

Benchmarking a Parallel Coupled Model

J. Walter Larson, Robert L. Jacob, and Everest T. Ong

Mathematics and Computer Science Division, Argonne National Laboratory

Anthony Craig, Brian Kauffman, and Thomas Bettge

Climate and Global Dynamics Division, National Center for Atmospheric Research

Yoshikatsu Yoshida

Central Research Institute of Electrical Power Industry

Junichiro Ueno, Hidemi Komatsu, and Shin-ichi Ichikawa

Computational Science and Engineering Center, Fujitsu Limited

Clifford Chen

Fujitsu America

Patrick H. Worley

Computer Science and Mathematics Division, Oak Ridge National Laboratory

Abstract

The need for high throughput in parallel coupled climate models gives impetus to the requirement for detailed performance measurements to inform configuration of these models for production. We have devoted considerable attention to the problem of benchmarking a particular parallel coupled model--the Community Climate System Model (CCSM). We will define parallel coupled models, and describe in general terms various aspects of coupled model architecture. We then present work in progress to assemble performance measurements for the various components of CCSM, and describe how these results can be applied to allocate resources to CCSM's major subsystems. We then describe manual instrumentation using MPE, and present results of an MPE-based performance study. We conclude by summarizing current results and outline an agenda for future work.

What is a Parallel Coupled Model?

The dramatic increases in computational power parallel computing offers is enabling researchers to shift focus from the simulation of individual subsystems in isolation toward more realistic simulations comprising numerous mutually interacting subsystems. We call these more complex systems *parallel coupled models*. Parallel coupled models present numerous challenges in *parallel coupling*, including—but not limited to—parallel data transfer, intermesh interpolation, variable transformations, time synchronization of exchanged data, and merging of multiple data streams, all while maximizing overall performance.

A climate system model [1-3] is an excellent example of a coupled model, typically comprising atmosphere and ocean general circulation models (GCMs), a dynamic-thermodynamic sea ice model, a land-surface model, and a river transport model. Interactions between these component models are often managed by a special component called a *flux coupler* [4].

Parallel Coupled Model Architecture

Parallel Coupled models may be classified using the following characteristics

- **Resource allocation** (i.e., MPI processes)
 - A single shared pool of processes
 - Distinct pools of processes, one pool per component
 - Clusters of components, each of which is assigned a distinct pool of processes
 - Overlapping pools of processes
- **Component scheduling**
 - Sequential (event-loop)
 - Concurrent (all components running simultaneously)
- **Number of executable images**
 - Single
 - Multiple

Examples

- CCSM
 - Multiple executables (atmosphere, ocean, sea ice, land, and flux coupler)
 - Concurrent component scheduling
 - Each component resides on a distinct pool of MPI processes
- Parallel Climate Model
 - Single executable
 - Sequential component scheduling
 - All components share the same pool of MPI processes
- Fast Ocean-Atmosphere Model (FOAM)
 - Single Executable
 - Clusters of components, each cluster assigned a distinct pool of MPI processes
 - Each cluster scheduled concurrently, with sequential component execution within each cluster

The Parallel Coupled Model Benchmarking Problem

What we wish to measure and maximize is the overall *throughput* of the coupled model

- Relatively straightforward for a single executable, sequentially scheduled parallel coupled models like PCM
 - Scaling and load balance of individual component models
 - Costs of coupling code
- More complex for concurrently scheduled parallel coupled models
 - Scaling and load balance of each component and the coupler
 - Intercomponent communications costs
 - Delays due to intercomponent data dependencies

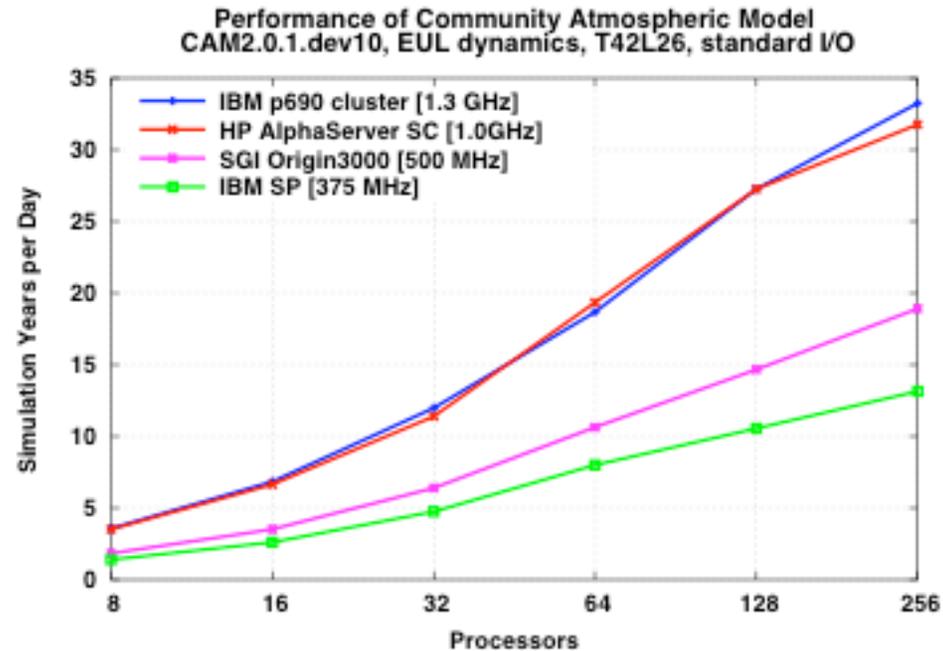
Benchmarking CCSM

- Component model performance measurements and scaling
 - Community Atmosphere Model (CAM)
 - Parallel Ocean Program (POP)
 - Common Land Model (CLM)
- Coupling benchmarks
 - MCT interpolation benchmark
- Measurement of overall model performance
 - MPE instrumentation
 - End-to-end throughput

CCSM Production Benchmark

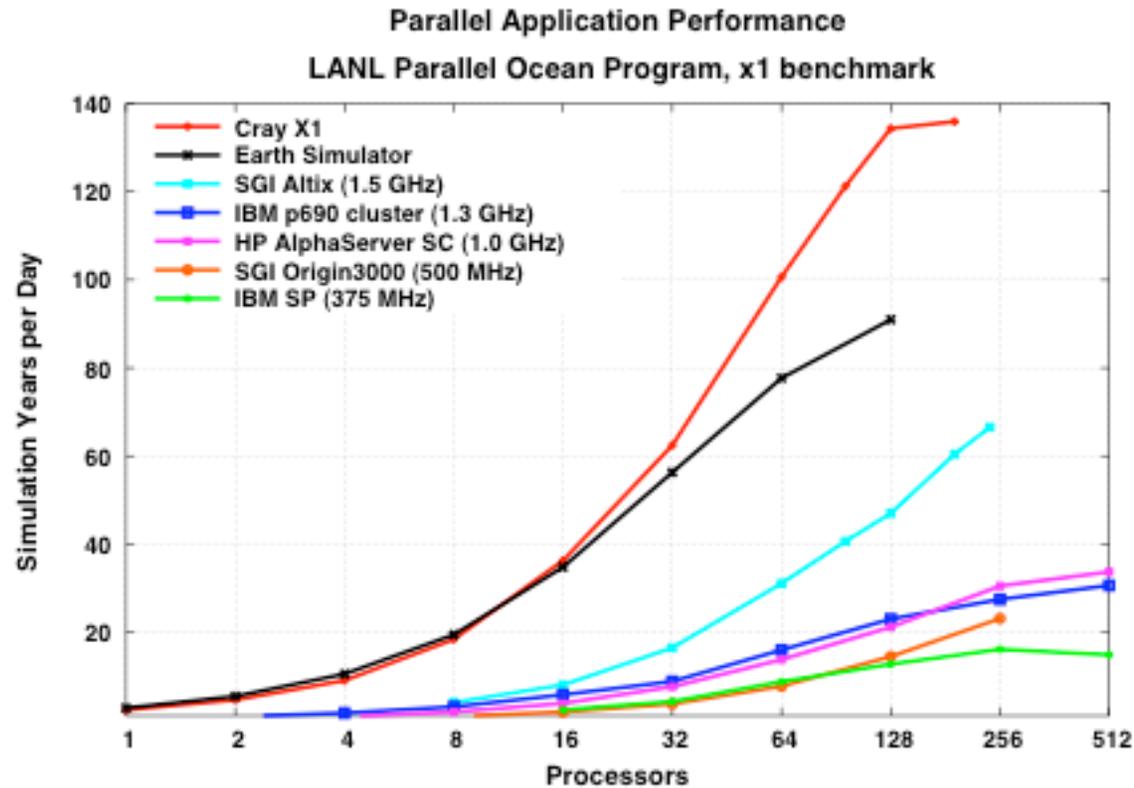
- Forward integration of relatively coarse models
 - Atmosphere and land on T42 grid (128 longitudes x 64 latitudes x 26 vertical levels)
 - Ocean and sea ice - 1 degree (320x384, L40)
- Finite difference and spectral, explicit and implicit methods, vertical physics, global sums, nearest neighbor communication
- I/O not a bottleneck (5 GB output / simulated year)
- Restart capability (750 MB)

Atmosphere Model Performance

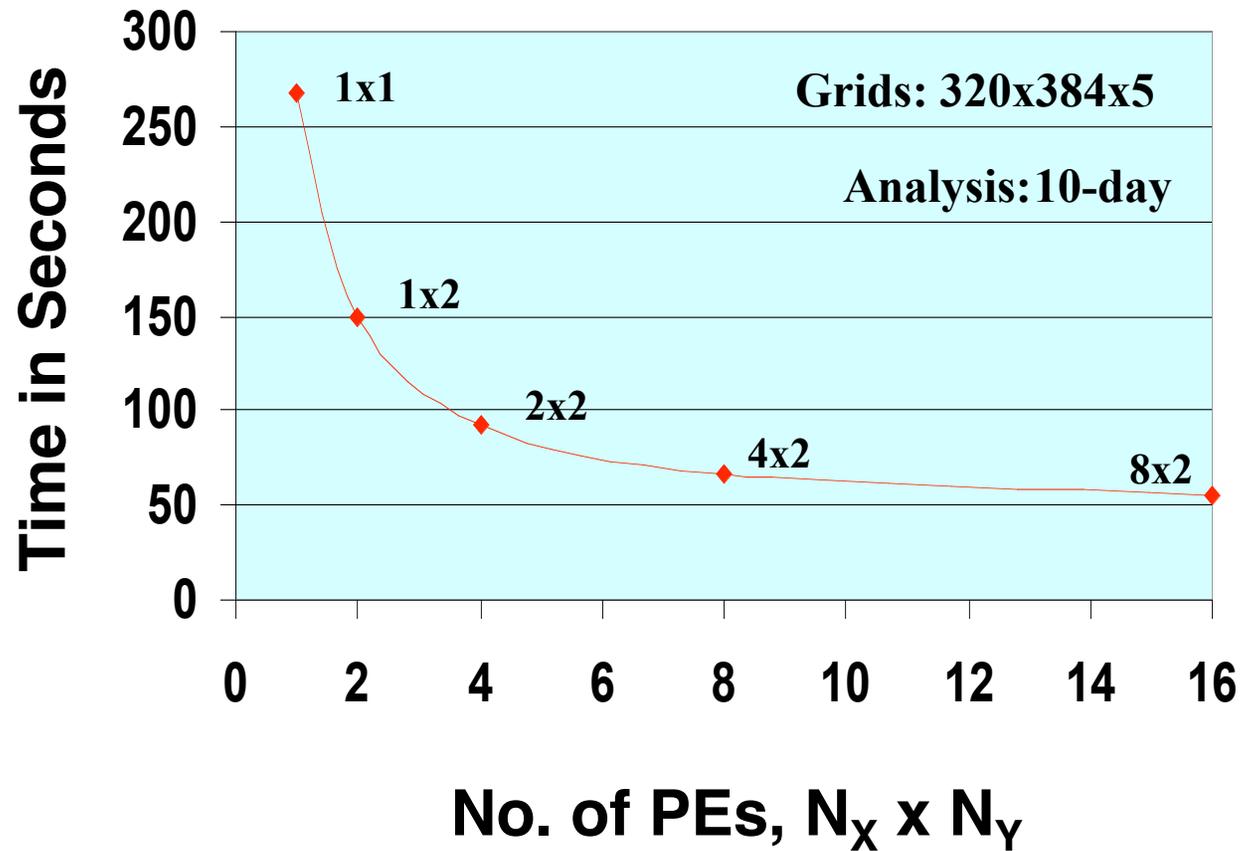


Ocean Model Performance and Scaling

Courtesy of PW Jones, PH Worley,
Y Yoshida, JB White III, J Levesque

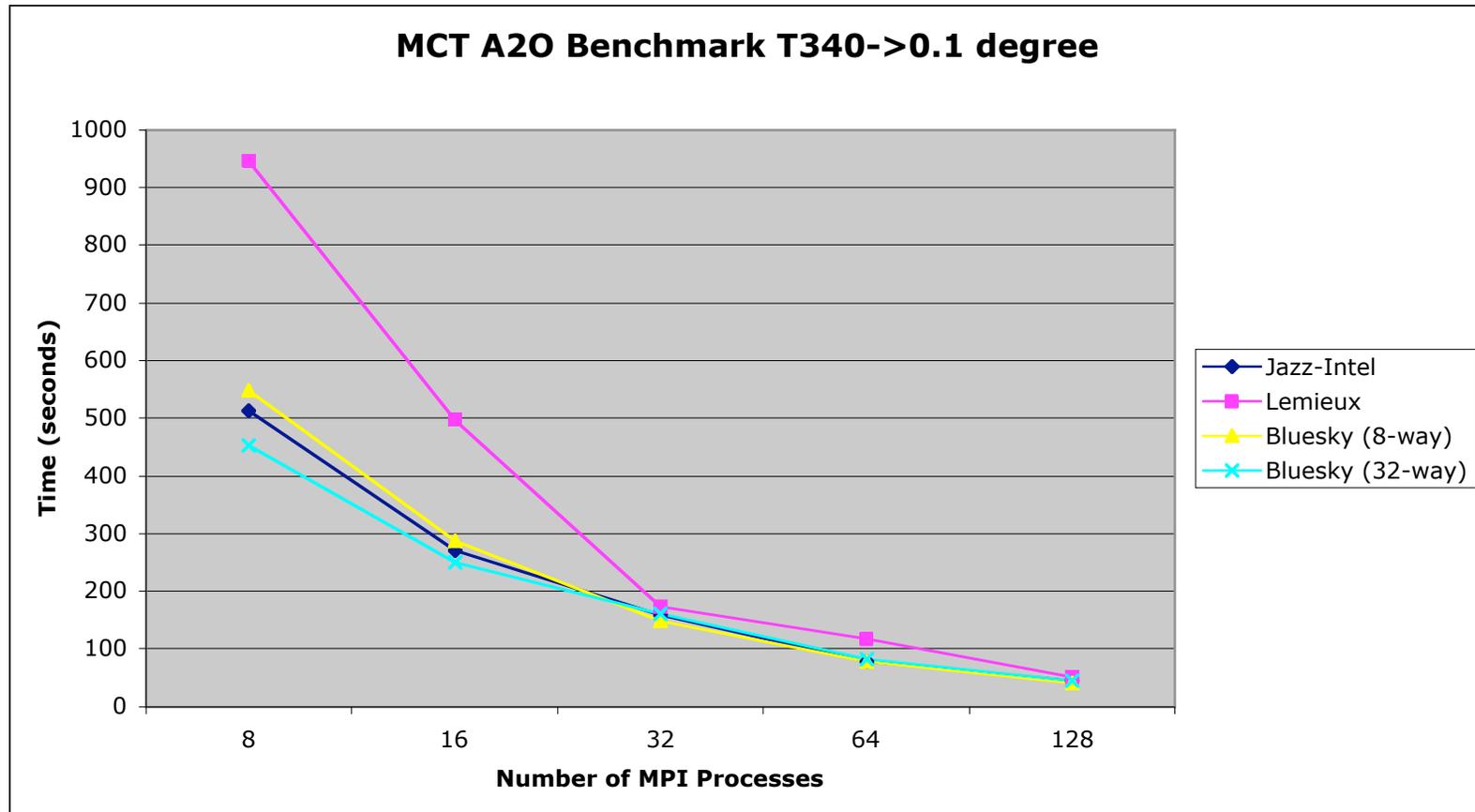


Standalone CICE at Earth Simulator



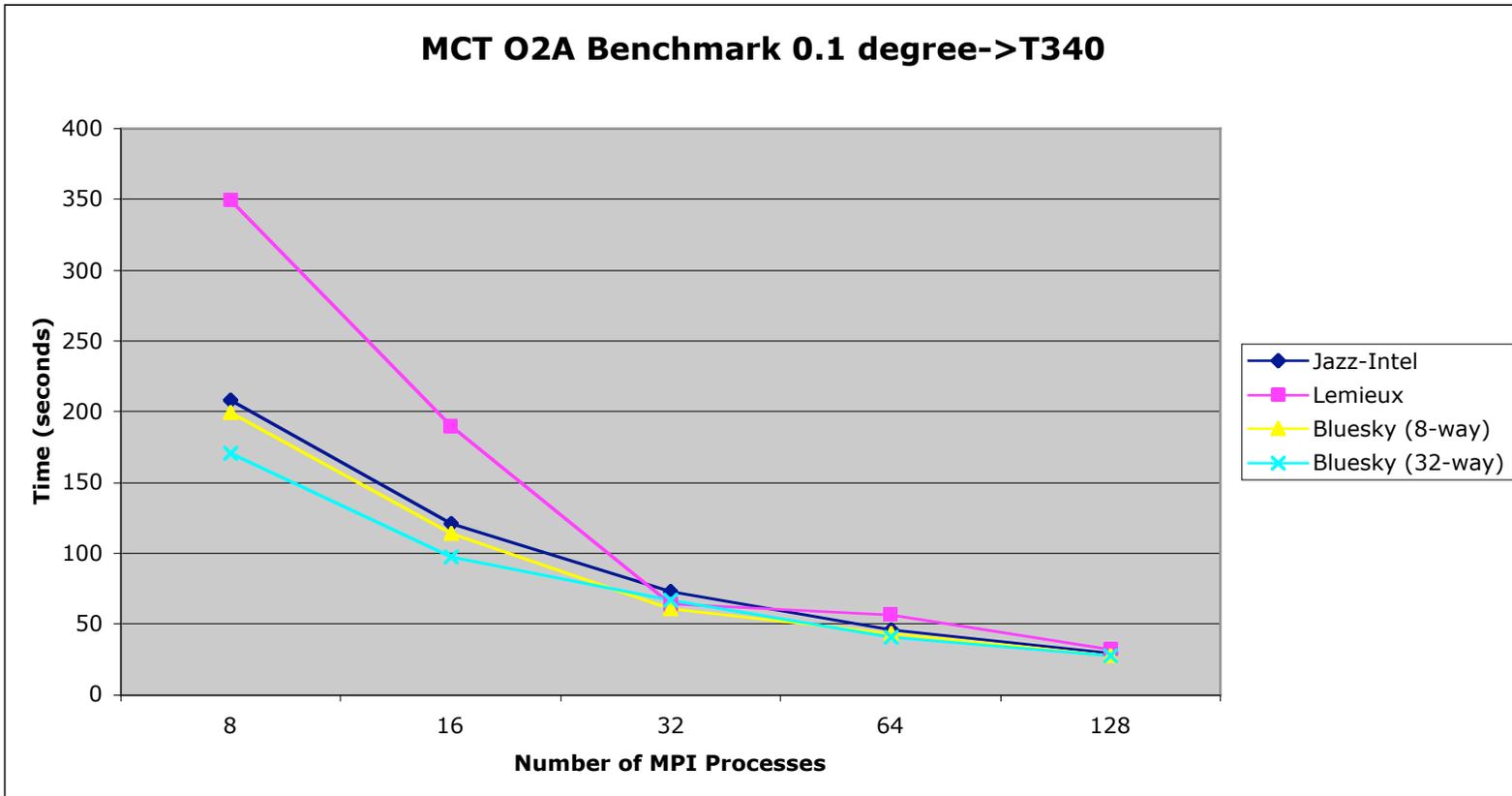
Source Code: As of Oct. 15, 2003

Intergrid Interpolation - A20



Measured across ten model days

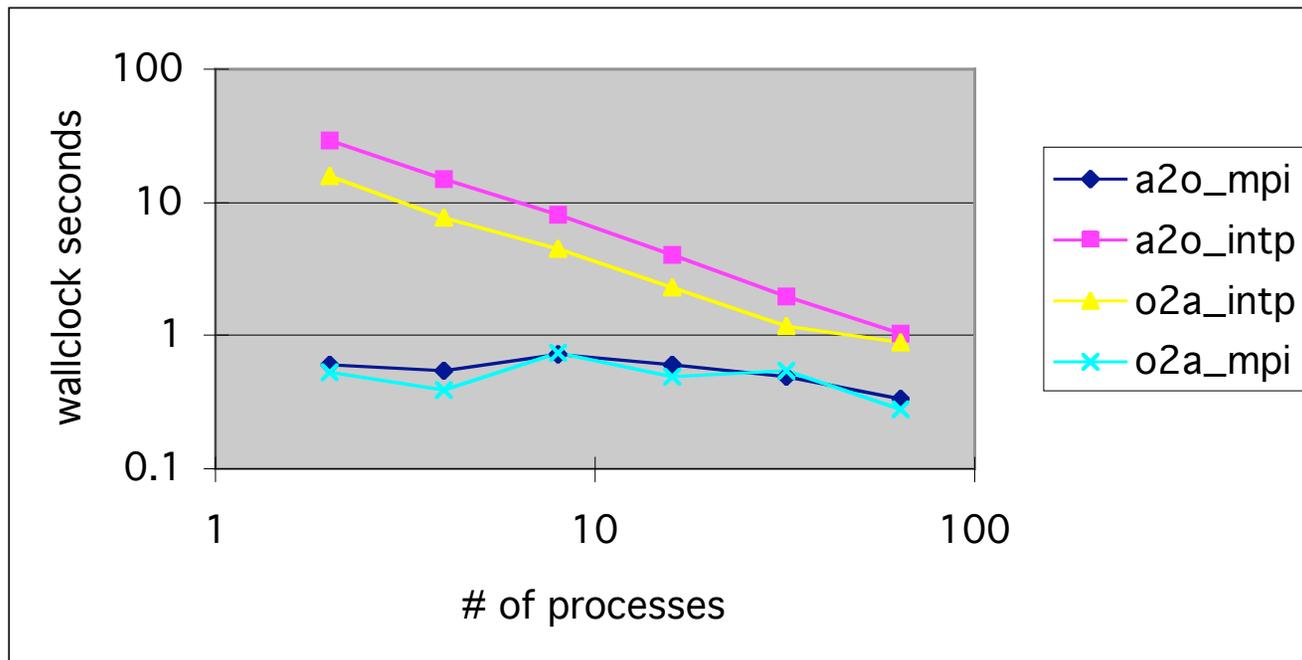
Intergrid Interpolation - O2A



Measured across ten model days

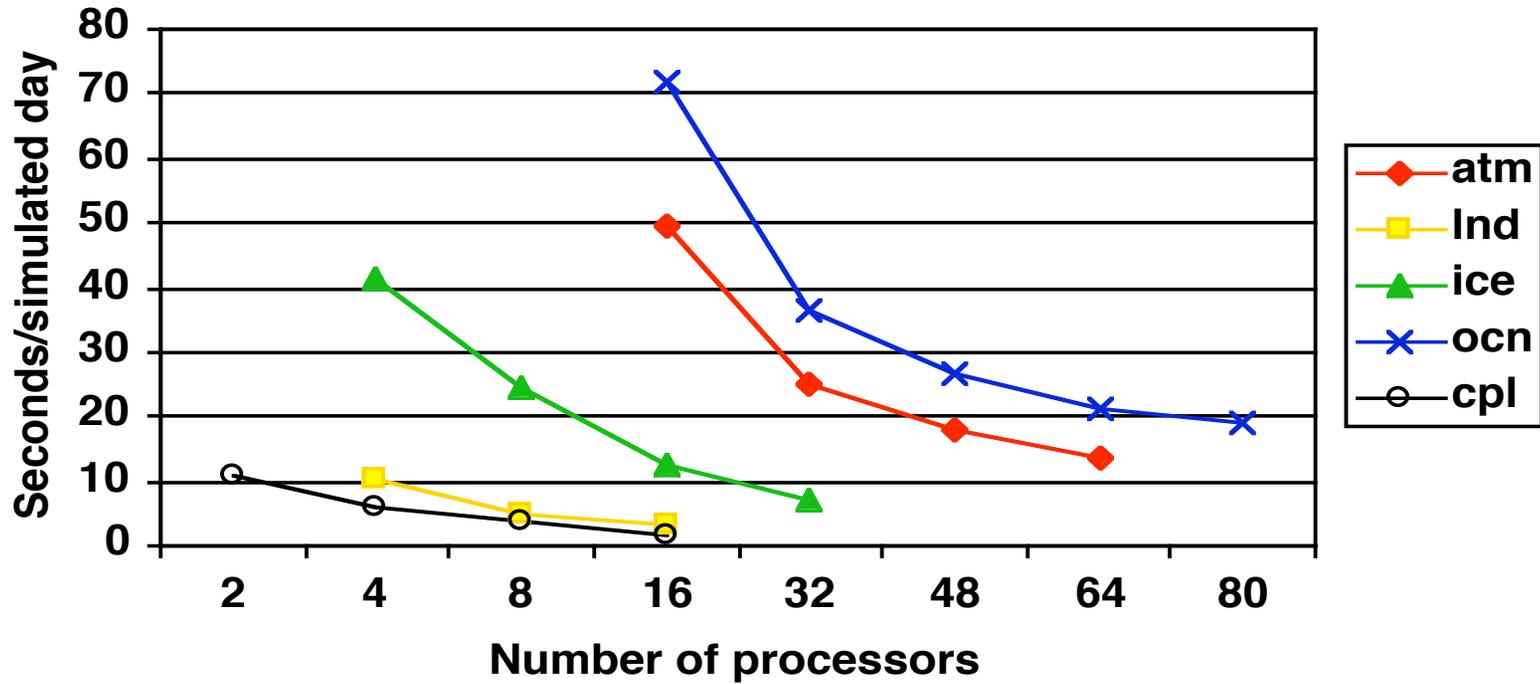
High-Resolution MCT Benchmark on the Earth Simulator

Plotted below is the scaling of the MCT interpolation benchmark for a T340 atmosphere and 0.1 degree POP ocean, measured over the course of one model day.

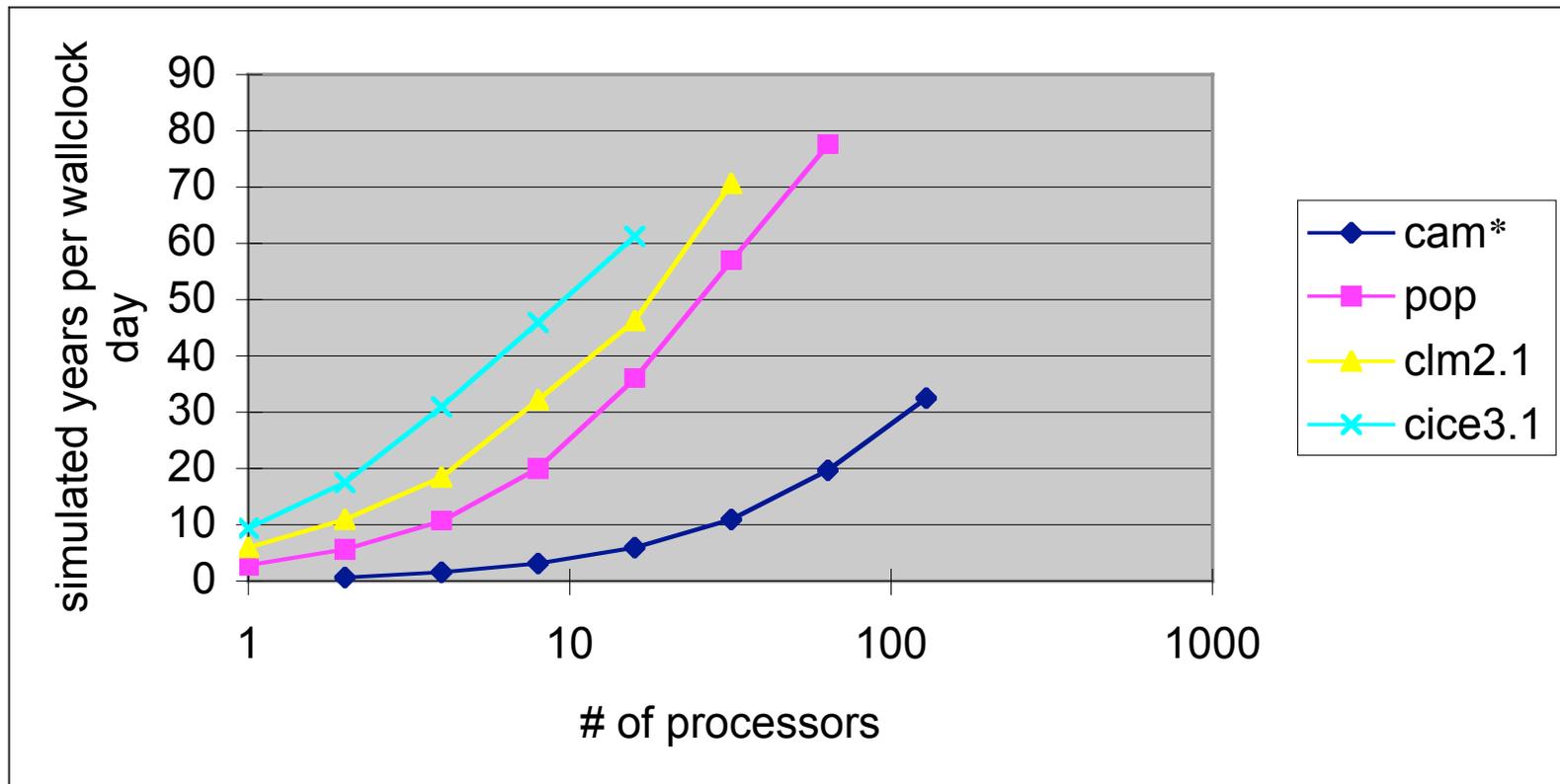


CCSM Component Scaling

CCSM2_2_beta08
T42_gx1v3
IBM Power4, bluesky



CCSM Component Scaling on the Earth Simulator



* Atmosphere performance has been improved beyond the figures shown above. On 64 PEs, CAM achieves 69 model years/day.

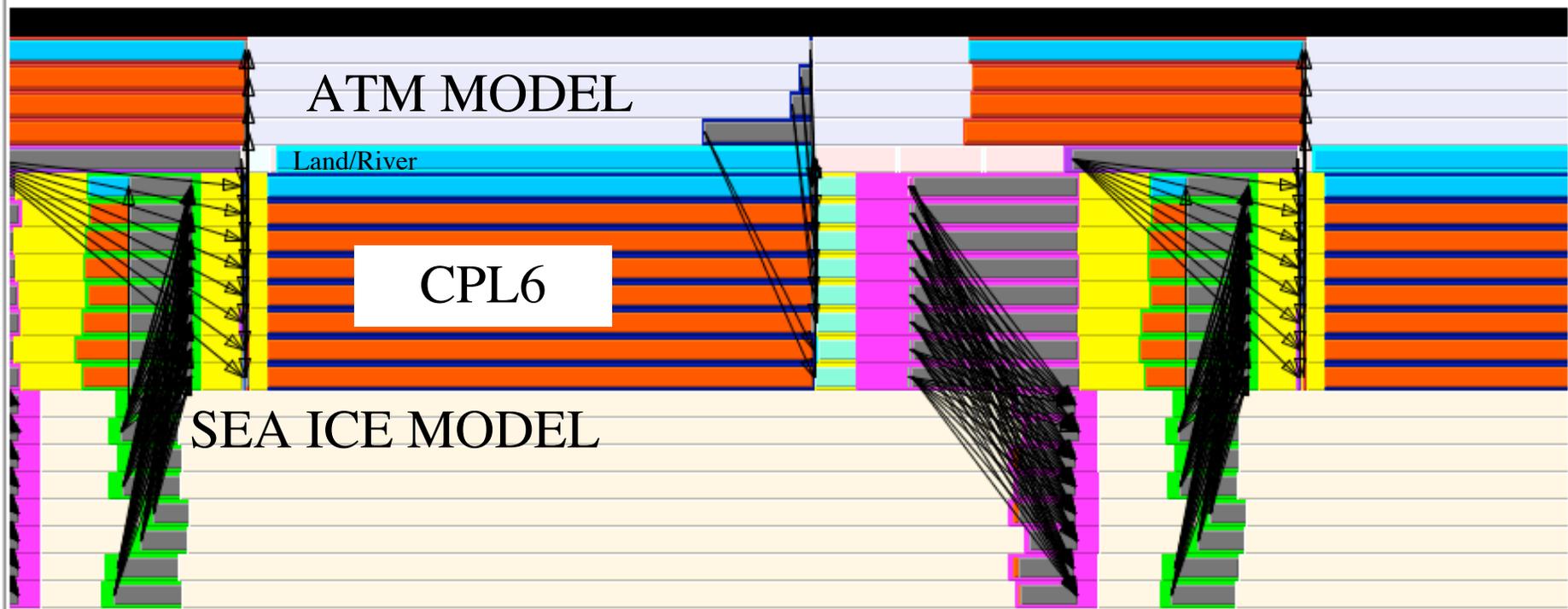
Instrumentation of CCSM with MPE

MPE - MultiProcessing Environment library

- Comes **free** with MPICH. Can be built on top of any vendor MPI implementation.
- Allows the user to automatically visualize all MPI calls and also add custom timers for sections of code. **We have added custom MPE timers to the coupler-related sections of CCSM.**
- MPE outputs a single binary log file containing information from all MPI processes. Viewed with GUI program called **Jumpshot**
- MPE is developed by Rusty Lusk, Bill Gropp and Anthony Chan at Argonne National Laboratory.
- See www.mcs.anl.gov/perfvis for more information.

Zoom Level	Global Min Time	View Init Time	Zoom Focus Time	View Final Time	Global Max Time	Time Per Pixel
8	0.0856329799	217.7749466841	218.726216431	219.6774856593	409.0244104266	0.0021843159

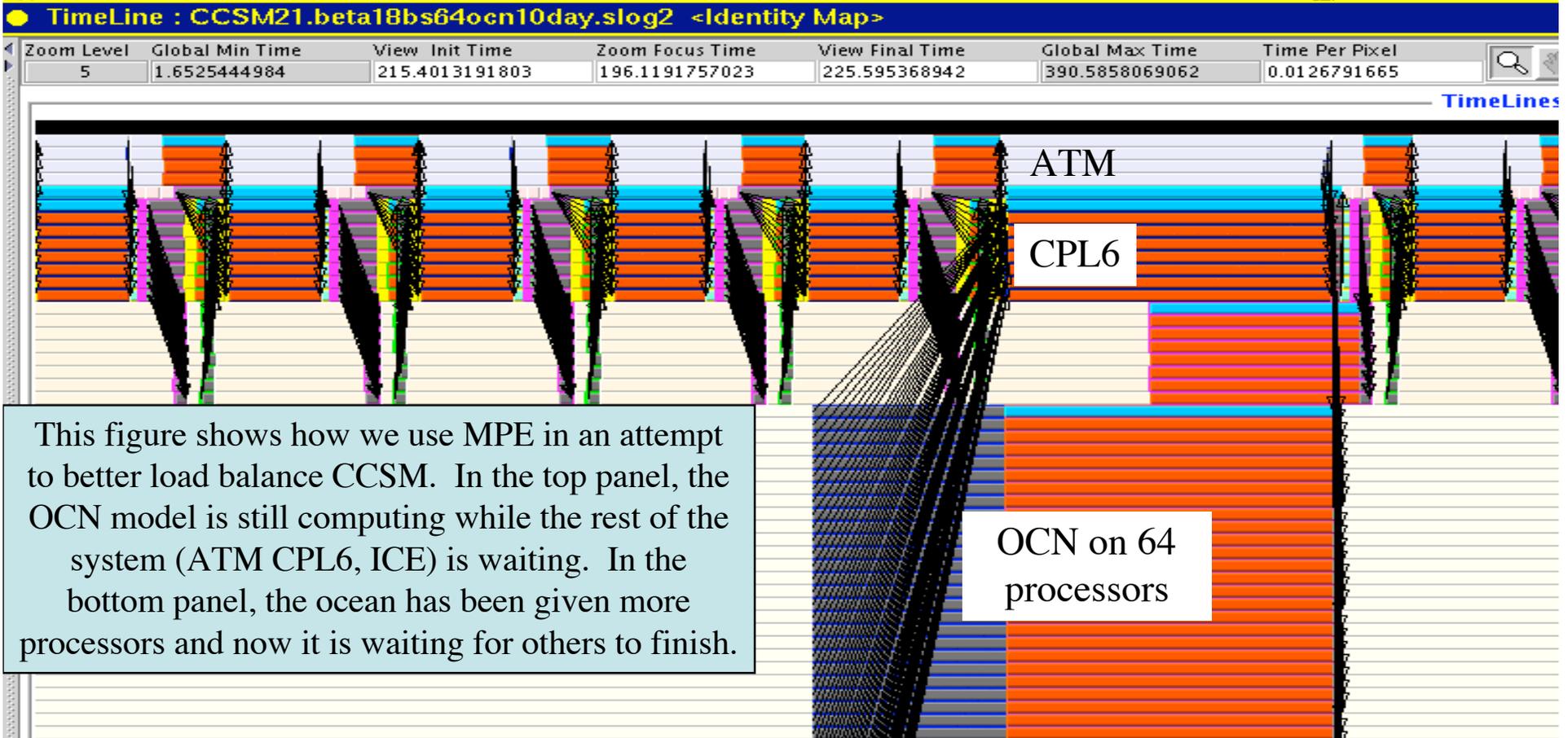
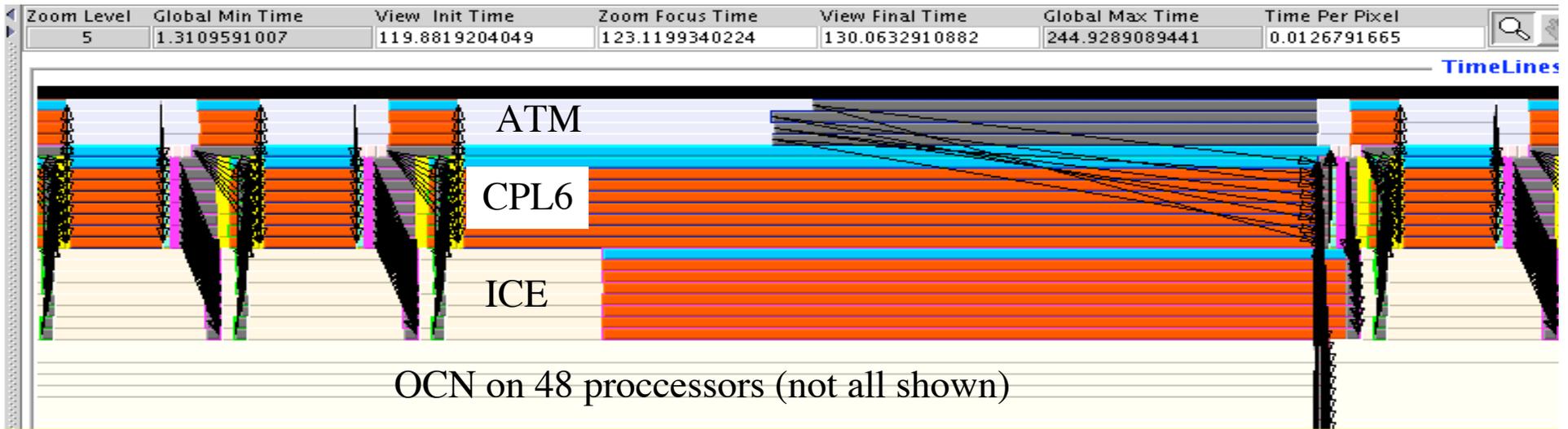
TimeLines



OCEAN MODEL

This figure (from Jumpshot) shows the interaction between components within CCSM for one simulated hour. Each horizontal bar is an MPI processes. Wait time (blue and red) is caused by data dependencies between component models. Computation is shown in white/pastel colors. In some places, the atmosphere, sea ice, ocean and land models are computing simultaneously as intended. Messages passed between coupler and components are indicated by arrows.





This figure shows how we use MPE in an attempt to better load balance CCSM. In the top panel, the OCN model is still computing while the rest of the system (ATM CPL6, ICE) is waiting. In the bottom panel, the ocean has been given more processors and now it is waiting for others to finish.

CCSM Load Balance and Intercomponent Interactions

CCSM2_2_beta08
IBM Power4, bluesky

processors

64 ocn

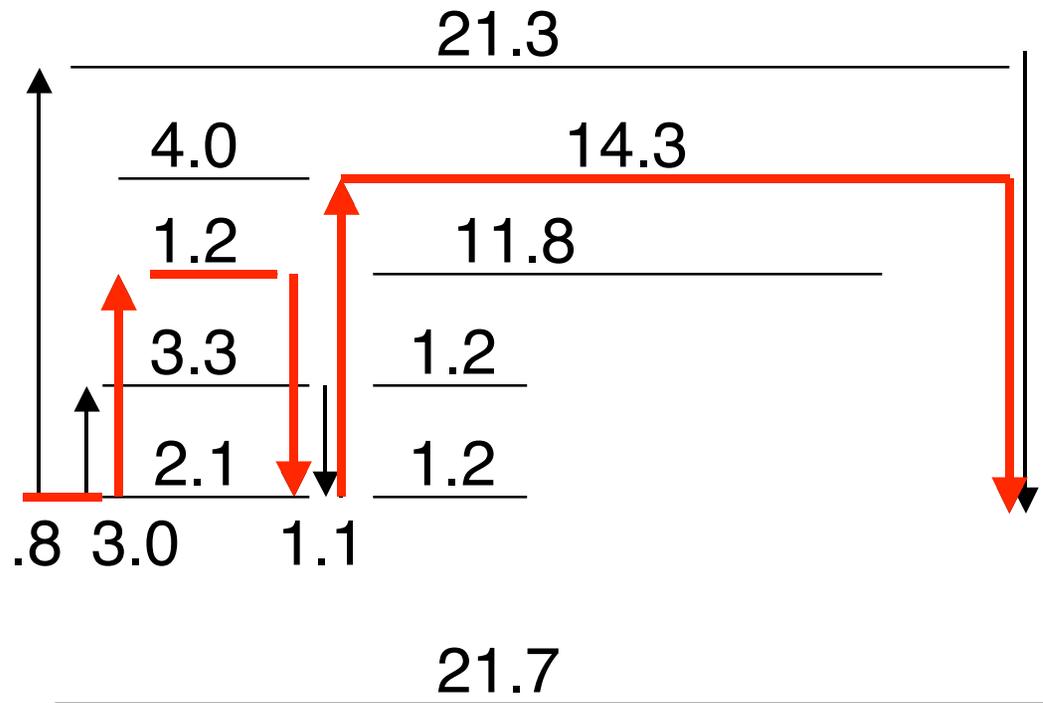
48 atm

16 ice

8 Ind

16 cpl

152 total



Seconds per simulated day

CCSM Throughput vs Resolution

Measured on IBM power4 system, bluesky, as of 9/1/2003

Atmosphere Resolution	Ocean Resolution	Number of Processors	Throughput (yrs/day)
T31 (3.7°)	3°	68	12.0
T42 (2.8°)	1°	152	10.9
T42	1°	120	8.6
T42	1°	104	7.2
T85 (1.4°)	1°	200	4.0
T170 (0.7°)	1°	400	2.0 (estimate)

CCSM Throughput vs Platform

T42 atmosphere 1° ocean

Platform	Number of Processors	Throughput (yrs/day)
IBM power3 wh (NCAR)	104	3.5
IBM power3 nh (NERSC)	112	4.0
SGI O3K (LANL)	112	4.0
Linux cluster* (ANL)	104	4.0-7.0 (estimate)
HP/CPQ Alpha (PSC)	104	6.0 (estimate)
IBM power4 (NCAR)	104	7.2
IBM power4 (NCAR)	152	10.9
NEC Earth Simulator	160	60.0 (estimate)
Cray X1 (ORNL)	160	60.0 (estimate)

* Argonne jazz cluster, 2.4Ghz Pentium

Conclusions / Future Work

We have presented preliminary results in our ongoing work to benchmark and tune the performance of CCSM. We have a comprehensive set of CCSM's components RISC-based platforms. We have preliminary component scaling results for vector platforms, but are still in the process of tuning many of CCSM's components for vector platforms such as the Earth Simulator and Cray X-1. We have demonstrated that scaling estimates can be used to guide resource allocation for CCSM, but that intercomponent interactions and data dependencies complicate the process sufficiently that it is hard to make accurate predictions of overall throughput using these figures. We have shown that MPE instrumentation of the CCSM has been helpful in refining these estimates. Finally, we have presented overall throughput figures for CCSM that are measured on some platforms, and *estimated* for the Cray X-1 and the Earth Simulator. Future work on this project will be devoted to refining our data collection process, collection of new data as optimization proceeds for the X1 and the ES platforms, and exploration of the applicability of run-time instrumentation tools such as Tuning and Analysis Utilities (TAU) [9].

Acknowledgements

The authors would like to thank several people for contributing their time and useful advice: Phil Jones of Los Alamos National Laboratory, J.B. White of Oak Ridge National Laboratory, and John Levesque of Cray, Incorporated. The authors also thank their respective employers for their support of this work.

Many of the authors receive support through the *Collaborative Design and Development of the Community Climate System Model for Terascale Computers* project, which is funded through the US Department of Energy's Climate Change Prediction Program, a part of the DOE Scientific Discovery through Advanced Computing (SciDAC) initiative. Some of the authors are supported by the *Project for the Sustainable Coexistence of Humans, Nature, and the Earth*, which is funded by the Ministry of Education, Culture, Sports, Science and Technology of the Government of Japan.

Argonne National Laboratory is managed by the University of Chicago for the US Department of Energy under contract W-31-109-ENG-38. Oak Ridge National Laboratory is managed by UT/Batelle for the US Department of Energy under contract DE-AC-05-00OR22725. The National Center for Atmospheric Research is managed by the University Corporation for Atmospheric Research for the US National Science Foundation.

References and Notes

- [1] National Center for Atmospheric Research (NCAR) Community Climate System Model (CCSM), <http://www.cesm.ucar.edu/models> .
- [2] Bettge, T., Craig, A., James, R., Wayland, V., and Strand, G. (2001): “The DOE Parallel Climate Model (PCM): The Computational Highway and Backroads,” Proceedings of the International Conference on Computational Science (ICCS) 2001, V.N. Alexandrov, J.J. Dongarra, B.A. Juliano, R.S. Renner, and C.J.K. Tan (eds), Springer-Verlag LNCS Volume 2073, pp 149-158.
- [3] Jacob, R.L., Schafer, C., Foster, I., Tobis, M., and Anderson, J., (2001): “Computational Design and Performance of the Fast Ocean-Atmosphere Model, Version One,” Proceedings of the International Conference on Computational Science (ICCS) 2001, V.N. Alexandrov, J.J. Dongarra, B.A. Juliano, R.S. Renner, and C.J.K. Tan (eds), Springer-Verlag LNCS Volume 2073, pp 175-184.
- [4] For examples, see the Web sites for the CCSM coupler (<http://www.cesm.ucar.edu/models/cpl6>) and the European Centre for Research and Advanced Training in Scientific Computation (CERFACS) Ocean-Atmosphere-Sea-Ice-Surface (OASIS) coupler (<http://www.cerfacs.fr/globc/software/oasis>).

[5] For example, see the Fifth-generation Penn State/NCAR Mesoscale Model (MM5) benchmark Web site (<http://www.mmm.ucar.edu/mm5/mpp/helpdesk/20030305.html>).

[6] Larson, J.W., Jacob, R.L., Foster, I.T., and Guo, J. (2001): “The Model Coupling Toolkit,” Computational Design and Performance of the Fast Ocean-Atmosphere Model, Version One,” Proceedings of the International Conference on Computational Science (ICCS) 2001, V.N. Alexandrov, J.J. Dongarra, B.A. Juliano, R.S. Renner, and C.J.K. Tan (eds), Springer-Verlag LNCS Volume 2073, pp 185-194

[7] MCT Web site (<http://www.mcs.anl.gov/mct>).

[8] Ong, E.T., Larson, J.W., and Jacob, R.L. (2002): “A Real Application of the Model Coupling Toolkit,” Proceedings of the International Conference on Computational Science (ICCS) 2002, C.J.K. Tan, J.J. Dongarra, A.G. Hoekstra, and P.M.A. Sloot (Eds.), LNCS Volume 2330, Springer-Verlag, pp. 748-757.

[9] TAU Web site <http://www.cs.uoregon.edu/research/paracomp/tau/tautools/>