Vectorizing the CCSM: SciDAC's Contribution

(and other feats and marvels of software engineering)

Patrick H. Worley
Oak Ridge National Laboratory

Fifth Science Team Meeting
DOE Climate Change Prediction Program
October 19, 2004
Seattle Hilton Hotel
Seattle, Washington

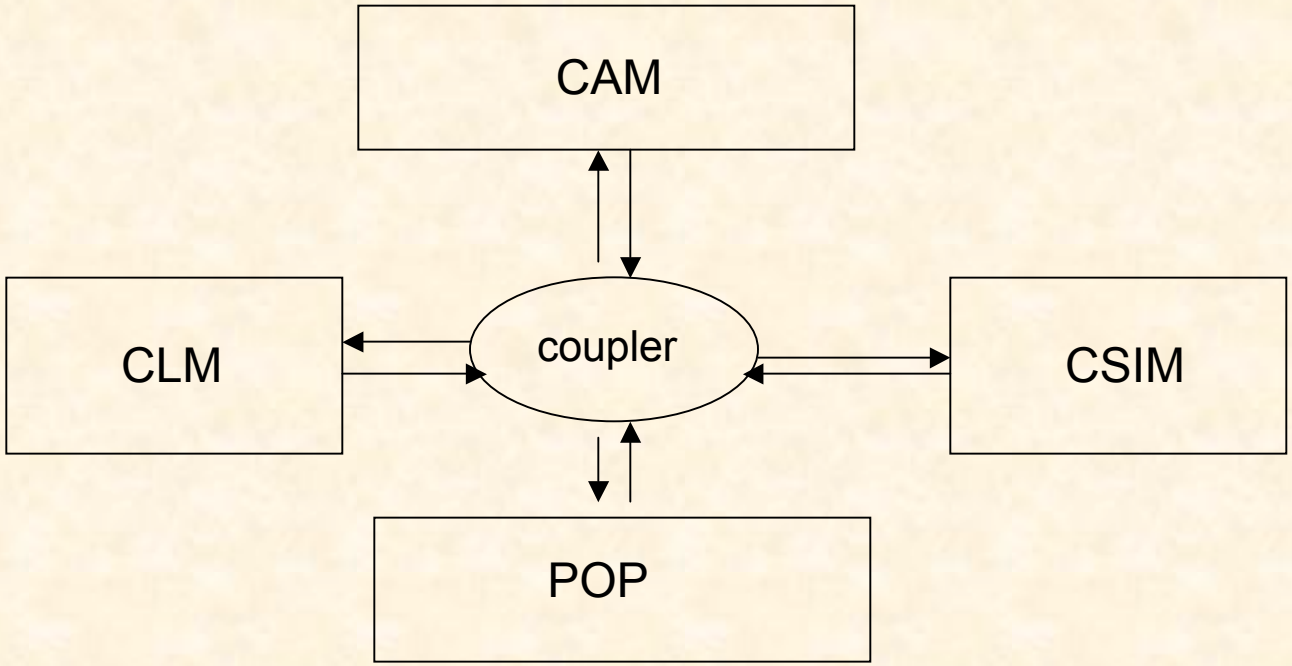
Acknowledgements

- 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.
- These slides have been authored by a contractor of the U.S. Government under contract No. DE-AC05-00OR22725. 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
- Oak Ridge National Laboratory is managed by UT-Battelle, LLC for the United States Department of Energy under Contract No. DE-AC05-00OR22725.

CCSM Overview

- CCSM, the Community Climate System Model is a coupled model for simulating the earth's climate system.
  - Developed at NCAR with significant collaborations with DOE, NASA and the university community
- Components in CCSM3 include
  - Atmospheric Model – CAM 3.0
  - Ocean Model – modified version of POP 1.4.3
  - Sea Ice Model – CSIM5
  - Land Model – CLM2
  - Coupler - CPL6

CCSM Components



Vectorization Process

- For each component model
  - Port to new (vector) system
  - Optimize performance (including vectorization)
  - Merge subset of modifications back into development trunk
  - Validate/Evaluate updated model on all target ("category I") platforms
- For the CCSM
  - Import updated component models
  - Port and optimize scripts and other CCSM infrastructure to new system
  - Verify that CCSM runs correctly in all required configurations
  - Validate climate produced by CCSM
  - Tune configuration to optimize performance on new system

Merge Guidelines and Process

- Cannot degrade performance significantly on other target systems
  - Allowable degradation depends on perceived importance (availability) of given platform for science.
- Cannot alter solution (bit-for-bit) on other platforms
  - Can be relaxed when climate validation needs to be repeated on other platforms anyway.
- For CAM and CLM, solution must be independent of number of processors (i.e., reproducibility).
- Limited amounts of architecture-dependent code allowed (i.e., no large scale #ifdef NEC/CRAY/IBM sections)
  - This is for code maintainability. What is or is not permitted varies among the CCSM working groups.
- Actual merge process consists of making a proposal to the relevant component Change Review Board, followed by some period of negotiation.

Vectorization Teams

Incomplete list of individuals and projects/institutions involved in porting and optimizing CCSM components on the Cray X1 and the Earth Simulator (ES40). Reverse alphabetical order used, and does not reflect value of contributions.

- CAM
  - *SciDAC*: Drake, Ham, Hoffman, Mirin, Worley; *NCAR*: Eaton; *NEC*: Parks, Snyder; *Cray*: Carpenter, Cordery, Levesque
- CICE/CSIM
  - *SciDAC*: Lipscomb; *NCAR*: Schramm; *Fujitsu/NEC*: Chen; *Cray*: Carpenter
- CLM2
  - *SciDAC*: Hoffman; *ORNL*: White; *NCAR*: Vertenstein; *NEC*: Parks; *CRIEPI*: Kitabata;
  - See poster presentation on CLM2 optimizations.

Vectorization Teams (cont.)

- POP
  - *SciDAC*: Jones, Worley; *NCAR*: Norton; *CRIEPI*: Yoshida; *Cray*: Carpenter, Levesque
- MCT/CPL6
  - *SciDAC*: Jacob, Larson, Ong; *NCAR*: Craig; *CRIEPI*: Yoshida, …; *Cray*: Carpenter, Cordery
- CCSM on the X1
  - *SciDAC*: Drake, Ham, Hoffman, Worley; *NCAR*: Carr, Henderson; *Cray*: Carpenter, Cordery

Why Vectorization?

- Availability
  - Earth Simulator: peak 40 TFlop/sec
  - Cray X1 at ORNL: current peak 6.4 TFlop/sec; increasing to over 12 TFlop/sec in 1/05 .
- Performance
  - CCSM IPCC runs >4 times faster on Earth Simulator

    (14 yrs/day on 168 processors)

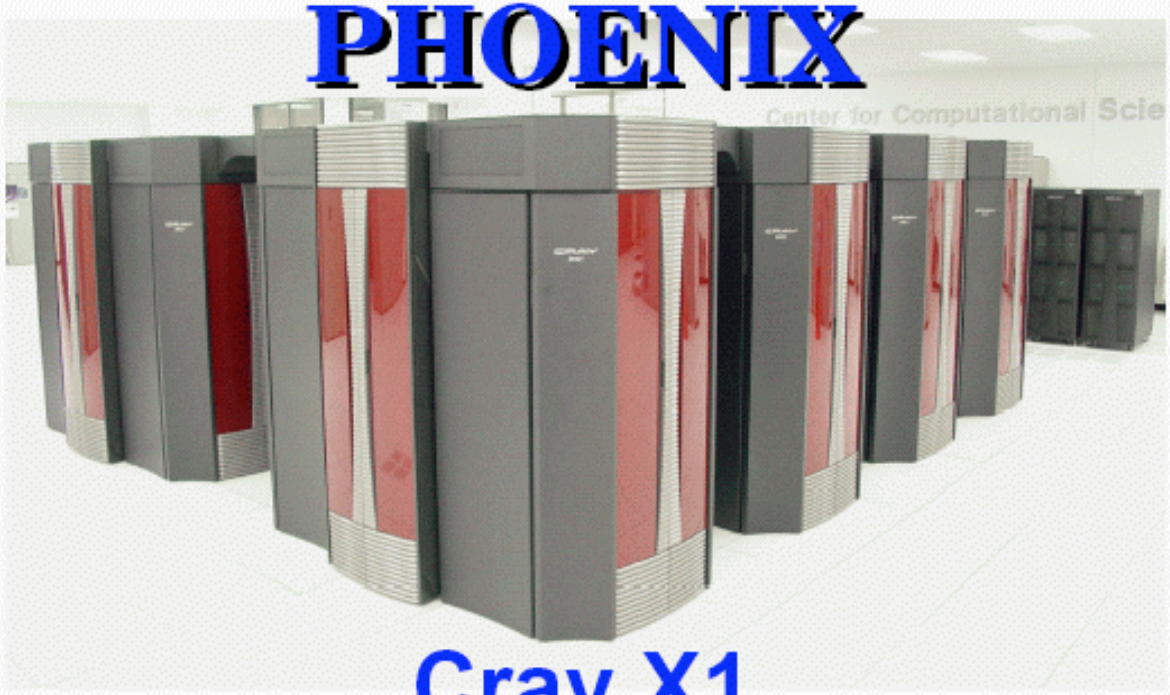
    than on p690 cluster

    (2 yrs/day on 96 processors)
  - See next few slides.

Earth Simulator

Cray X1

PHOENIX



Cray X1

- 128 SMP nodes
- 4 Multi−Streaming Processors (MSPs) per node
- 4 Single Streaming Processors (SSPs) per MSP
- Two 32−stage, 64−bit wide vector units running at 800 MHz and one 2−way superscalar unit running at 400 MHz per SSP
- 2 MB E−cache per MSP
- 16 GB of memory per node

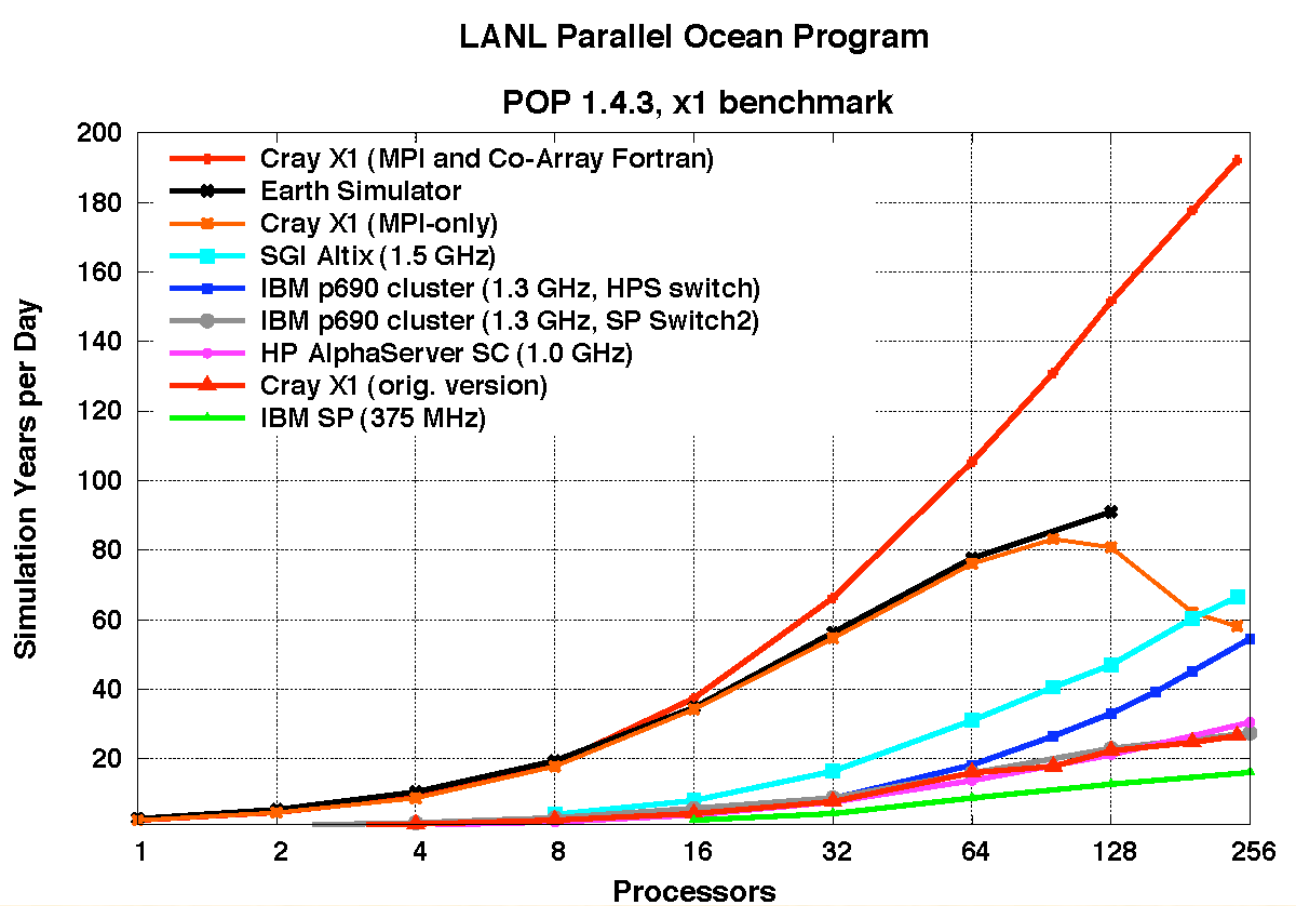
512 processors (MSPs), 2048 GB of memory, and 6400 GFlop/s peak

POP Simulation Rate: x1 benchmark

Comparing performance and scaling across platforms.

- Earth Simulator results courtesy of Dr. Y. Yoshida of the Central Research Institute of Electric Power Industry
- IBM SP results courtesy of Dr. T. Mohan of Lawrence Berkeley National Laboratory

using NEC port of POP on Earth Simulator and on X1, not all of which changes have been accepted into CCSM.

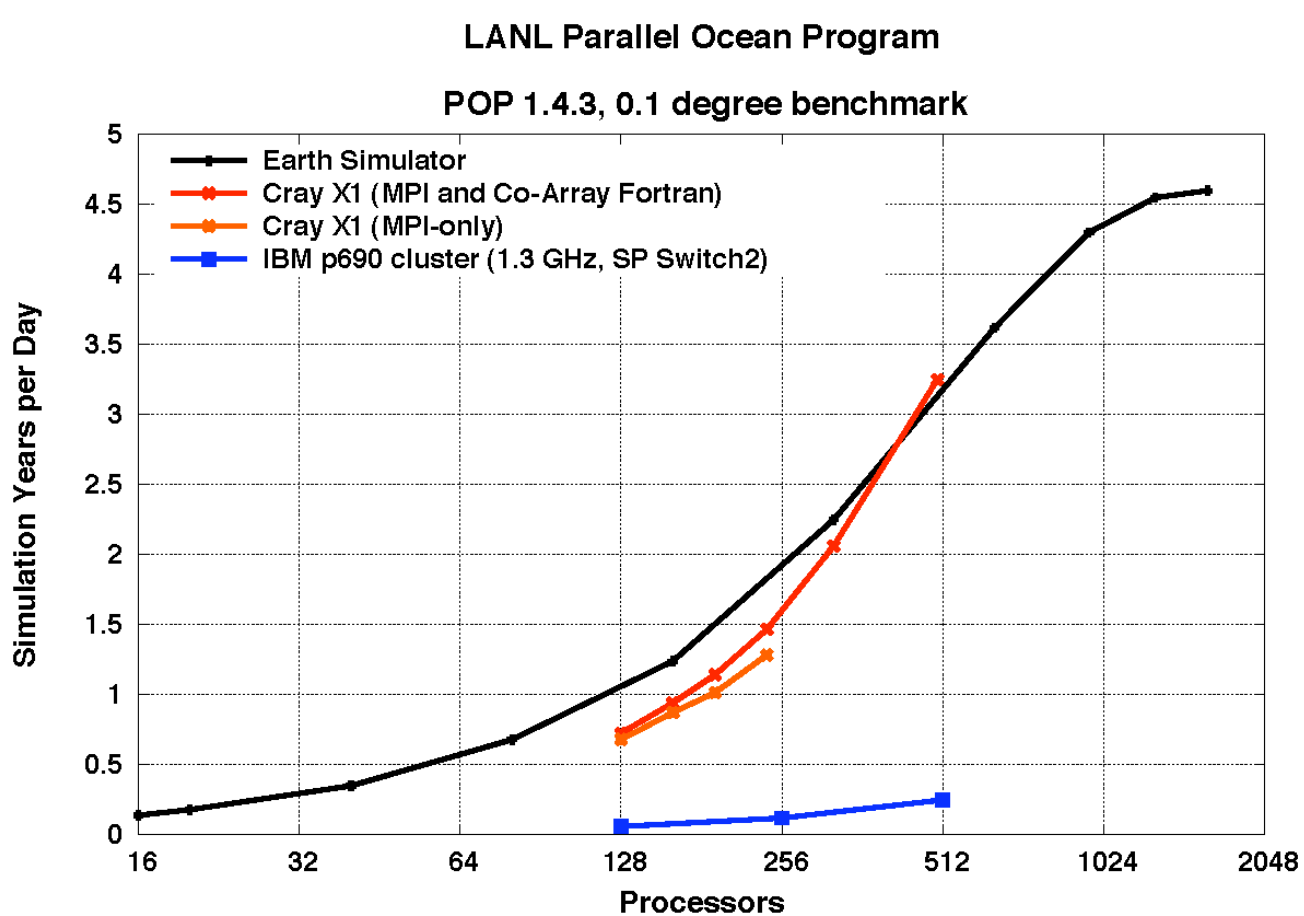


POP Simulation Rate: 0.1 benchmark

Comparing performance and scaling across platforms for a 0.1 degree benchmark problem (3200 x 2400 x 40 grid).

- Earth Simulator results courtesy of Dr. Y. Yoshida (collected Spring, 2003)
- X1 results collected Summer, 2004

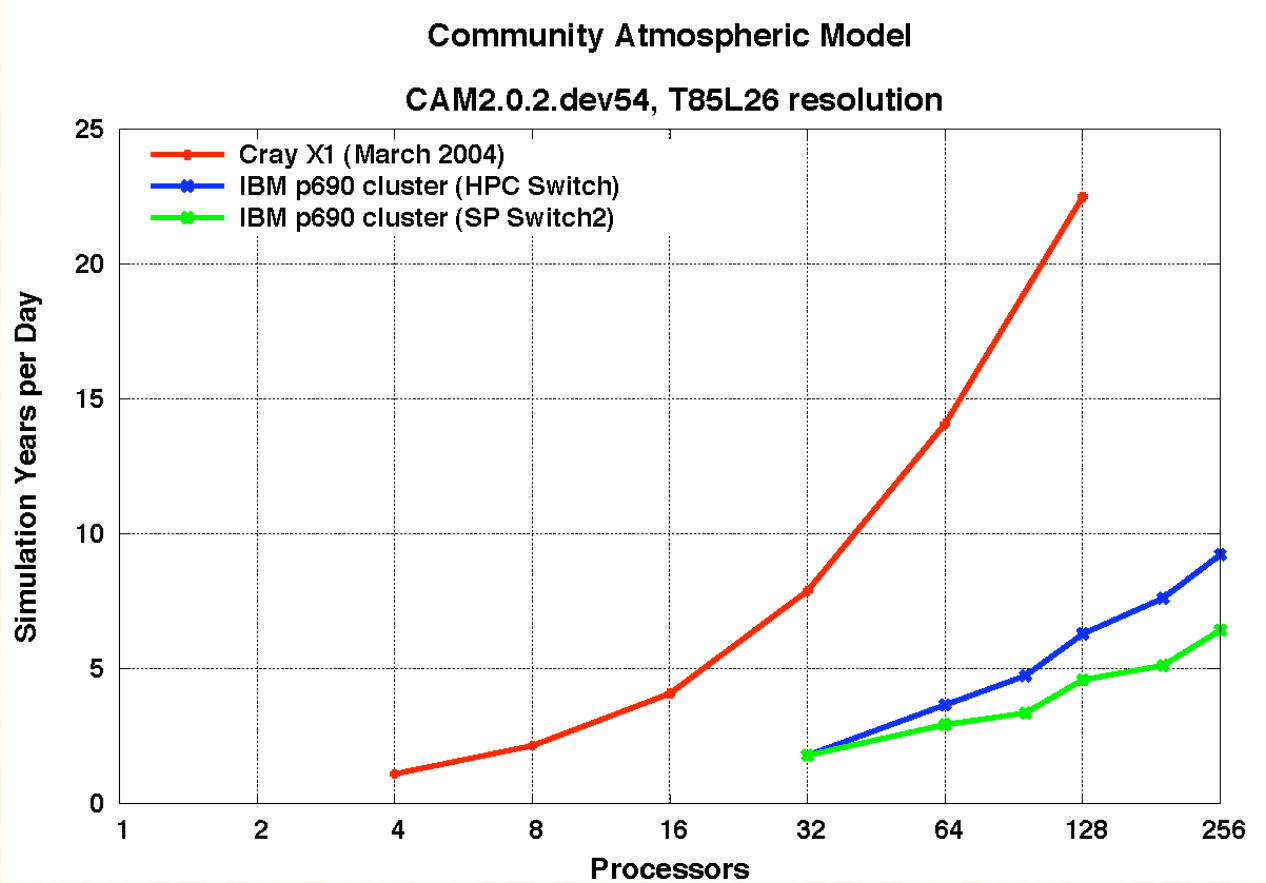
X1 performance is good compared to IBM (>13X), but lags behind that of the Earth Simulator until reach 512 processors. Larger problem size changes importance of Cray-specific vector optimizations.



CAM Simulation Rate: T85L26 benchmark

Comparing performance and scaling across platforms for a T85L26 benchmark problem (256 x 128 x 26 grid).

X1 performance is good compared to IBM (> 3.6X for 128 processors), with additional optimizations planned.



Approximate Timelines

=> December 2003:

- Component model vectorization

=> April, 2004:

- Merge of vector versions into development branch, including basic support for the X1
- CAM/CLM2 standalone model (spectral Eulerian dycore) validated on the Earth Simulator and X1

=> June, 2004:

- CCSM validated on Earth Simulator and achieves required percentage of vectorization
- CCSM3 released, including basic support for X1

=> October, 2004:

- Verification on X1 almost complete
- CCSM3 validated on the X1

Remaining X1 Verification Issues

- Model requires a particular (old) version of system software (compilers and MPI libraries).
- Model time in POP and CSIM4 suddenly becomes corrupted after approximately 10 simulation years.
- Performance variability is unexpectably high.
- Answers change slightly (round-off level) when using load balancing.
- Performance timers in coupler are broken.
- Need to harden run scripts for ORNL environment.
- Job log file contains output from all coupler processes, not just the master.
- Long term archiving script is not working correctly.
- Ice model validation may need to be revisited.

Immediate X1 Tasks

- Complete CCSM3 verification.
- Optimize CCSM3 configuration.
- Revisit component optimizations.
- Complete CAM FV and SLD optimization and validation.
- Revisit component optimizations for high resolution runs and for new components (e.g., POP 2.0).

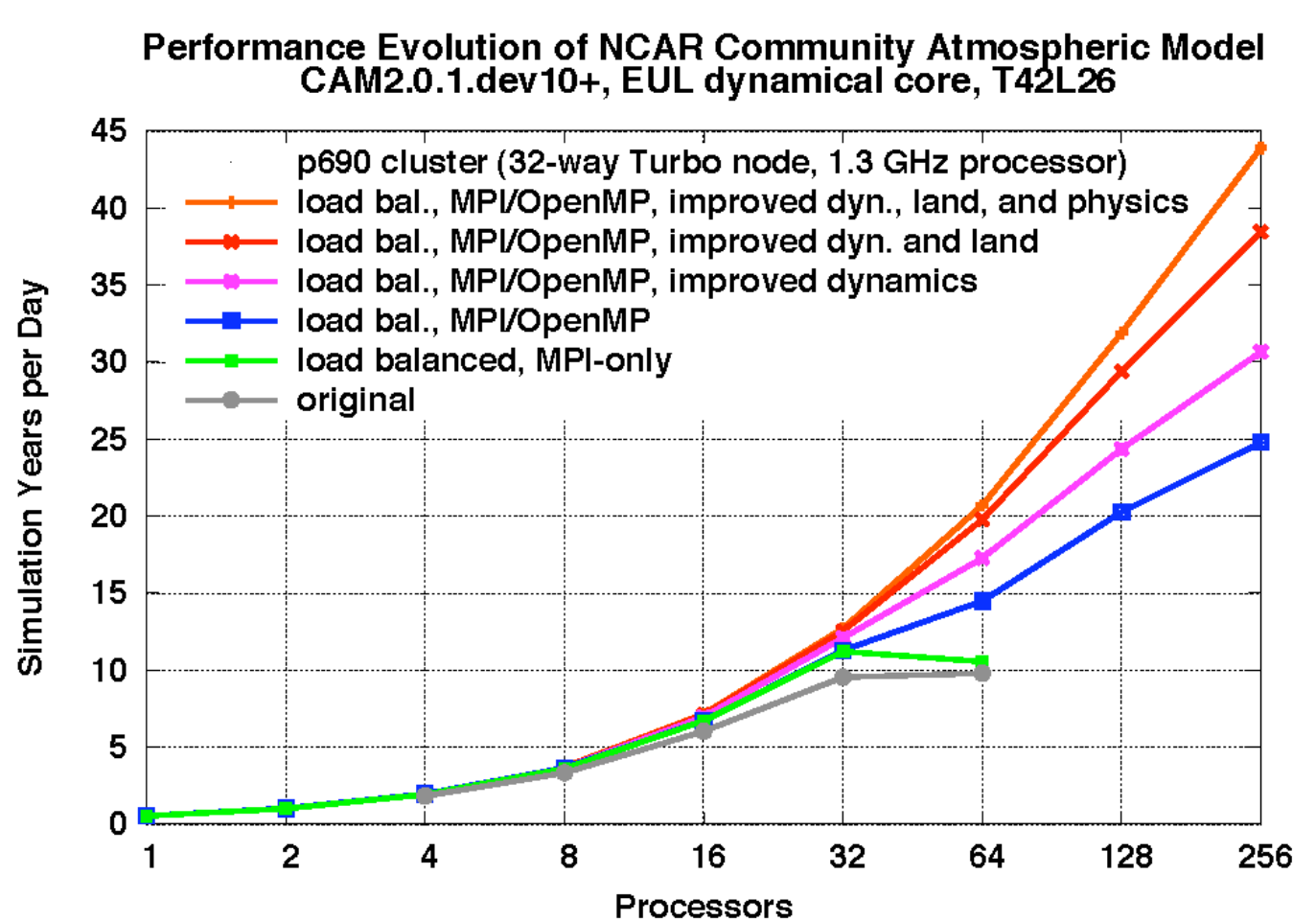
Impact of SE on Performance Portability

- Run- and compile-time choice of inner loop lengths can be used to optimize for vector systems (longer inner loops) and nonvector systems (shorter inner loops for cache blocking).
- Runtime load balancing options:
  - minimal load balancing / no communication overhead
  - better load balancing / higher communication overhead
  - best load balancing / highest communication overhead

  can be used to optimize performance on systems with higher or lower performing communication networks, with special options designed for SMP clusters.
- Runtime choice of MPI-1 collectives, MPI-1 two-sided, MPI-2 one-sided, and Co-Array Fortran implementations of interprocess communication.
- Hybrid MPI/OpenMP with load-time choice of number of MPI processes and OpenMP threads.

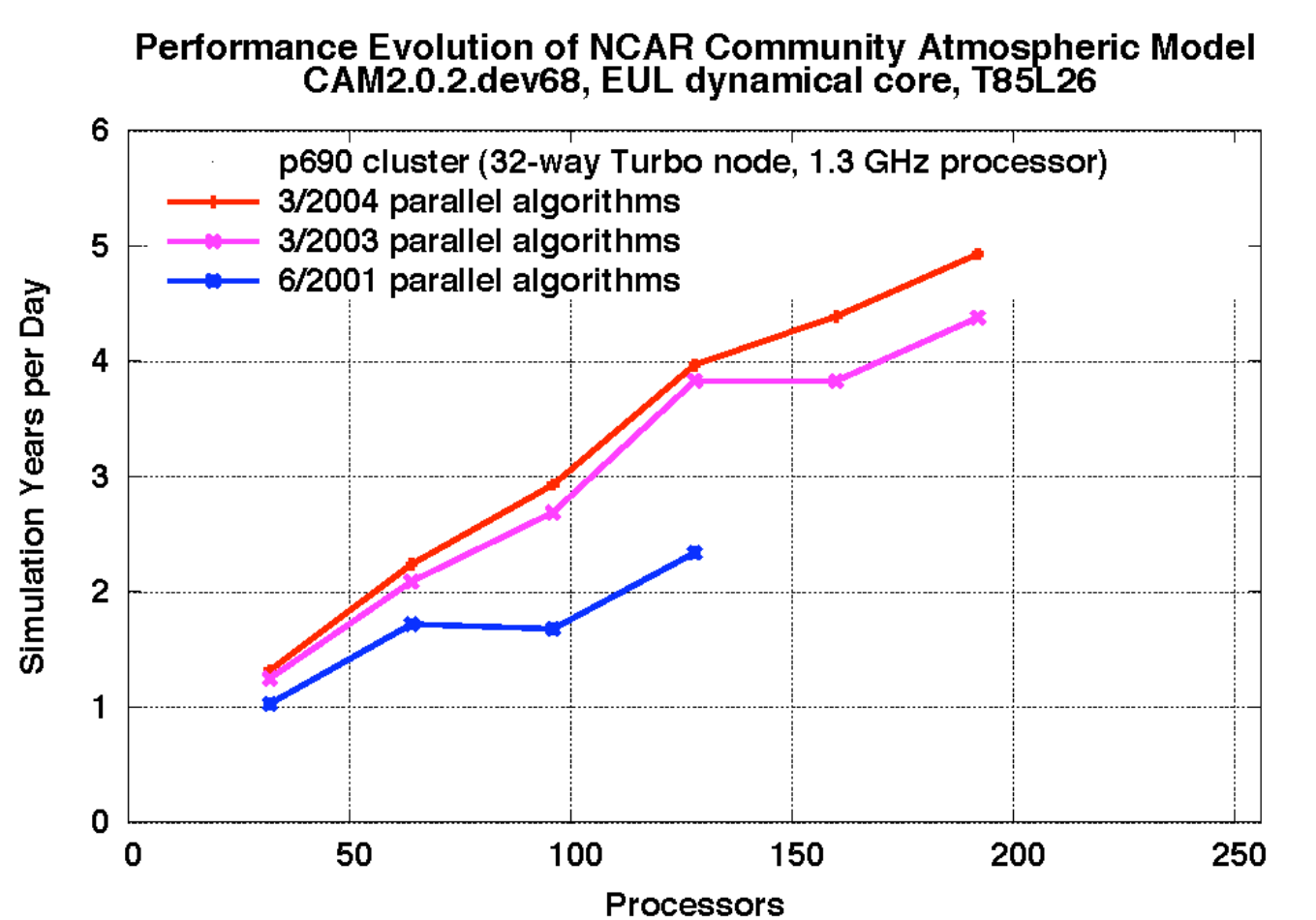
Evolution of Atmospheric Model Performance as of April 2003

Last year's results for 64x128x26 grid and 3 advected fields. Since then, resolution has increased to 128x256x26, 11 advected fields and parameterizations for many new processes.



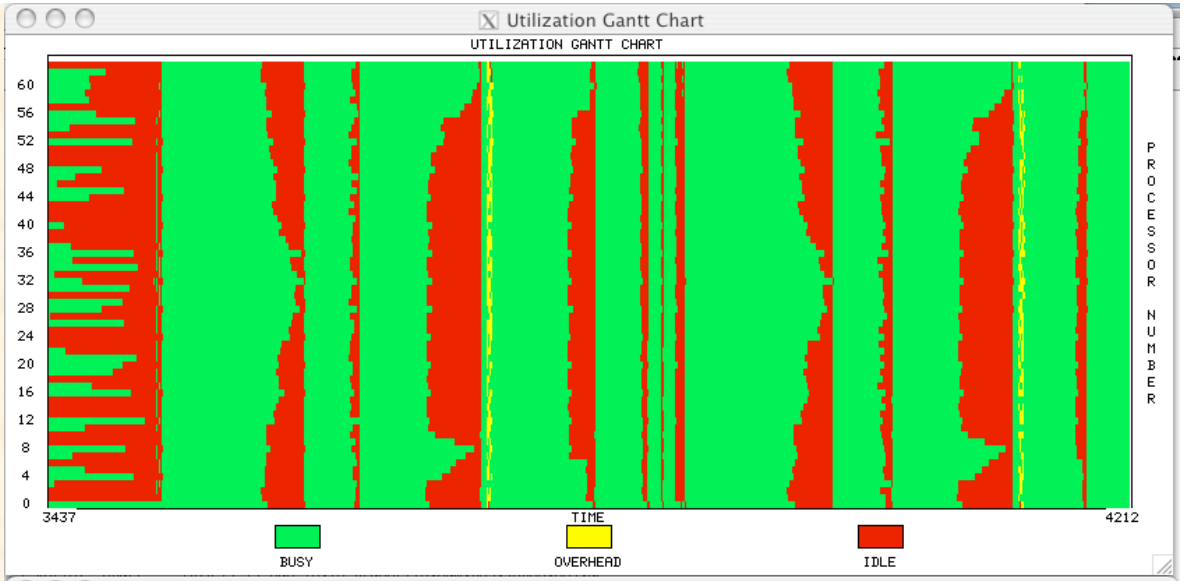
Evolution of Atmospheric Model Performance as of March 2004

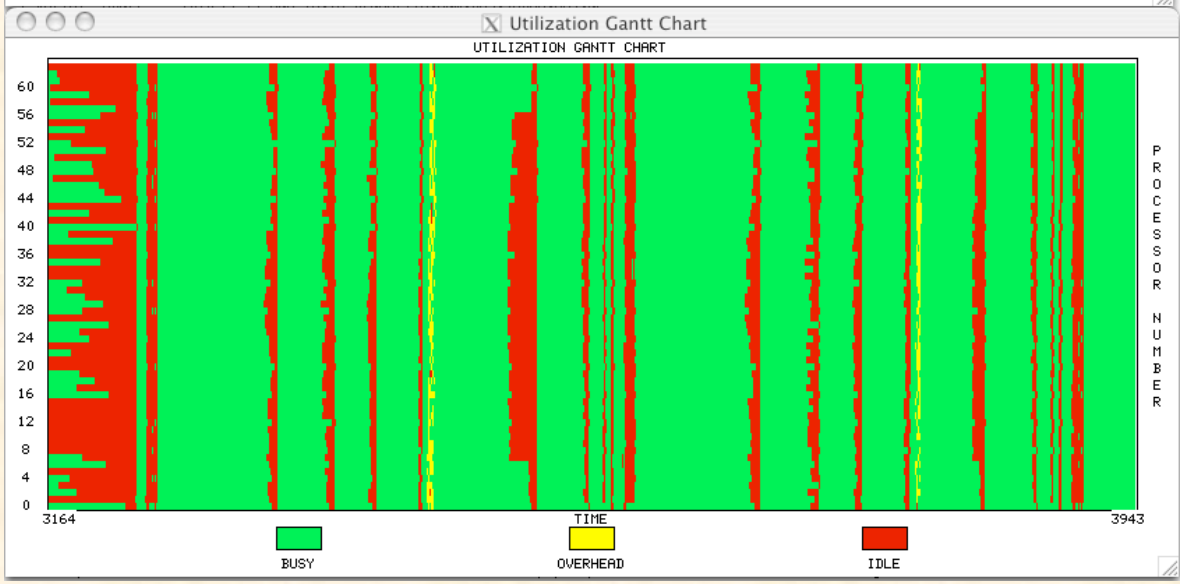
New resolution decreased simulation rate significantly. Previous optimizations continue to be important, but new opportunities for optimization arise as science in code evolves and as code is ported to new platforms.



Performance Impact of Load Balancing

Compute (Busy) / Communicate (Overhead and Idle) Gannt charts without and with load balancing. Load balancing is (unexpectedly) even more important on the X1 than on other systems. Because of the good interprocessor communication performance, the optimal static load balancing was the best choice.





Conclusions

- CCSM3 ported, optimized, and validated on both Cray X1 and Earth Simulator. Verification (bug fixes) continues on the X1.
- Prior performance portability options have proven very important for successful vector ports, enabling good vector performance without degrading performance on nonvector systems.
- Performance optimization continuing on X1, focusing in particular on high(er) resolution models.