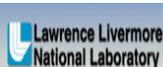


Performance Engineering of the Community Atmosphere Model

Patrick H. Worley
Oak Ridge National Laboratory

Arthur A. Mirin
Lawrence Livermore National Laboratory

11th Annual CCSM Workshop
June 20-22, 2006
The Village at Breckenridge
Breckenridge, CO



Data Assimilation Office



Overview

- Recent SciDAC-sponsored Community Atmosphere Model (CAM) software and performance engineering activities:
 - Introducing and maintaining performance optimizations on both vector and nonvector computer systems as code evolves
 - Improving portability and performance portability
 - Measuring and analyzing performance
- Poster Outline
 - Performance impacts of recent CAM modifications
 - CAM performance tuning options
 - CAM performance and performance analysis
 - Performance issues and future activities

Community Atmosphere Model (CAM)

Atmospheric global circulation model

- 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) dycoreall 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 (in a “coupler”)
- Developed at NCAR, with contributions from the Department of Energy (DOE) and the National Aeronautics and Space Administration (NASA).

CAM Performance Experiments

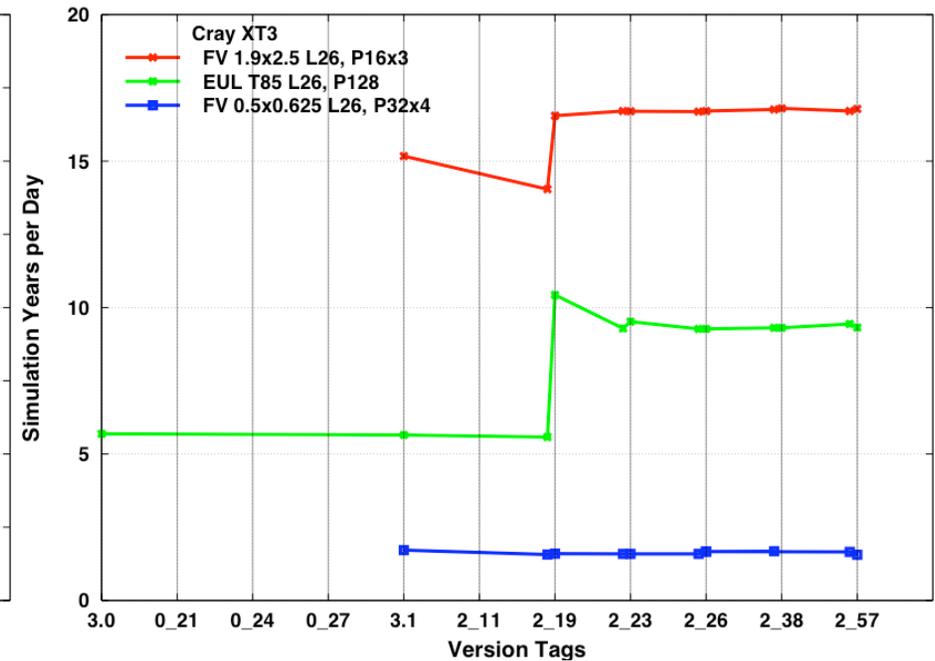
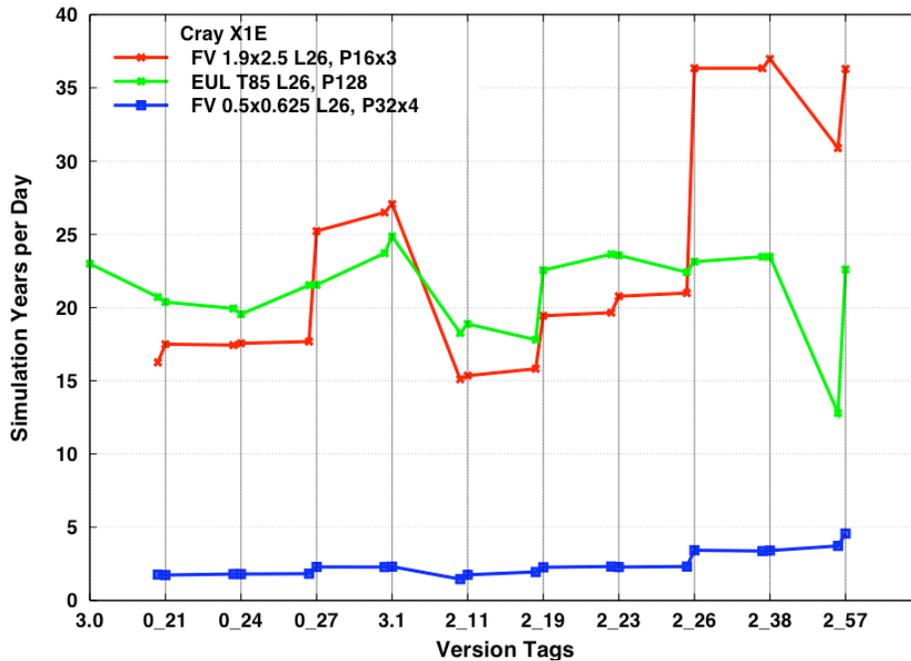
1. Spectral Eulerian dycore running on T85L26 computational grid
 - 128x256x26 (latitude by longitude by vertical) grid
 - Current production dynamical core and grid resolution in CCSM
2. Finite Volume dycore running on 1.9x2.5 degree horizontal grid with 26 vertical levels
 - 96x144x26 (latitude by longitude by vertical) grid
 - Finite volume dycore is the preferred (required among current options) dycore for atmospheric chemistry due to its conservation properties. 1.9x2.5 degree resolution is the initial CCSM production grid size.
3. Finite Volume dycore running on 0.5x0.625 degree horizontal grid ('D grid') with 26 vertical levels
 - 361x576x26 (latitude by longitude by vertical) grid
 - 15 times larger than FV production grid resolution.

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 Altix 3700** at ORNL: 128 2-way SMP nodes and a NUMAflex fat-tree interconnect supporting cccNUMA global shared memory. Each processor is a 1.5 GHz Itanium 2 with a 6 MB L3 cache.
- **SGI Altix 3700 Bx2** at NASA: 1024 2-way SMP nodes and a NUMAflex fat-tree interconnect supporting NUMA global shared memory. Each processor is a 1.6 GHz Itanium 2 with a 9 MB L3 cache.

CAM Performance History: X1E and XT3



- Performance impact of recent SciDAC check-ins on the Cray X1E and XT3.
- Not all check-ins improved performance, nor were expected to - some improved portability, added new performance tuning options, or fixed bugs.
- Maintaining performance as CAM evolves is (and will be) as important as further performance improvements.

CAM Performance Portability 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.

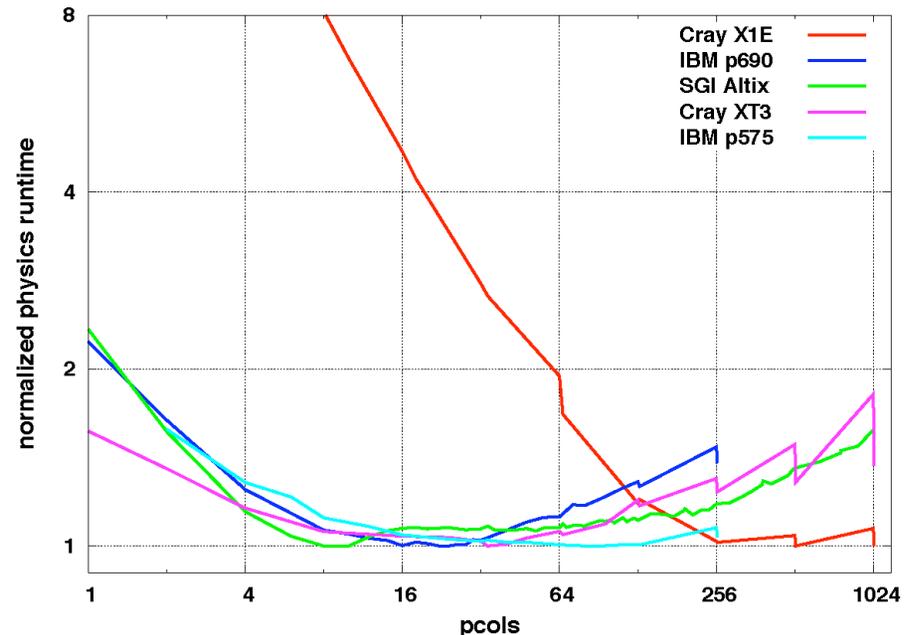
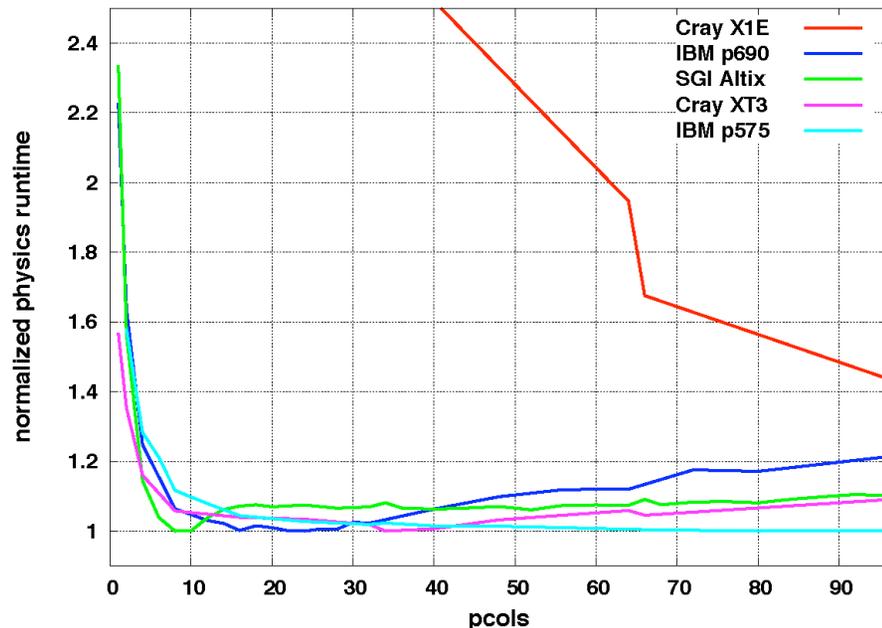
No optimal solution for all desired (platform,problem,processor count) specifications. Approach: compile-time and runtime optimization options.

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.

Performance Tuning Physics Data Structures



pcols parameter determines vector length and cache locality in physics.

Altix: minimum at pcols = 8

p575: minimum at pcols = 80

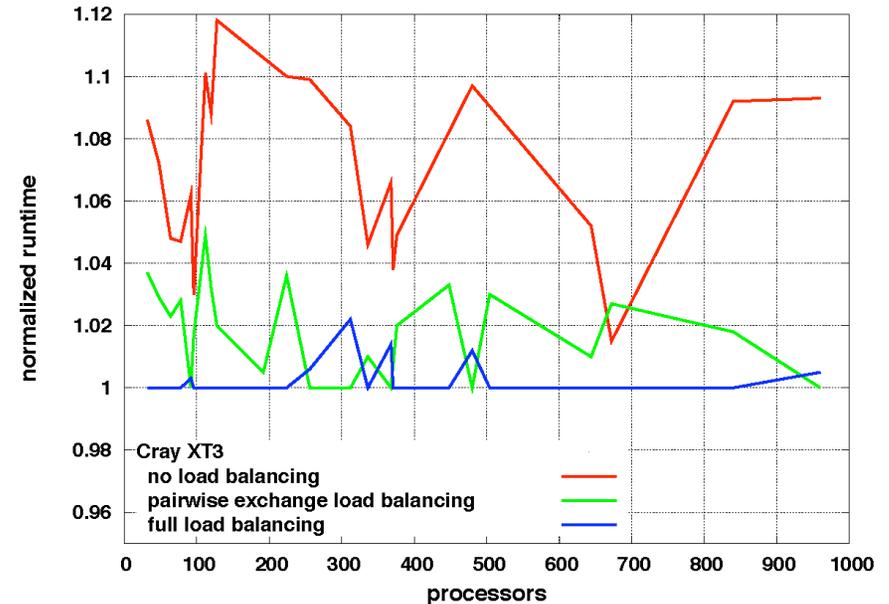
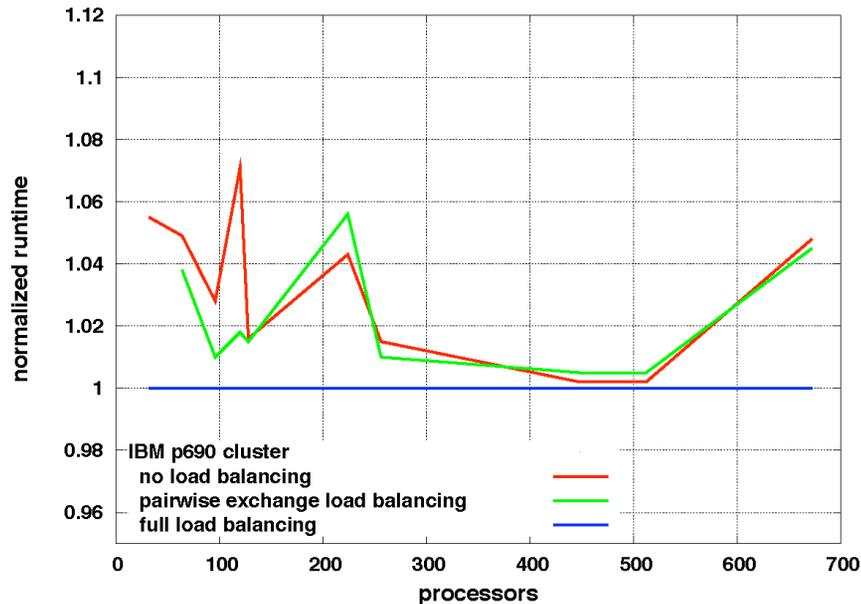
p690: minimum at pcols = 24

X1E: minimum at pcols = 514 or 1026

XT3: minimum at pcols = 34

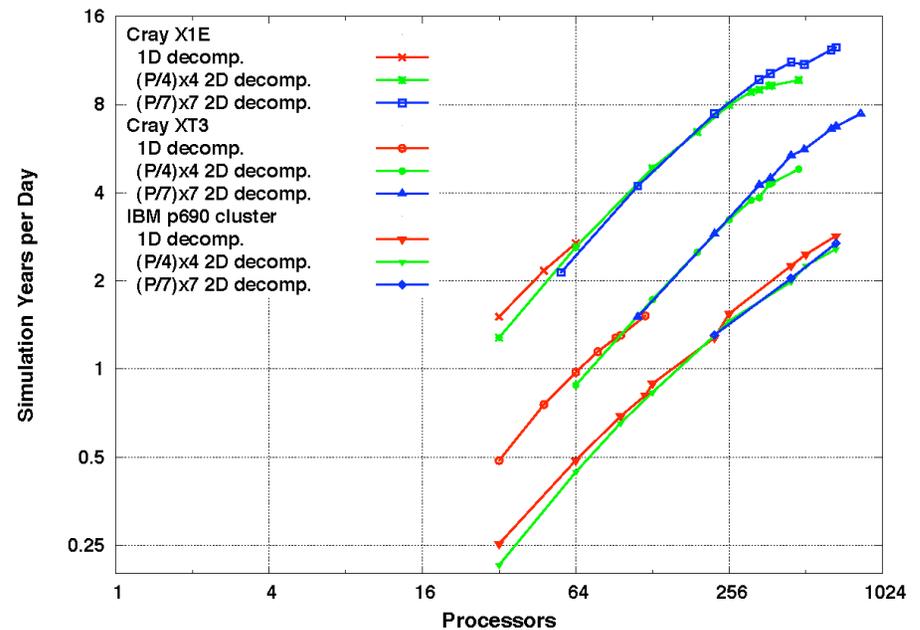
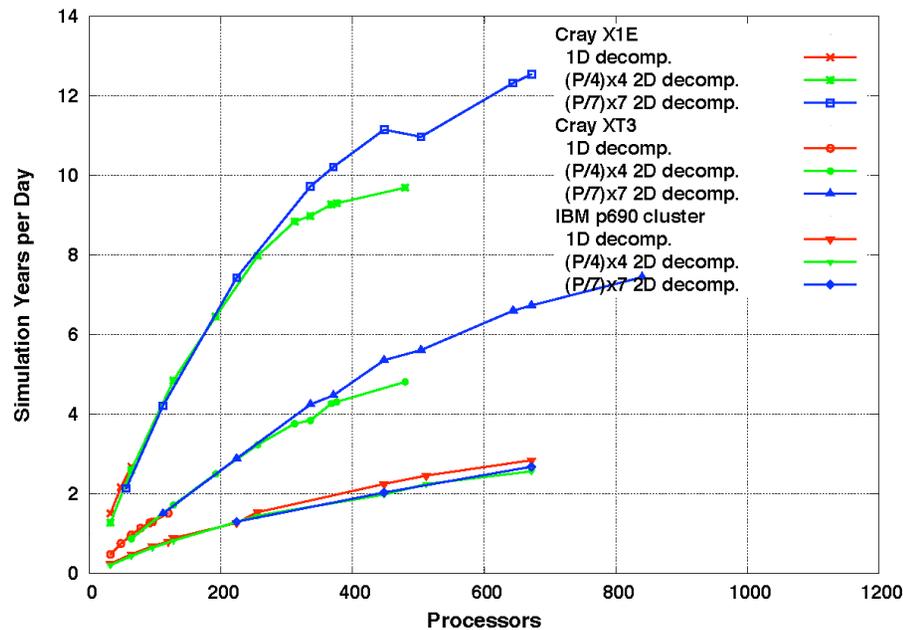
pcols <= 4 bad for all systems.

Evaluating Load Balancing: FV / D Grid



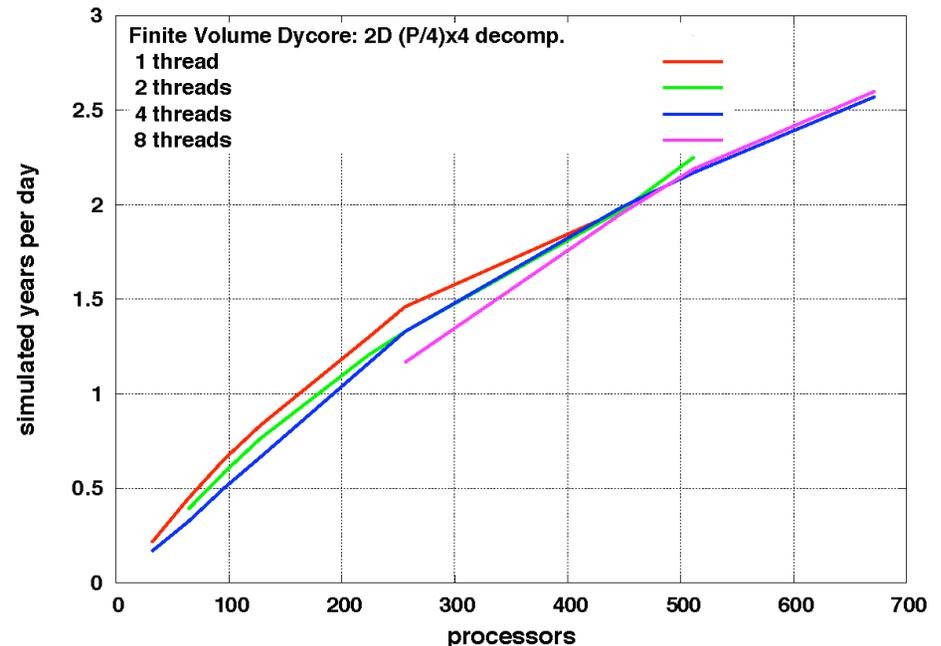
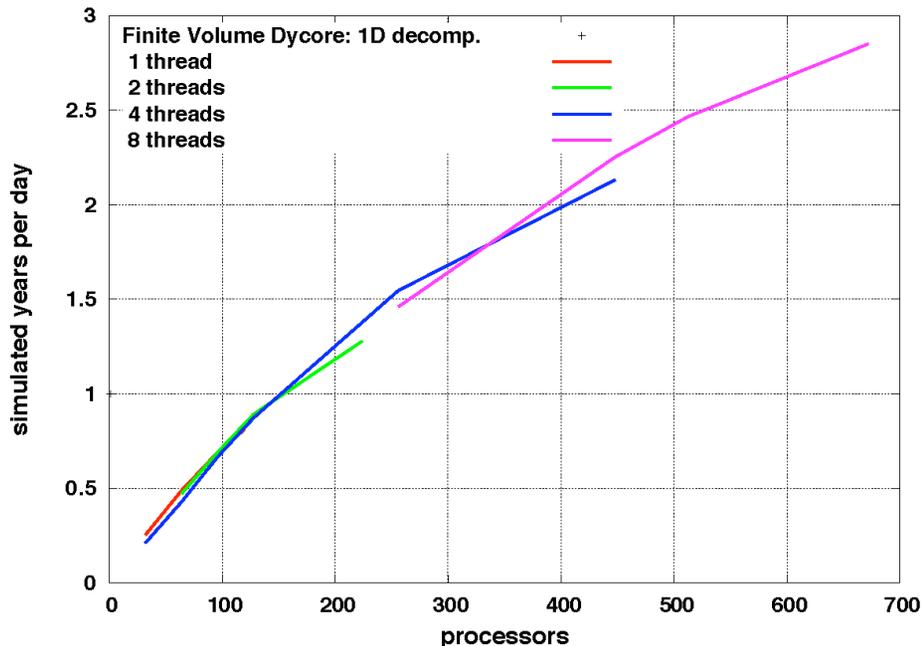
- On the IBM p690, full load balancing is always best, but no load balancing is, at worst, only 7% slower.
- On the Cray XT3, full load balancing is usually best, but pairwise exchange load balancing is competitive. No load balancing is usually more than 4% slower, and as much as 12% slower.
- The best example of the advantage of full load balancing is on the Cray X1E. It is so much better that it was clear early on in the tuning process and the other load balancing options were “pruned” from the search tree.

1D vs. 2D Decompositions: FV / D Grid



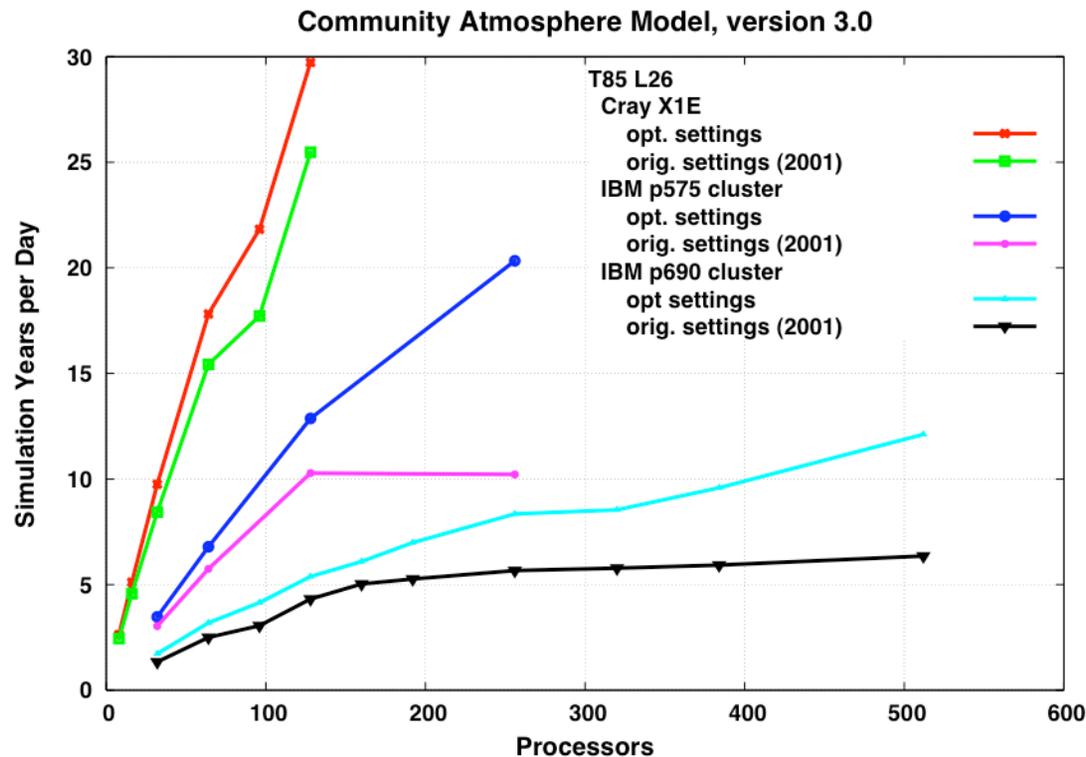
- 2D decomposition is superior to 1D when number of MPI processes decomposing latitude in 1D exceeds some limit, ~70 on two Cray systems (not using OpenMP). Similarly, (P/7)x7 decomposition outperforms (P/4)x4 when the number of MPI processes decomposing latitude in (P/4)x4 exceeds ~70.
- On IBM p690, OpenMP and the 1D decomposition is superior to 2D decompositions up to 672 processors, even when the optimum for 1D uses more threads per process than the optimum for 2D. (Note: 1D never uses more than 84 MPI processes.)

MPI vs. OpenMP: FV / D Grid on p690 cluster



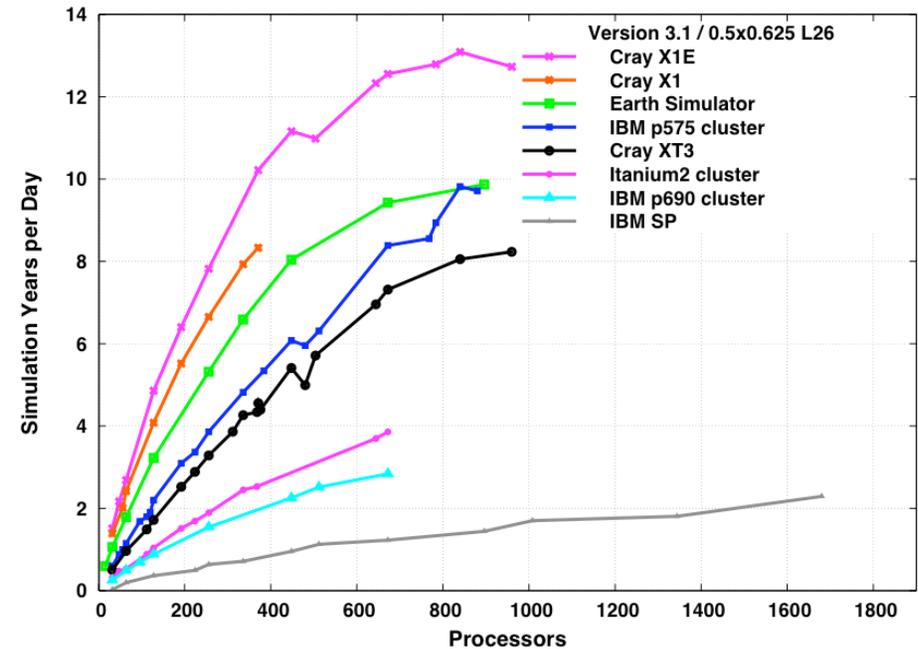
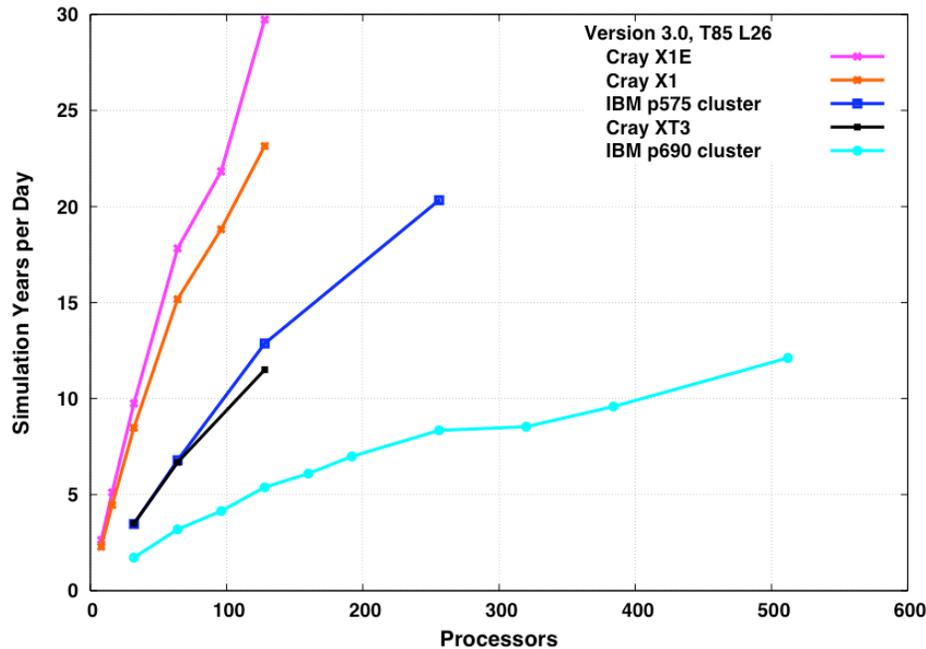
- Primary utility of OpenMP is to extend scalability.
- For 2D decomposition, fewer threads is generally better for the same number of processors.
- For 1D decomposition, it is more important not to let the number of processors decomposing latitude exceed approx. 64 than it is to minimize the number of threads.

Performance Tuning Impact: EUL / T85L26



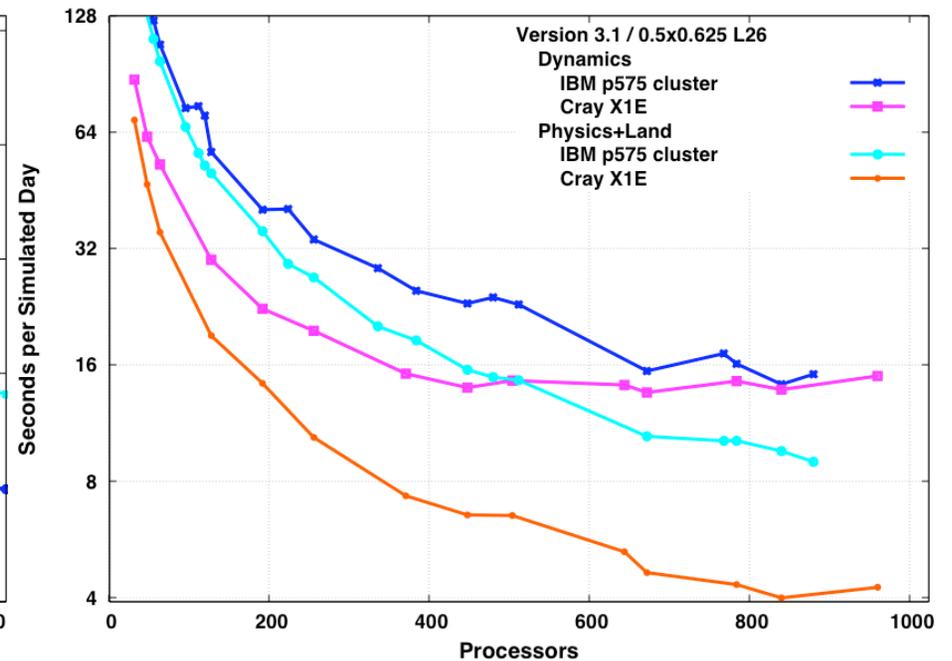
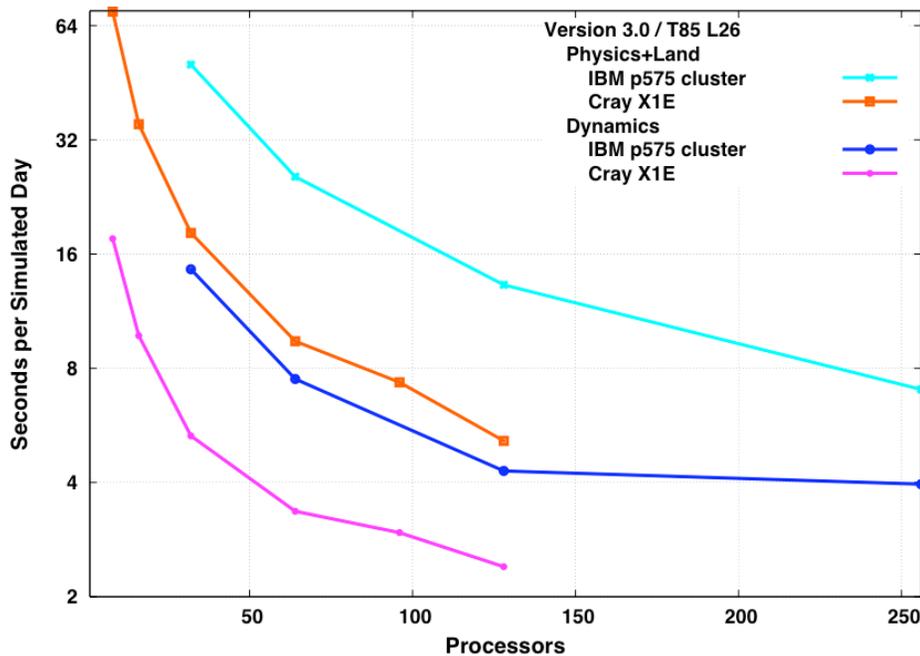
- Original settings - tuning option settings representative of what was used in 2001.
- Big win on IBM, primarily due to ability to use OpenMP parallelism in physics when dynamics parallelism exhausted (at approximately 128 processors).
- As vector length for original settings is near optimal on the X1E, performance difference is primarily due to load balancing.

Platform Comparisons



- Earth Simulator results courtesy of D. Parks. SP results courtesy of M. Wehner.
- Maximum number of MPI processes is 128 for T85 L26 and 960 for 0.5x0.625 L26. IBM systems and Earth Simulator use OpenMP to increase scalability.
- Recent performance optimizations backported into CAM 3.0 and CAM 3.1 for these experiments.

Performance Diagnosis: p575 vs. X1E



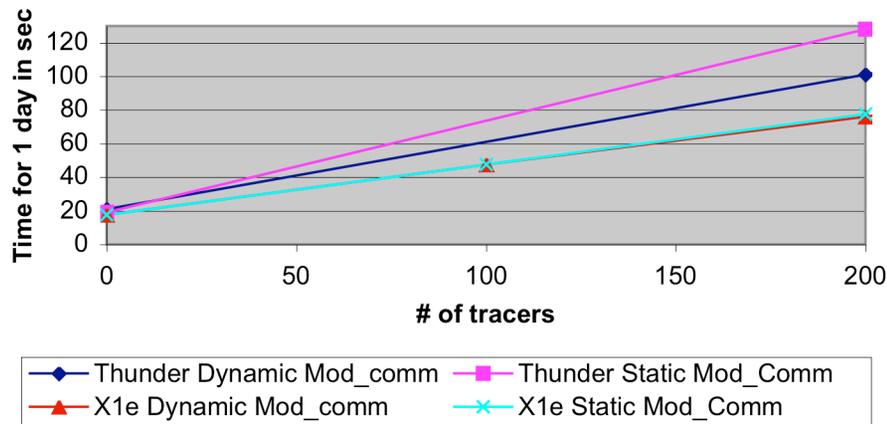
- For T85 L26, physics more expensive than dynamics. For 0.5x0.625 L26, the reverse is true.
- For T85 L26, OpenMP parallelism improves physics scalability on p575, but has no effect on dynamics performance.
- For 0.5x0.625 L26, dynamics performance on p575 approaches that of X1E. However, p575 uses OpenMP and X1E does not in these experiments. P575 using a coarser dynamics domain decomposition for same processor count.

Current Limits to FV Scalability

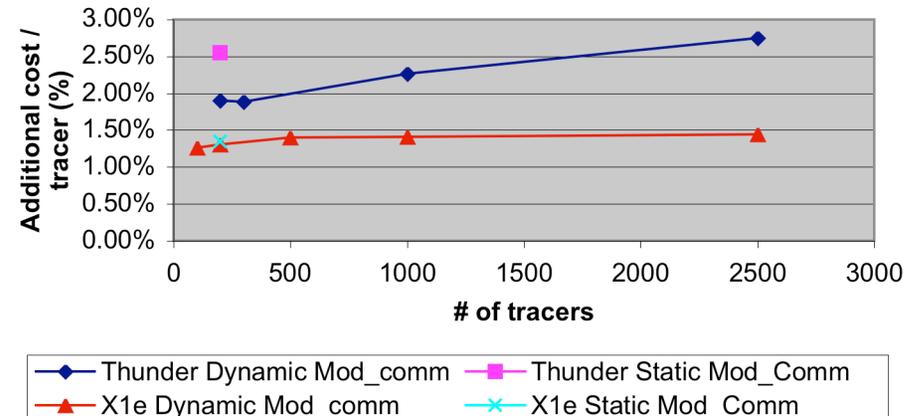
1. Polar filter introduces load imbalances, especially on vector systems (because the small number of short FFTs do not vectorize well).
2. Requirement that at least 3 latitudes and 3 vertical levels be present in each “block” of domain decomposition within FV dycore. For D grid with 26 vertical levels, 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 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

Tracer scaling (up to 200 tracers)



Tracer scaling factor



- 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.

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.
- 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.