PARALLELIZATION

of an ocean general circulation model on the CRAY T3D system

Josef M. Oberhuber
Deutsches Klimarechenzentrum GmbH (DKRZ)
Hamburg, Germany

Klaus Ketelsen
Cray Research, Inc.

Some evidence suggests that increasing amounts of atmospheric pollution are threatening the delicate balance of the Earth's ecosystems. High-resolution ocean models used in coupled atmosphere-ocean-land simulations are vital to predicting climate changes and their time scales. Their size and complexity demand such vast amounts of computing resources that costs have been prohibitive and turnaround times impractical to date. A cooperative project completed recently between DKRZ, the German Climate Computing Center in Hamburg, and Cray Research, during which the ocean general circulation model OPYC—already implemented on a CRAY C90 system—was ported to a CRAY T3D massively parallel processing system, meant an important step forward in finding a practical, affordable solution for computationally intensive ocean modeling.

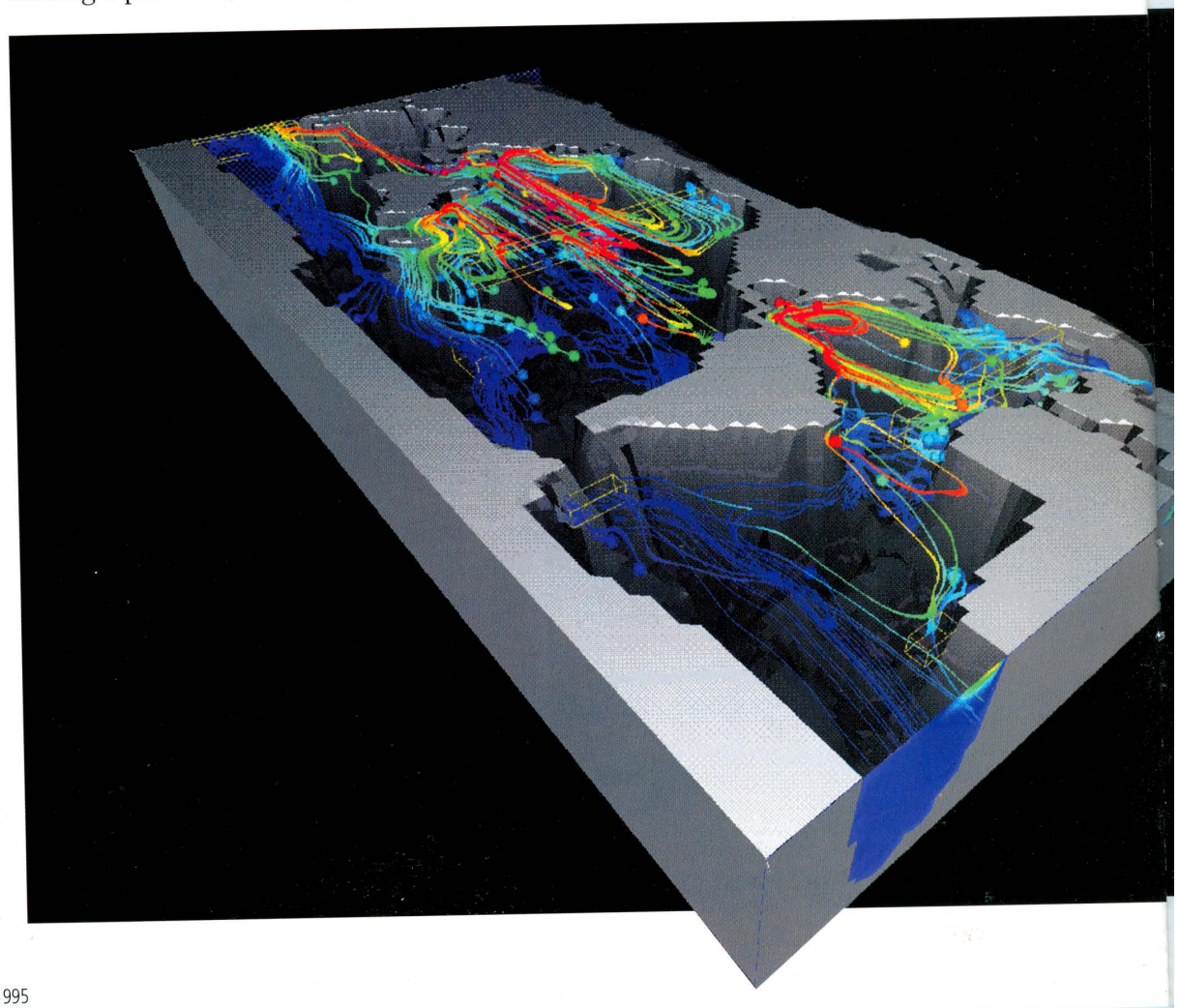
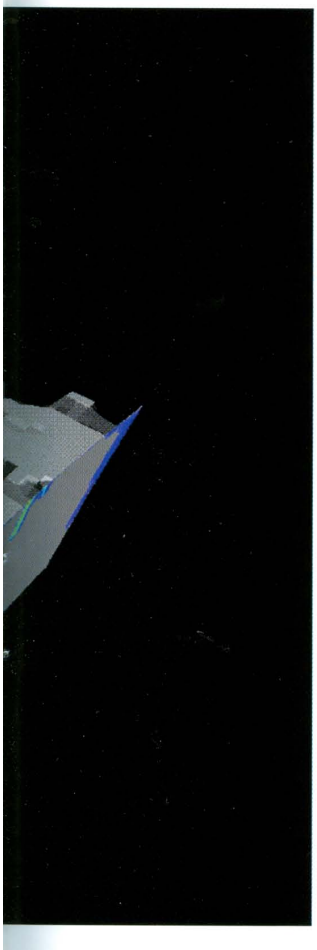


Figure 1. OPYC, a state-of-the-art system to simulate ocean general circulation, reveals a three-dimensional flow in the global ocean basin visualized by particle trajectories, where particles have been released at specified locations. Trajectories in blue indicate low temperature, red indicates high temperature.

Rising sea levels and global warming caused by increasing amounts of greenhouse gases released into the atmosphere are examples of pressing global environmental issues for which decision makers must find quick solutions to preserve life on planet Earth. To document the delicate balance of the Earth's ecosystems and provide decision makers with predictions about the consequences of disturbing that balance, environmental researchers use computer simulation models that span time scales ranging from days to millennia.

In recent years, ocean models have become an important part of these environmental studies because of the ocean's pivotal role in climate variability. Ocean models are used to study the variability of ocean circulation across years, centuries, and millennia in order to understand the physical causes of El Niños (warm events in the equatorial Pacific) or of warm or cold periods in the Earth's history, for example, and to study their influence on global ocean circulation. Extremely high-resolution ocean models are capable of resolving oceanic eddies and allow studying their influence on global ocean circulation or the natural variability of the ocean, while low-resolution models are used to study climate variability on extremely long time scales. Used in coupled atmosphere-ocean-land models, ocean models are vital to predicting climate changes and their time scales resulting from any kind of increased atmospheric pollution.

Even when run on highest-performance parallel-vector supercomputers, the sheer size and complexity of high-resolution ocean models make computation costs prohibitive and turnaround times impractical when used for time scales of centuries. For example, a high-resolution global ocean model requires a resolution of 0(10 km), which results in a 3600 x 1800 horizontal grid size times approximately 40 levels vertically.



Today, massively parallel processing (MPP) systems, such as Cray Research's CRAY T3D system, hold the promise of providing the vast amounts of computing resources required for ocean modeling. To begin tapping these powerful resources, researchers at the Deutsches Klimarechenzentrum (DKRZ), the German Climate Computing Center in Hamburg, and computer specialists from Cray Research are cooperating on a project to port OPYC[1] (Ocean model with isoPYCnal coordinates), an ocean general circulation model already implemented on DKRZ's CRAY C90 system—and on many other Cray Research systems worldwide—to the CRAY T3D system.

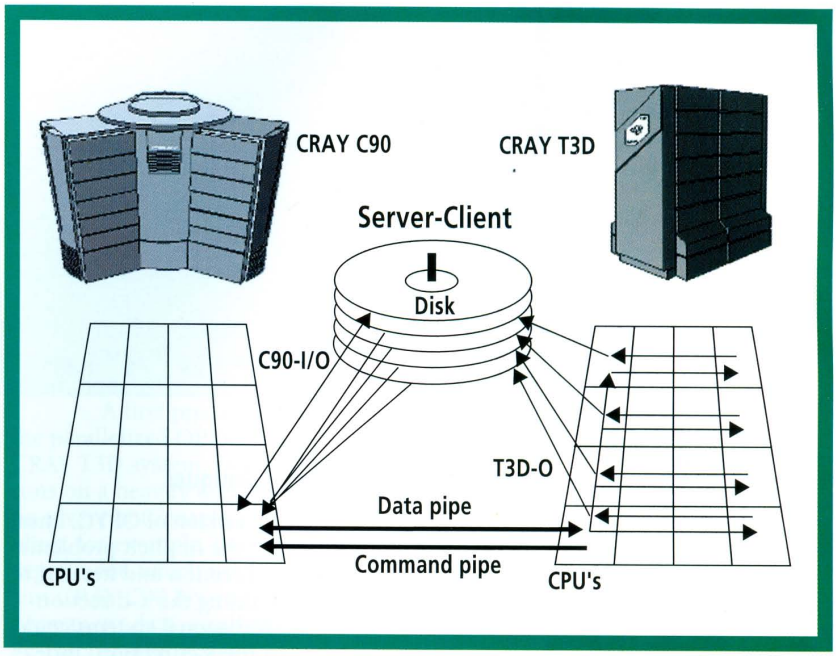


Figure 2. A client-server concept was used during parallelization of the ocean general circulation model OPYC on the CRAY T3D system. The diagram shows the flow of various data streams. A data and a command pipe are used to initialize model variables on the CRAY T3D system and to extract the final ocean state for a subsequent restart of the model. Intermediate data frequently generated by the ocean model are saved on disk. These data, which are still partitioned in one geographical direction, are asynchronously read by the CRAY C90 system and converted to a final file format.

The ocean model

The ocean general circulation model OPYC is a state-of-the-art system for simulating large-scale ocean circulation, as shown in Figure 1. It uses primitive equations in flux form with free surface formulation. The density field is computed from temperature and salinity via a general state equation for seawater. Horizontally, the model uses a regular, staggered grid on spherical coordinates. Vertically, isopycnals (planes of constant potential density) are used as Lagrangian coordinates. In addition to standard parameterizations for convection, isopycnal and diapycnal mixing, a mixed layer model simulates the depth of a turbulent surface layer. A dynamic-thermodynamic sea-ice model predicts ice and snow cover according to a comprehensive ice rheology. The OPYC model uses a fully implicit time integration scheme. This leads to nonlinear equations for the three-dimensional wave equation, for advection and diffusion of momentum, mass, heat, and salt, and for thickness and concentration of ice and snow and the related momentum.

Parallelization on the CRAY T3D system

The basic strategy for parallelizing the time-stepping part is the (x,y)-domain decomposition. Columns (i.e., the z-direction) are located on the same processing element (PE). Each PE carries additional latitudes and longitudes for boundary values, as does the CRAY C90 version for the single domain. In addition, the client-server concept is used as a guideline (Figure 2).

Table 1. Benchmarks for T106 model without and with postprocessing.

PEs	npe_x	npe_y	Without postprocessing Time (sec)	Without postprocessing MFLOPS	With postprocessing Time (sec)	With postprocessing MFLOPS	With postprocessing MFLOPS/PE	With postprocessing Speedup
16	4	4	2459	273	2474	271	16.9	16.0
32	4	8	1234	541	1250	534	16.7	31.6
64	8	8	675	990	680	983	15.3	57.9
128	8	16	346	1930	346	1930	15.1	114.4
256	16	16	198	3368	200	3338	13.0	196.9
512	16	32	97	6885	101	6646	13.0	393.8

PARALLELIZATION

Changes in the iterative techniques

As in the CRAY C90 version of OPYC, linear equation systems for each of the implicit problems are solved directly in the x-direction and iterated in the y-direction. After partitioning the x-direction into npe_x subdomains, the tridiagonal systems could be solved across all PEs in the x-direction. Independent of the chosen algorithm, this would result in a number of sequential operations across all PEs in the x-direction to solve the linear equations. Consequently, one may expect that such an algorithm will not scale well when the code is run on a truly massively parallel computer with 512 PEs, for example. This was demonstrated by Eltgroth,[2] who developed an efficient technique for the solution of linear equations for MPP systems.

Such a technique has already been applied to OPYC's sea-ice model, which, as part of the coupled atmosphere-ocean-land model at Lawrence Livermore National Laboratory (LLNL) in Livermore, California, is coupled to the Modular Ocean Model (MOM) from Geophysical Fluid Dynamics Laboratory (GFDL) in Princeton, New Jersey. Eltgroth showed that such a scheme scales reasonably only up to 0(100) PEs.

To achieve good scaling for a larger number of PEs, we developed an engineering type of approach. Instead of solving one tridiagonal equation across all PEs in the x-direction, we formulated a complete linear equation system for each PE and coupled the equations by a back-and-forth shifting of the grid point data relative to the virtual address space. This data shifting is carried out after each iteration. Thus, boundary values located near the interface between adjacent PEs will be located in the center of the PE grid point space and thus in the center of the linear equation system during the next iteration. This method results in a convergence rate similar to a fully direct method and yields surprisingly few changes to the version implemented on the CRAY C90 system.

Heterogeneous environment

The code was split into client and server components. The server remains on the CRAY C90 system and performs the initialization phase and postprocessing. In addition, the server supplies the client with atmospheric forcing data required periodically each month. The server reads these data and sends them via a named pipe to PE(0,0). PE(0,0) distributes the data to all other PEs. The client runs on the CRAY T3D system and carries out the computation. During the initialization phase the server fills all COMMON blocks and sends a copy to PE(0,0) on the CRAY T3D system via a named pipe. PE(0,0) partitions the data in the x- and y-directions and distributes the respective parts to the other PEs. At the end of the client program all modified COMMON blocks are returned to the server. These data are used to generate a restart file. Using the client-server concept for porting the code offers the following advantages:

- The number of routines that had to be ported was minimized. Only the computationally intensive part of OPYC runs on the CRAY T3D system. On the other hand, if the model is initialized once from scratch, the expensive and hardly parallelizable routines for data preprocessing, etc., run on the CRAY C90 system.
- The CRAY T3D version of OPYC is transparent to the user. All files that communicate with the outside world are created on the CRAY C90 system. From the user's point of view there is no difference using either the combined CRAY C90/T3D version or the full CRAY C90 version.

Communication

The parallel part of OPYC is running on the CRAY T3D system as a single program multiple data (SPMD) code. Explicit shared memory message passing is used for communication on the CRAY T3D system. Communication between different PEs is required in three different cases:

- Zonal and meridional boundary conditions
- Data shift in the x-direction to improve convergence of the solvers
- Global sums

An analysis of the network activity on 32 PEs showed that 1.5 percent of the total CRAY T3D time is used for global sums, 2 percent for zonal and meridional boundary conditions, and 8 percent for the data shift mechanism. While the network is active, 964.4 Mbytes/s are moved between all PEs.

Data postprocessing

The output of postprocessing data has been optimized to meet the requirements of a large CRAY T3D system with 512 PEs. To take better advantage of the I/O bandwidth of the CRAY T3D system, the first PE, i.e., PE(0,y), of all y-groups collects the data of its group and writes them onto an intermediate disk file. The server collects the postprocessing data from the intermediate files, converts the format from IEEE to Cray-binary format, and merges them into the final postprocessing file. This asynchronous procedure offers the advantage that the server may pick up the data from distributed files whenever the CRAY C90 system's scheduling provides the server with some resources.

Optimization for the DEC Alpha chip

To improve the performance of OPYC, optimization work has been done to take into account the architecture of Digital Equipment Corporation's Alpha chip. Alpha is a RISC processor with a slow hardware floating point divide. The main goals of the optimization work have been

1. Reducing memory traffic. The following techniques were used:
   - □ Loop collapsing
   - □ Storing intermediate results in temporary variables
   - □ Reordering the sequence of computation
2. Better use of cache, for example, due to swaps of loop indices to achieve an increment of one in the innermost loop.
3. Reducing the number of divides by rearranging loops or exchanging divides by a constant for a multiply by a constant.

Benchmark tests

Runs with a variety of partitions (Table 1) for the global ocean model with T106 resolution (320 x 160 horizontal grid points) show that the performance does not drop significantly for up to 512 PEs. There are only small differences when data postprocessing is switched on or off. After completing the Alpha chip optimization we achieved 292 MFLOPS on 16 PEs, 593 MFLOPS on 32 PEs, 1029 MFLOPS on 64 PEs, and 2173 MFLOPS on 128 PEs for a lower-resolution T42 (128 x 64 horizontal grid points) model. This means that for this problem 23 PEs on the CRAY T3D system are equivalent to one CRAY C90 CPU. The T106 model ran with 7.4 GFLOPS on the CRAY T3D system with 512 PEs.

Conclusion

During parallelization of the ocean general circulation model OPYC, the client-server layout was found to be useful because of the small amount of work involved in porting the entire code and because of the user transparency achieved.

Use of an implicit time step scheme in combination with the line-relaxation method and the data shift mechanism resulted in a code that involves surprisingly few changes relative to the CRAY C90 version. The network activity caused by the data shift mechanism was found to be tolerable on the CRAY T3D system and scales very well.

Data postprocessing was found to be crucial for production runs. An asynchronous write from the client and read into the server was found to be the best strategy. The server reconstructs all full-domain data in a CRAY C90-compatible format. Pipes are used only to initialize the CRAY T3D system, to return data that the server requires for generating the restart file, and to supply the client with monthly data. It was not found useful to transfer data frequently across pipes because the server on the CRAY C90 usually is not running in dedicated mode.

A first production run demonstrated that the parallelized OPYC runs efficiently on a 32-PE CRAY T3D system. Despite the fact that the server runs on a heavily used CRAY C90 system at DKRZ, no severe bottlenecks, caused by CRAY C90 scheduling, for example, have appeared. This scientific experiment, which used approximately 1500 equivalent CRAY C90 hours and produced 100 Gbytes of permanent data, confirmed the reliability of Cray Research's MPP system.

Porting the general ocean circulation model OPYC to the CRAY T3D system is an important step toward finding a practical, affordable solution for computationally intensive ocean modeling. Improved turnaround and greater model precision are indispensable for finding timely answers to critical environmental problems. ▬

Acknowledgments

Josef Oberhuber thanks Cray Research for its support for this project, both for access to a CRAY T3D system at Cray Research's computing facility and for the time provided by Klaus Ketelsen. The authors acknowledge the support of Michael Böttinger from the Visualization Group at DKRZ for generating the three-dimensional image using the Data Visualizer from Wavefront Technologies, Inc. in Santa Barbara, California.

About the authors

Josef M. Oberhuber is a modeler within the Model Application and Development group at DKRZ in Hamburg, Germany, who focuses on ocean modeling and coupled ocean-atmosphere-land models for climate predictions. Oberhuber received a Ph.D. degree in meteorology in 1984. He developed an ocean general circulation model at the Max Planck Institute for Meteorology and the Meteorological Institute of the University, both in Hamburg, Germany.

Klaus Ketelsen is an applications analyst in the Marketing and Sales Support group at Cray Research GmbH, Germany, focusing on porting application codes to the CRAY T3D system. He received an advanced degree in electrical engineering from the University of Hannover in 1973 and has worked in the petroleum industry for nine years.

References

1. Oberhuber, J. M., *The OPYC Ocean General Circulation Model*, Technical Report No. 7, Deutsches Klimarechenzentrum, Hamburg, 1993, p. 130.
2. Eltgroth, P. G., *Two Portable Parallel Tridiagonal Solvers*, UCRL-JC-118017, Lawrence Livermore National Laboratory, 1994, p. 9.