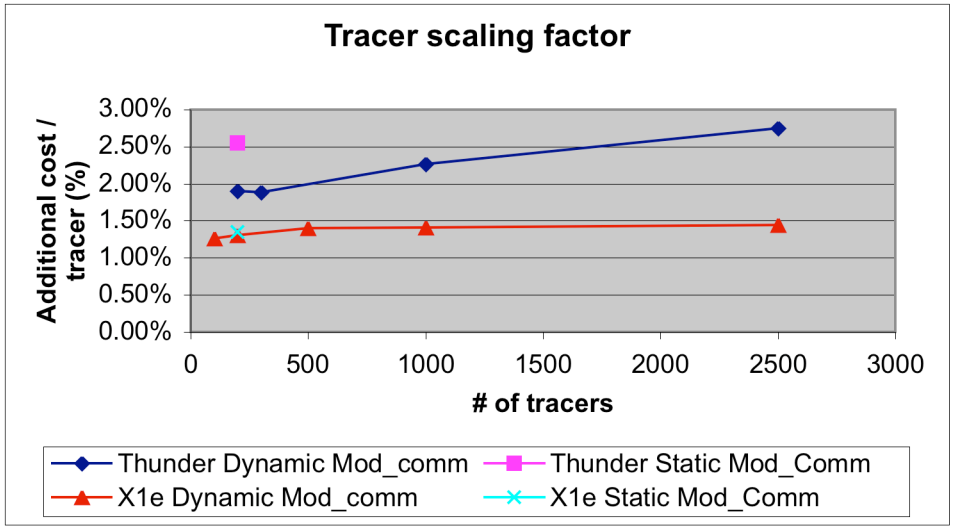
- Earth Simulator results courtesy of D. Parks. SP results courtesy of M. Wehner.
- Maximum number of MPI processes is 960 for 0.5x0.625 L26. IBM systems and Earth Simulator use OpenMP to increase scalability.
- Recent performance optimizations backported into CAM 3.1 for these experiments.

Current Limits to FV Scalability

1. Polar filter introduces load imbalances, especially on vector systems (because the small number of short FFTs do not vectorize well).
2. Requirement that at least 3 latitudes and 3 vertical levels be present in each "block" of domain decomposition within FV dycore. For D grid with 26 vertical levels, limit is 120x8 processor grid, or 960 MPI processes. For 1.9x2.5 degree grid, limit is 32x8 processor grid, or 256 MPI processes. (Note that the vertical level decomposition restriction can probably be relaxed, but performance is likely to be poor when using fewer than 3 levels per block.)
3. Physics can use many more processors, but is currently limited to the same number of MPI processes as the dynamical core. OpenMP can be used to assign more processors to physics than to dynamics, mitigating this to some degree. There is also some OpenMP parallelism available within the dynamics.
4. On vector systems, additional parallelism in physics is of limited utility, as vector length drops below 220 for D grid when using more than 960 processors (and drops below 110 in radiation routines).

New Issue: Tracer Transport





- Atmospheric chemistry introduces not only additional cost in the physics, but also requires the advection of many new fields.
- Experiment measures CAM runtime when advecting additional fields for 1x1.25 grid using a 48x7 processor grid on the LLNL Itanium2 cluster and a 48x4 processor grid on the X1E.
- Cost per tracer is minimally 1-2%, so more than doubles CAM runtime when advecting 100 new fields.

New and Planned Activities

1. Dynamics Scaling
   - Generalize dynamics/physics interface to support dycores not using lon/lat grid. Investigate new, more scalable dycores, for example, FV on a cubed sphere computational grid.
2. Physics Scaling
   - Add support to use different numbers of MPI processes in the dynamics and in the physics (generalizing current OpenMP approach to pure MPI codes).
3. Atmospheric chemistry
   - Investigate parallelizing advection over species and optimizing interprocess communication.
   - Investigate 3D decomposition for chemistry.
4. Increasing model resolution exacerbates the current I/O bottlenecks and memory impact of the (few) remaining global arrays.
   - Investigate parallel I/O on target platforms.

Summary

- Over the last 5 years, SciDAC-funded activities have improved performance of CAM significantly, by
  - a factor of 4.5 on IBM p690 cluster for T42 L26;
  - a factor of at least 2 on current IBM SMP clusters for T85 L26 due to parallel algorithm improvements;
  - a factor of over 3 on the XT3 and X1E for 0.5x0.625 L26 due to the 2D domain decomposition;
  - a factor of 5 for T85 L26, 1.9x2.5 L26, and 0.5x0.625 L26 on the X1E from vectorization modifications and parallel algorithm improvements;
  - a factor of at least 4 from porting to new platforms.
- Performance engineering activities included
  - improving serial performance and decreasing parallel overhead;
  - adding compile- and runtime options that improve performance portability;
  - porting and optimizing on new platforms;
  - interface design that allows new dycores to be integrated with CAM efficiently.
- Increases in problem complexity and porting to petascale computer architectures will require further improvements in scalability in order to maintain required simulation throughput.

Source Material

A. Mirin and W. Sawyer, *A Scalable Implementation of a Finite-Volume Dynamical Core in the Community Atmosphere Model*, International Journal for High Performance Computer Applications, 19(3), August 2005, pp. 203-212.

W. Putman, S-J. Lin, and B-W. Shen, *Cross-Platform Performance of a Portable Communication Module and the NASA Finite Volume General Circulation Model*, International Journal for High Performance Computer Applications, 19(3), August 2005, pp. 213-223.

L. Oliker, 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*, in Proceedings of the ACM/IEEE International Conference for High Performance Computing, Networking and Storage (SC05), Seattle, WA, November 12-18, 2005.

P. Worley, *Benchmarking using the Community Atmosphere Model*, in Proceedings of the 2006 SPEC Benchmark Workshop, Austin, TX, January 23, 2006.

P. Worley, *Performance of the Community Atmosphere Model on the Cray X1E and XT3*, in Proceedings of the 48th Cray User Group Conference, Lugano, Switzerland, May 8-11, 2006.

P. Worley and J. Drake, *Performance Portability in the Physical Parameterizations of the Community Atmospheric Model*, International Journal for High Performance Computer Applications, 19(3), August 2005, pp. 187-201.

P. Worley, A. Mirin, J. Drake, and W. Sawyer, *Performance Engineering in the Community Atmosphere Model*, in Proceedings of the 2006 SciDAC Conference, Denver, CO, June 26-29, 2006.

Acknowledgements

- Research sponsored by the Atmospheric and Climate Research Division and the Office of Mathematical, Information, and Computational Sciences, Office of Science, U.S. Department of Energy under Contract No. DE-AC05-00OR22725 with UT-Battelle, LLC and Contract No. W-7405-Eng-48 with the University of California Lawrence Livermore National Laboratory, and under the auspices of the NASA Goddard Space Flight Center, NASA Task 00-9103-01a.
- These slides have been authored by contractors of the U.S. Government under contracts No. DE-AC05-00OR22725 and No. W-7405-Eng-48, and are released as LLNL Report UCRL-PRES-222116. 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 research 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, and 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.