Investigating Solution Convergence in a Global Ocean Model Using a 2048-Processor Cluster of Distributed Shared Memory Machines

Investigating solution convergence in a global ocean model using a 2048-processor cluster of distributed shared memory machines

Chris Hill[a,*], Dimitris Menemenlis[b], Bob Ciotti[c] and Chris Henze[c]

[a]*Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Room 54-1515, 77 Massachusetts Avenue, Cambridge, MA 02139, USA*

[b]*Jet Propulsion Lab, California Institute of Technology, CA, USA*

[c]*NASA Advanced Supercomputing, Ames Research Center, CA, USA*

Abstract. Up to 1920 processors of a cluster of distributed shared memory machines at the NASA Ames Research Center are being used to simulate ocean circulation globally at horizontal resolutions of 1/4, 1/8, and 1/16-degree with the Massachusetts Institute of Technology General Circulation Model, a finite volume code that can scale to large numbers of processors. The study aims to understand physical processes responsible for skill improvements as resolution is increased and to gain insight into what resolution is sufficient for particular purposes. This paper focuses on the computational aspects of reaching the technical objective of efficiently performing these global eddy-resolving ocean simulations. At 1/16-degree resolution the model grid contains 1.2 billion cells. At this resolution it is possible to simulate approximately one month of ocean dynamics in about 17 hours of wallclock time with a model timestep of two minutes on a cluster of four 512-way NUMA Altix systems. The Altix systems' large main memory and I/O subsystems allow computation and disk storage of rich sets of diagnostics during each integration, supporting the scientific objective to develop a better understanding of global ocean circulation model solution convergence as model resolution is increased.

Keywords: Computational fluid dynamics, ocean modeling, parallel computing

1. Introduction

In this article we describe technical aspects of global ocean model configurations with resolutions up to 1/16° (≈ 5 km) that exploit a testbed 2048 Itanium-2 processor SGI Altix system at the NASA Ames Research Center. The size of this system makes possible global ocean simulations at resolutions that until now have been impractical. The calculations we describe employ an Adams-Bashforth time-stepping procedure on a finite volume grid with up to 1.25 billion grid cells. This workload is spread evenly over 1920 of the Altix processors so that each individual processor is responsible for simulating around 586,000 grid cells (corresponding to a surface region roughly 210 × 210 km in extent). Decomposing the workload over this many processors yields a setup that, with extensive diagnostics and analysis options included, uses about 870 MB of main memory per processor and can integrate forward at a rate of around 5 μs/timestep/gridcell. With a timestep of two minutes this performance allows a year of simulation to be completed in under ten days.

The model configurations we employ are significant in that, at the resolutions the Altix system makes possible, numerical ocean simulations begin to truly represent the key dynamical process of oceanic meso-scale turbulence. Meso-scale turbulence in the ocean is, in some ways, the analog of synoptic weather fronts in the atmosphere. However, because of the density characteristics of seawater, the length scale of turbulent eddy

[*] Corresponding author. Tel.: +1 617 253 6430; Fax: +1 617 253 4464; E-mail: cnh@mit.edu.

ISSN 1058-9244/07/$17.00

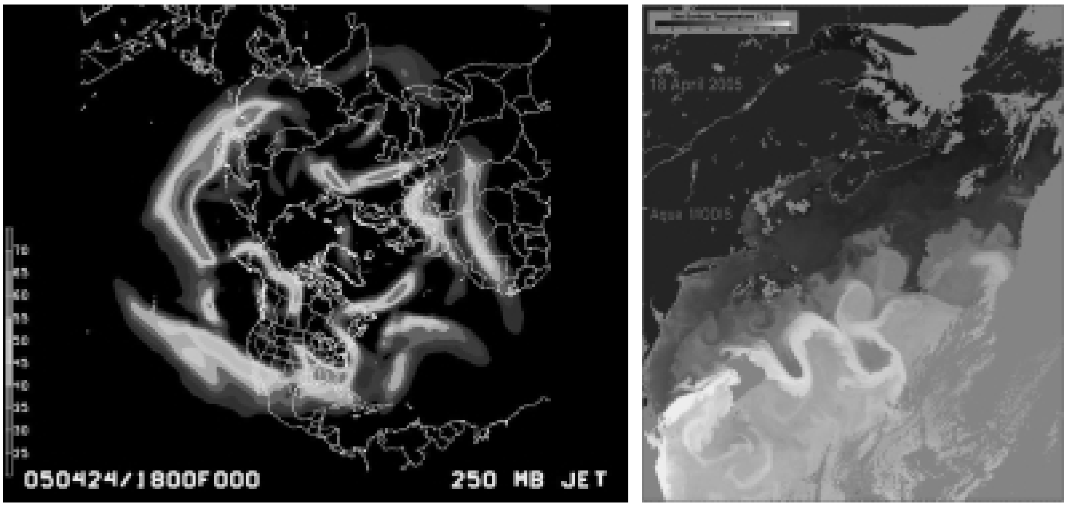

Fig. 1. Atmosphere and ocean eddies. Eddies occur in both the atmosphere and ocean but with vastly different length scales. The left panel shows a polar view of the atmospheric jet stream meandering as it circulates the northern hemisphere. The length scales of the meanders are several thousand kilometers. The right panel shows satellite imagery of Gulf Stream meanders off the U.S. East Coast. In places the length scales of the meanders are as small as ten kilometers.

phenomena in the ocean is around ten or less kilometers. In contrast, in the atmosphere, where the same dynamical process occurs, it has length scales of thousands of kilometers – see Fig. 1. Although it has been possible to resolve ocean eddy processes well in regional ocean simulations [12] and process studies [7,14,28] for some time, global scale simulations that resolve or partially resolve the ocean's energetic eddy field are still rare [17,23] because of the immense computational challenge they represent.

Physically ocean meso-scale eddies contribute to regulating large-scale ocean transports of momentum, heat and fresh water, all of which play an important role in creating the Earth's climate. Eddies also interact, non-linearly, with the larger scale general circulation thereby modifying an ocean simulation's mean state and altering its transient response to perturbations. Through these processes, and through the processes by which ocean energy is dissipated at basin margins, meso-scale eddies play a significant, complex, and non-linear role in climate. Process studies [14] suggest that resolutions of a few kilometers are needed to adequately resolve eddy processes. Consequently ocean modelers are continually on the look out for technological means to increase the resolution of global-scale integrations in order to converge on an adequate representation of these non-linear behaviors. The work we describe here creates the technological machinery for examining how much resolution is needed for eddies, and their impacts, to be accurately represented in a global ocean model.

The article is organized as follows. We first review prior work in Section 2. A description of the underlying mathematical and numerical basis of our model together with a discussion of its computational formulation is in Section 3. Section 4 provides details on how the 2048 shared memory Altix cluster is configured, looks at how our parallel algorithm maps to the system, and describes the resultant performance. In Sections 5 and 6, we summarize the prospects for using a large Altix system for optimal representation of eddy terms in a global ocean model.

2. Prior work

Eddying calculations first started to appear around 1990 [2,4]. These initial calculations were facilitated by the availability of vector machines such as the Cray YMP with sustained memory bandwidths of several gigabytes/second per processor. Global eddying calculations were first carried out [30] around this time too. During this period computational resources supported simulations that permitted eddies, but did not allow models to resolve the scales shown in Fig. 1. More recently [12,25,29,31] have undertaken large-scale regional simulations that are termed eddy-resolving in that they resolve the first baroclinic Rossby radius of deformation in the ocean [5], a physical estimate of a typical ocean eddy size. In these regional studies authors find that the transition from eddy-permitting to eddy-resolving produces qualitative improvements in the fit between simulations and observations. Following these and other studies [17] reported on a global $\frac{1}{10}^\circ$ simulation on an IBM Power 3 system and [23] described a global $\frac{1}{10}^\circ$ simulation on the Earth Simulator in Japan. These studies further reinforce the notion that resolving eddy processes provides a significant boost in model skill. In this paper we report on the computational aspects of a study that uses a cluster of four 512-way NUMA Altix systems for global eddy resolving simulations at up to $\frac{1}{16}^\circ$ resolution. The study is aimed at developing a clearer understanding of processes underlying skill improvements that eddy resolving models show. To develop this understanding we perform a series of heavily instrumented simulations that produce more than a terabyte of output for each year simulated and that repeatedly process more than one billion grid cells in a time-stepping procedure. In doing this we gain insights into the performance of a clustered Altix system with 2048 CPUs for ocean simulation.

3. Algorithm

The Masschusetts Institute of Technology General Circulation Model (MITgcm) algorithm is rooted in the incompressible form of the Navier-Stokes equations for fluid motion in a rotating frame of reference [11,20,21]. The model supports different numerical configurations including hydrostatic and non-hydrostatic simulations. A generic, height like, vertical coordinate allows the fluid equations to be solved in either height or pressure coordinates, making the code amenable to both ocean and atmospheric simulations [9,19]. Adjoint forms of the code are automatically maintained [8] supporting advanced assimilation [32] and dynamical analysis [10, 18]. The configuration used here is hydrostatic and the underlying continuous equations can be written as partial differential equations, in an Eulerian form, that govern the three-dimensional temporal and spatial evolution of velocity u, potential temperature θ, salinity s and pressure $g\rho_{ref}\eta + p_{hyd}$ within the fluid

$$\frac{\partial \vec{\mathbf{u}}_h}{\partial t} + \nabla_h(g\rho_{ref}\eta + p_{hyd}) = \vec{\mathbf{G}}_{\vec{\mathbf{u}}_h} \tag{1}$$

$$\frac{\partial \gamma}{\partial t} = G_\gamma \text{ where } \gamma = \theta, s, \lambda \tag{2}$$

$$\nabla^2\eta + \frac{\partial \eta}{\partial t} = \nabla \cdot \widehat{G_{\vec{\mathbf{u}}_h}} \tag{3}$$

$$\frac{\partial p_{hyd}}{\partial z} = -g\rho(\theta, s, z) \tag{4}$$

$$\frac{\partial w}{\partial z} = -\nabla_h \overrightarrow{\mathbf{u}_h} \tag{5}$$

In Eq. (1) $g\rho_{ref}\eta$ and p_{hyd} are the surface pressure (due to fluid surface elevation η) and the internal hydrostatic pressure (due to the mass of the fluid above a particular depth), respectively. The subscript $_h$ denotes operations in the horizontal plane. Hydrostatic pressure relates to fluid potential temperature, θ, salinity, s and depth z according to the function ρ in (4) multiplied by gravity g. The term $\vec{\mathbf{G}}_{\vec{u}_h}$ in Eq. (1) incorporates sources and sinks of momentum, due, for example, to surface winds or bottom drag together with terms that represent coriolis forces and the transport of momentum by the fluid flow. Equation (2) captures the time evolution of the ocean heat content, θ, and salt, s and tracers, λ, respectively. Right-hand-side terms, $G_{\theta,s,\lambda}$, include source terms due to surface fluxes and transport by the fluid flow.

To evaluate Eq. (1) we require knowledge of the pressure field terms in the fluid, $g\rho_{ref}\eta + p_{hyd}$. The hydrostatic part of the pressure field can be found by integrating the fluid density relation $\rho(\theta, s, z)$ down from the surface, where $p_{hyd} = 0$. The surface pressure is the product of the fluid free surface elevation η multiplied by gravity g and a reference density ρ_{ref}. Invoking the vertically integrated form of the continuity condition for an incompressible fluid with a free surface, $\nabla \cdot \widehat{\vec{\mathbf{u}}} + \frac{\partial \eta}{\partial t} = 0$, yields a diagnostic relation Eq. (3) for η. Here $\widehat{\phi}$ indicates a vertical integral from the bottom of the ocean to the surface. The vertical velocity w is found by integrating Eq. (5) vertically from the bottom of the ocean.

Equations (1) and (2) are stepped forward explicitly in time using an Adams-Bashforth procedure that is second order accurate. The equations are discretized in space using a finite volume based technique [1] yielding a solution procedure that requires at each time step explicitly evaluated local finite volume computations and an implicit two-dimensional elliptic inversion.

To solve Eqs (1)–(5) numerically we write them in terms of discrete time levels n, $n + \frac{1}{2}$ and $n + 1$

$$\frac{\overrightarrow{\mathbf{u}_h}^{n+1} - \overrightarrow{\mathbf{u}_h}^{n}}{\Delta t} + \nabla_h(\eta^{n+1} + p_{hyd}^{n+\frac{1}{2}}) = \overrightarrow{\mathbf{G}}_{\overrightarrow{\mathbf{u}}_h}^{n+\frac{1}{2}} \tag{6}$$

$$\frac{\gamma^{n+1} - \gamma^n}{\Delta t} = G_\gamma^{n+\frac{1}{2}} \text{ where } \gamma = \theta, s, \lambda \tag{7}$$

$$\nabla^2\eta^{n+1} + \frac{\eta^{n+1} - \eta^n}{\Delta t} = \nabla \cdot \widehat{\overrightarrow{\mathbf{G}}_{\vec{\mathbf{u}}_h}}^{n+\frac{1}{2}} + \frac{\widehat{\overrightarrow{\mathbf{u}_h}^{n}}}{\Delta t} - \widehat{p_{hyd}^{n+\frac{1}{2}}} \tag{8}$$

$$\frac{\partial p_{hyd}^{n+1}}{\partial z} = -g\rho(\theta^{n+1}, s^{n+1}, z) \qquad (9)$$

$$\frac{\partial w^{n+1}}{\partial z} = -\nabla_h \overrightarrow{\mathbf{u}_h}^{n+1} \qquad (10)$$

The $^{n+\frac{1}{2}}$ terms in Eqs (6)–(10) denote centered in time values that are explicitly evaluated from known values at time levels n and $n-1$ using Adams-Bashforth extrapolation, $\phi^{n+\frac{1}{2}} = (\frac{3}{2}+\epsilon)\phi^n - (\frac{1}{2}+\epsilon)\phi^{n-1}$. Time levels are separated by a dimensional time Δt. The computational algorithm then proceeds as a series of $n_{\text{final}} - n_{\text{initial}}$ time steps according to algortihm 6.

Algorithm 6 shapes the computational formulation of MITgcm. Lines 2: and 5: of the algorithm involve integrals along the vertical axis of the discrete domain. Typically these integrals contain dependencies that prevent efficient parallel decomposition along this axis. Accordingly the parallel formulation takes a global finite volume domain with $N_x \times N_y \times N_z$ cells in the two horizontal directions, x and y, and the vertical direction, z, respectively, and decomposes it into $N_{sx} \times N_{sy}$ sub-domains each of size $(S_{nx} + 2 \times O_x) \times (S_{ny} + 2 \times O_y) \times N_z$ such that $S_{nx} \times N_{sx} = N_x$ and $S_{ny} \times N_{sy} = N_y$. The O_x and O_y values are overlap region finite volume cells that are added to the boundaries of the subdomains to hold replicated data from neighboring subdomains. The values of O_x and O_y are chosen such that the explicit steps 2: and 4: in algorithm 6 can be evaluated locally by each subdomain independent of others. Each computational process integrating forward the MITgcm is then given a static set of one or more subdomains. The process is then responsible for evaluating all the terms required by algorithm 6 for its subdomains and for cooperating with other processes to maintain coherent data in the overlap region finite volume cells. This yields a decomposed computational procedure outlined in algorithm 6. In algorithm 6 a single time-step is split into a series of *compute*, *exchange*, and *sum* phases. *Compute* phases contain only local computation (predominantly arithmetic and associated memory loads and stores) and I/O operations. Performance of *compute* phases is sensitive to the volume of I/O and computation involved, local CPU and memory capabilities of the underlying system, and to the system I/O capacity. *Exchange* phases involve point-to-point communication between neighbor processes. Their performance hinges on the performance of the systems interconnect and inter-process communication software stack. *Sum* phases involve all subdomains collectively combining locally calculated 8-byte floating point values to yield a single global sum. The *sum* phases are sensitive to how system performance for collective communication scales with processor count.

4. Running on the Altix system

The system we are using consists of four interconnected 512 processor SGI Altix machines each with a separate OS image. Each Altix system uses 1.6 GHz Itanium 2 processors arranged in pairs and is interconnected with SGI's BX2 NUMALink 4 technology through a front side bus interface. Within each 512-way system NUMAlink supports non-uniform access global shared memory [33] through memory coherence technology rooted in the distributed shared memory approaches of the Stanford DASH project [16]. The network fabric within an Altix system is a fat-tree arrangement [3,15] based on 8-port router chips. This provides for 64 bisection links on a 512-way system giving a total bisection bandwidth of 400TB/s. Each CPU has memory banks populated with 2 GB of 166MHz DDR memory, so that the core memory of a 512-way unit is 1TB. On the Stream benchmark [24] a 512-way system achieves a sustained memory bandwidth for the triad value of just over 1TB/s [6]. Each 512 way system runs a version of the Linux 2.4 kernel with specific SGI modifications that supports scaling to large numbers of shared memory processors [27].

We scale our runs to 1920 CPUs using a cluster of four 512-way systems interconnected with a single NUMAlink fat-tree. The SGI Array Services and Message Passing Tookit systems are available on the Altix system, providing optimized MPI implementations that can take advantage of the coherent shared memory within each 512-way unit and transparently adapt to non-coherent memory communication, still over NUMAlink, between systems. The *exchange* and *sum* phases in the MITgcm algorithm are, therefore, configured to use the systems native MPI implementation for the runs we describe.

The most challenging numerical experiment we have undertaken is a $\frac{1}{16}^{\circ}$ near global simulation. In this section we examine the performance of that simulation on the Altix system. The analysis makes use of operation counts based on code inspection as well as runtime processor statistics provided by the pfmon utility [13] which reads Itanium hardware counters and per MPI process communication statistics available when the `-stats` option of the Message Passing Toolkit is enabled.

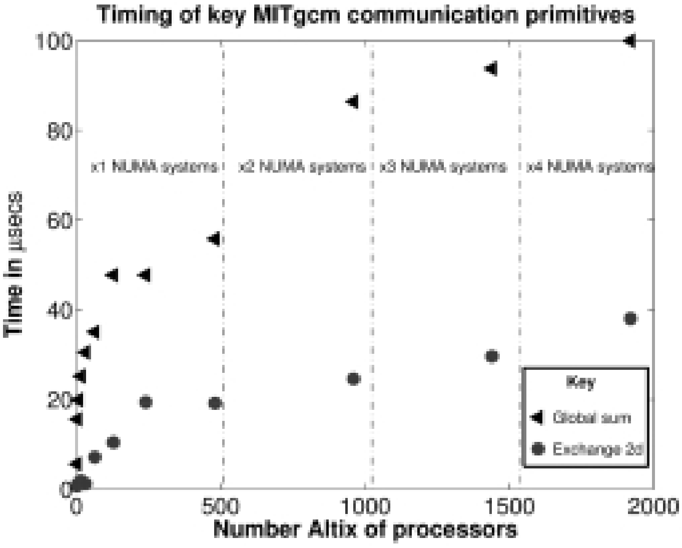

Fig. 2. Performance of key primitives used on the $\frac{1}{16}^{\circ}$ resolution simulation. The exchange times are for a sub-domain of size 96×136 with $O_x = O_y = 1$.

The grid spacing for this model configuration is $\frac{1}{16}^{\circ}$ longitudinally while latitude circle spacing is $\frac{1}{16}^{\circ}\cos(\mathrm{latitude})$. The model domain extends from $78.7°$S to $80°$N. This gives a global finite volume grid with grid cells that are 6.9 km × 6.9 km at the equator and 1.2 km × 1.2 km at the northern boundary. The number of surface grid cells is just over 25 million, and the configuration has 50 vertical levels, making the total number of cells in all three dimensions just over 1.25 billion. Each of the three dimensional fields that make up the description of the simulation domain and its time evolving state therefore requires 10 GB of storage. With the full set of runtime diagnostics that are required for a scientifically useful calculations there are about 180 active three dimensional fields, so that the memory footprint of our running calculation is around 1.8TB. Each time step of algorithm 6 requires computing approximately 3600 floating point operations for each grid point. Aggregated over a total of 1.25 billion grid cells a single time step entails approximately 4.5×10^{12} arithmetic operations.

At each timestep communication is necessary to update overlap regions in the sub-domains on which we compute (see Section 3). We use a sub-domain size of 96×136 with overlap regions $O_x = O_y = 3$. A total of 1920 sub-domains are needed to cover the full domain at this size so that a single three-dimensional *exchange* operation entails transferring a total of 1 GB of data. For the configuration used each timestep involves eighteen such operations, corresponding to steps 4: and 15: in algorithm 6.

Additionally an average of forty iterations in the conjugate gradient solver are required at each timestep. Each iteration entails a pair of two-dimensional exchange operations (each involving 20 MB of data transfer) and at least one global sum of an 8-byte floating point value from each sub-domain.

A large number of diagnostics are saved to allow analysis. In addition to basic state information high-frequency outputs of two-dimensional sea-surface elevation and bottom pressure that can be related to satellite measurements are saved along with three-dimensional product terms that cannot be derived from their separate components. The list of possible outputs is application dependent but at a minimum the I/O requirement is 300 GB per month of integration.

The 1.8 TB memory footprint of our calculation is larger than the main memory of a single 512-way box so for this study we examine processor counts of 960, 1440 and 1920 which correspond to running across two, three and four 512-way Altix boxes. We make use of a dedicated sets of CPUs and employ the Message Passing Toolkit MPI_DSM_DISTRIBUTE flags to ensure that when an MPI process is assigned to an Altix node, the CPU and memory on that node bind to that process.

Figure 2 shows the scaling behavior of two key communication primitives used in MITgcm, the *sum* operation and the two-dimensional form of the *exchange* operation. These times were obtained from an isolated kernel benchmark, taken from the MITgcm code, because the full, $\frac{1}{16}^{\circ}$, configuration will not fit on smaller numbers of processors. The times, shown in microsec-

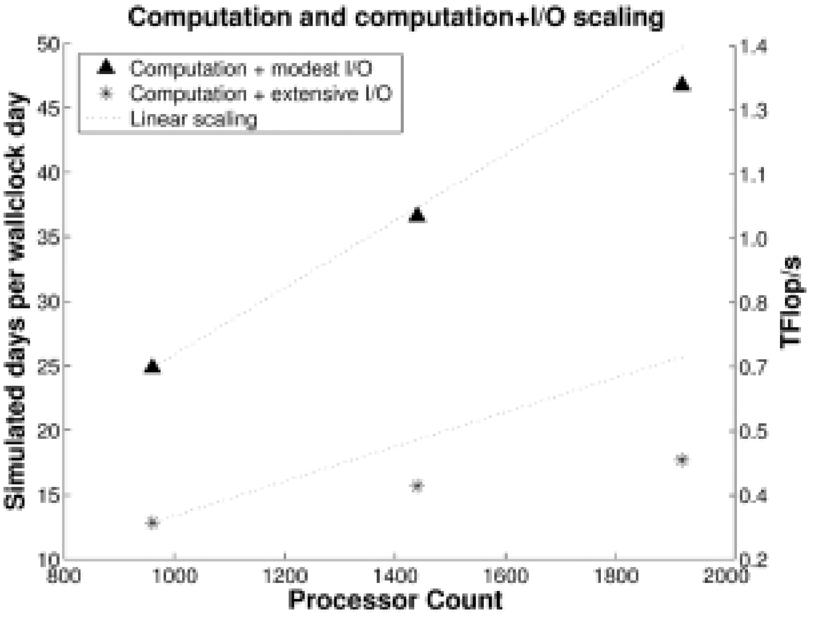

Fig. 3. Overall scaling and performance of the $\frac{1}{16}^\circ$ resolution simulation on 960, 1440 and 1920 processors.

onds, increase with processor count for both the *sum* operation and the two-dimensional *exchange*. However, it is encouraging to note that the increase in time when going between different numbers of Altix boxes is comparable to the increase associated with going from 1 to 512 CPUs. The *exchange* times stay relatively flat because although more data is exchanged between processors as the number of CPUs increases the available network bandwidth increases in proportion with CPU count.

Figure 3 shows the overall scaling and performance of a full simulation. Two scenarios are shown one with a modest I/O load (the upper line) and one with a very high I/O load (the lower line). The communication costs in Fig. 2 are only a small fraction of the total time in both cases. For the modest I/O case the peak performance at 1920 processors is sufficient to achieve 45 days of simulation per wall-clock day with an aggregate performance of 1.39 TFlop/s. The modest I/O case also scales well from 960 processors. For the intensive I/O scenario performance drops so that maximum throughput on 1920 processors is around 17 simulated days per wall-clock day with an aggregate performance of around 460 GFlop/s. For the intensive I/O scenario the scaling from 960 processors falls below the linear relation (the dotted line). However, at present, the MITgcm I/O routines being used have not been optimized for the Altix systems and so we expect the performance of the intensive I/O scenario to improve significantly. Overall these results are encouraging and are sufficient to enable us to proceed with some preliminary numerical experiments.

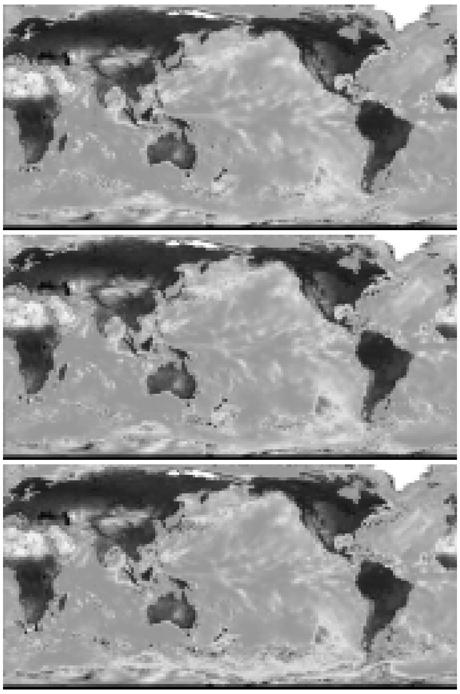

Fig. 4. One month sea-surface height differences. Upper $\frac{1}{4}^\circ$, middle $\frac{1}{8}^\circ$, lower $\frac{1}{16}^\circ$. Scale -0.5 m to 0.5 m.

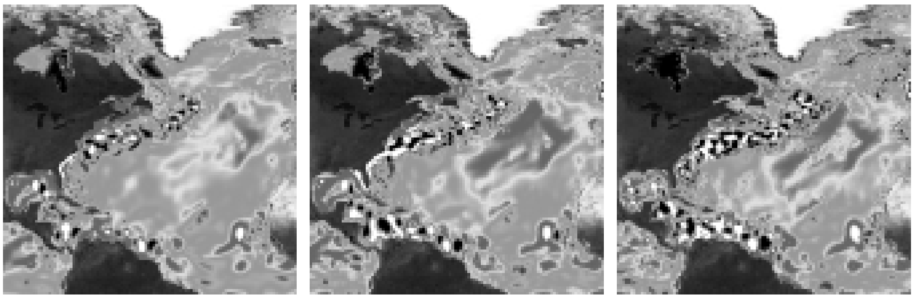

Fig. 5. Gulf Stream region sea-surface height difference plots at different resolutions for the same time period as Fig. 4. Left panel$\frac{1}{4}^{\circ}$, middle panel $\frac{1}{8}^{\circ}$, right panel $\frac{1}{16}^{\circ}$. Scale -0.125 m to 0.125 m.

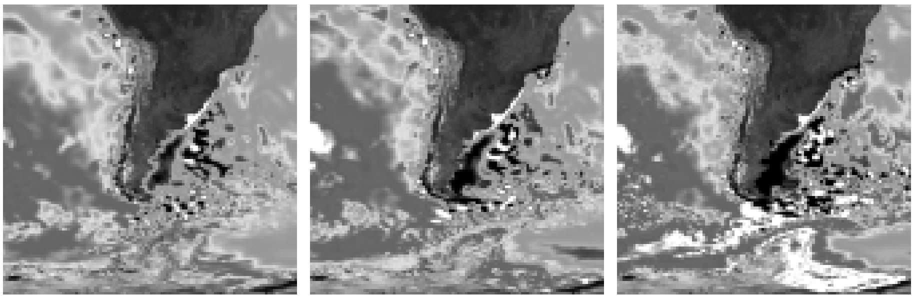

Fig. 6. Drake Passage region sea-surface height difference plots at different resolutions for the same time period as Fig. 4. Left panel$\frac{1}{4}^{\circ}$, middle panel $\frac{1}{8}^{\circ}$, right panel $\frac{1}{16}^{\circ}$. Scale -0.125 m to 0.125 m.

5. Preliminary results

Using the Altix configuration described earlier, we undertook a series of numerical simulations at $\frac{1}{4}^{\circ}$, $\frac{1}{8}^{\circ}$, and $\frac{1}{16}^{\circ}$ resolutions. The three cases are all configured as described in section 4 and, apart from differing resolutions, the model configurations are the same. Figures 4, 5 and 6 show significant changes in solution with resolution. Each plot shows the change in simulated sea-surface height over the same month. The plots capture changes due to eddy activity over a single month. Changes with resolution occur in the regions of the oceans where eddies are prevalent (such as the Gulf Stream, the Kurishio, the Aghullas, the Drake Passage and the Antarctic Circumpolar Current). For example in the Gulf Stream region the upper panel of figure 4 ($\frac{1}{4}^{\circ}$) shows a relatively small area of vigorous sea-surface height changes. The middle and lower panels ($\frac{1}{8}^{\circ}$ and $\frac{1}{16}^{\circ}$ respectively) show more extensive areas of changes but there are big shifts between the $\frac{1}{8}^{\circ}$ and the $\frac{1}{16}^{\circ}$ results.

Figure 5 and 6 are closeups of the Gulf Stream and Drake Passage. Key behaviors like how tightly waters “stick” to the coast, or how far energetic eddies penetrate the ocean interior change significantly between resolutions.

6. Conclusions

At first glance the three different resolution runs show significant differences. There does, however, seem to be smaller change between the $\frac{1}{8}^{\circ}$ and $\frac{1}{16}^{\circ}$ simulations. A next step is to undertake a fourth series of runs at even higher resolution, $\frac{1}{20}^{\circ}$ or $\frac{1}{32}^{\circ}$. Formally quantifying the changes between these runs would provide important information on whether ocean models are reaching numerically converged solutions.

Performance on the Altix shows it is well suited for addressing these questions. We monitored our code to be achieving about 722 Mflop/s/cpu when running on 1920 processors. This is 14% of the per CPU Top500 number achieved on the system [26]. Our code consists of predominantly BLAS1 class operations and cannot exploit the level of cache reuse that the Top500 Linpack benchmark achieves. The scaling we found across multiple Altix systems is encouraging and suggests that configurations that span eight or more Altix boxes, and that would therefore support $\frac{1}{20}^{\circ}$ and higher resolution, are today within reach.

Acknowledgements

This study is a contribution of the Consortium for Estimating the Circulation and Climate of the Ocean, Phase II (ECCO2) funded by the National Aeronautics and Space Administration (NASA).

Algorithms

Algorithm 1. MITgcm dynamical kernel algorithm

```
1. for $n = n_{\text{initial}}$ to $n_{\text{final}}$ do
2.     Finite volume calculations evaluate $n + \frac{1}{2}$ terms values using Adams-Bashforth extrapolation.
3.     Solve a two-dimensional implicit Helmholtz problem that is the finite volume discretization of Eq. (8) to yield $\eta^{n+1}$. This employs a preconditioned conjugate gradient solver [11,22].
4.     Update horizontal velocity, $\vec{\mathbf{u}}_h^{n+1}$, temperature, $\theta^{n+1}$ and salinity, $s^{n+1}$, according to Eqs (6) and (7).
5.     Vertically integrate finite volume forms of Eq. (9) and (10) to yield $p_{hyd}^{n+1}$ and $w^{n+1}$.
6. end for
```

Algorithm 2. MITgcm dynamical kernel decomposed computational procedure for a single time step n.

```
1.  for all subdomains of all processes do
2.      Compute terms corresponding to step 2: of algorithm 6.
3.  end for
4.  Neighbor subdomains exchange overlap regions over $N_z$ levels.
5.  while conjugate gradient solver for Eq. (8) is not converged do
6.      for all subdomains of all processes do
7.          Compute terms in the preconditioned conjugate gradient algorithm
8.          Neighbor subdomains exchange two-dimensional overlap regions for Helmholtz problem.
9.          All processes sum up terms in vector dot products to test for convergence and to update the conjugate gradient search directions.
10.     end for
11. end while
12. for all subdomains of all processes do
13.     Compute terms corresponding to steps 4: and 5: of algorithm 6.
14. end for
15. Neighbor subdomains exchange overlap regions over $N_z$ levels.
```

References

[1] A. Adcroft, C. Hill and J. Marshall, *Representation of topography by shaved cells in a height coordinate ocean model,* Monthly Weather Review, 1997, 2293–2315.

[2] C.W. Boening and R. Budich, Eddy dynamics in a primitive equation model: sensitivity to horizontal resolution and friction, *J Phys Oceanogr* **22** (1992), 361–381.

[3] G.A. Boughton, *Arctic Switch Fabric,* in Proceedings of the Second International Workshop on Parallel Computer Routing and Communication, volume 1417 of Lecture Notes in Computer Science, Springer-Verlag, Berlin (Germany), 1997, 65–74.

[4] K. Bryan and W. Holland, A high resolution simulation of the wind and thermohaline driven circulation of the North Atlantic Ocean, in: *Parameterization of small-scale processes,* P. Muller and D. Henderson, eds, Hawaii Institute of Geophysics Manoa, 1989, pp. 99–115.

[5] D. Chelton, R. DeSzoeke and M. Schalax. Geophysical variability of the first baroclinic rossby radius of deformation, *J Phys Oceanogr* **28** (1998), 433–460.

[6] B. Ciotti and B. Nelson, *STREAM benchmark of 512 CPU Altix,* http://www.cs.virginia.edu/stream/, 2003.

[7] T. Haine and J. Marshall, Gravitational, symmetric and baroclinic instability of the ocean mixed layer, *J Phys Oceanogr* **29** (1998), 634–658.

[8] P. Heimbach, C. Hill and R. Giering, Automatic generation of efficient adjoint code for a parallel navier-stokes solver, in: *Computational Science – ICCS 2002, volume 2331 of Lecture Notes in Computer Science,* J.J. Dongarra, P.M.A. Sloot and C.J.K. Tan, eds, Springer-Verlag, Berlin, Germany, 2002, pp. 1019–1028.

[9] C. Hill, A. Adcroft, D. Jamous and J. Marshall, *A strategy for terascale climate modeling,* in Proceedings of the Eight ECMWF Workshop on the Use of Parallel Processors in Meteorology, 1999.

[10] C. Hill, V. Bugnion, M. Follows and J. Marshall, Evaluating carbon sequestration efficiency in an ocean circulation model by ad-joint sensitivity analysis, *J Geophys Res* **109**(C11005), 2004. doi:10.1029/2002JC001598.

[11] C. Hill and J. Marshall, *Application of a Parallel Navier-Stokes Model to Ocean Circulation,* in Proceedings of Parallel Computational Fluid Dynamics: Implementations and Results Using Parallel Computers, 1995, 545–552.

[12] H.E. Hurlburt and P.J. Hogan, Impact of 1/8 to 1/64 resolution on gulf stream model-data comparisons in basin-scale subtropical atlantic ocean models, *Dyn Atmos and Oceans* **32** (2000), 283–329.

[13] S. Jarp, *A Methodology for using the Itanium 2 Performance Counters for Bottleneck Analysis,* HP Labs, 2002.

[14] H. Jones and J. Marshall, Restratification after deep convection, *J Phys Oceanogr* **27** (1997), 2276–2287.

[15] C.E. Leiserson, Fat-Trees: Universal Networks for Hardware-Efficient Supercomputing, *IEEE Transactions on Computers* **C34** (October 1985), 892–901.

[16] D. Lenoski, J. Laudon, K. Gharachorloo, W. Weber, A. Gupta, J. Hennessy, M. Horowitz and M. Lam, The Stanford Dash multiprocessor, *IEEE Computer* (March 1992), 63–79.

[17] M.E. Maltrud and J.L. McClean, An eddy resolving global 1/10 ocean simulation, *Ocean Modeling* **8** (2005), 31–54.

[18] J. Marotzke, R. Giering, K. Q. Zhang, D. Stammer, C. Hill and T. Lee, Construction of the Adjoint MIT Ocean General Circulation Model and Application to Atlantic Heat Transport Sensitivity, *J Geophys Res* **104**(C12) (1999), 29,529-29,549.

[19] J. Marshall, A. Adcroft, J.-M. Campin, C. Hill and A. White, Atmosphere-ocean modeling exploiting uid isomorphisms, *Monthly Weather Review* **132**(12) (2004), 2882–2894.

[20] J. Marshall, A. Adcroft, C. Hill, L. Perelman and C. Heisey, A finite-volume, incompressible navier stokes model for studies of the ocean on parallel computers, *J Geophys Res* **102**(C3) (1997), 5,753–5,766.

[21] J. Marshall, C. Hill, L. Perelman and A. Adcroft, Hydrostatic, quasihydrostatic and nonhydrostatic ocean modeling, *J Geophys Res* **102**(C3) (1997), 5,733–5,752.

[22] J. Marshall, H. Jones and C. Hill, Efficient ocean modeling using non-hydrostatic algorithms, *Journal of Marine Systems* **18** (1998), 115–134.

[23] Y. Masumoto et al., A fifty-year eddy-resolving simulaiton of the world ocean, *Journal of the Earth Simulator* **1** (April 2004), 35–56.

[24] J. McCalpin, *The STREAM benchmark web site,* http://www.cs.virginia.edu/stream/, 2005.

[25] J. McClean, P.M. Poulain, J.W. Pelton and M.E. Maltrud, Eulerian and Lagraangian statistics from surface drifters and a high-resolution POP simulation in the North Atlantic, *J Phys Oceanogr* **32** (2002), 2473–2491.

[26] H. Meuer and J. Dongarra, *The Top 500 benchmark web site,* http://www.top500.org, 2005.

[27] S. Neuner, Scaling Linux to New Heights: the SGI Altix 3000 System, *Linux Journal* (2003).

[28] A.J.G. Nurser and J.W. Zhang, Eddy-induced mixed layer shallowing and mixed layer/thermocline exchange, *J Geophys Res* **105** (2000), 21851–21868.

[29] A. Oschillies, Improved representation of upper-ocean dynamics and mixed layer depths in a model of the North Atlantic on switching from eddy-permitting to eddy-resolving grid resolution, *J Phys Oceanogr* **32** (August 2002), 2277–2298.

[30] A. Semtner and R. Chervin, Ocean general circulation from a global eddy-resolving model, *J Geophys Res* **97** (1992), 5493–5550.

[31] R.D. Smith, M.E. Maltrud, F. Bryan and M.W. Hecht, Numerical simulation of the North Atlantic at 1/10, *J Phys Oceanogr* **30** (2000), 1532–1561.

[32] D. Stammer, C. Wunsch, R. Giering, C. Eckert, P. Heimbach, J. Marotzke, A. Adcroft, C. Hill and J. Marshall, Volume, heat, and freshwater transports of the global ocean circulation 1993–2000, estimated from a general circulation model constrained by World Ocean Circulation Experiment (WOCE) data, *J Geophys Res* **108**(C1) (2003), 3007–3029.

[33] M. Woodacre, D. Robb, D. Roe and K. Feind, *The SGI Altix 3000 Global Shared-Memory Architecture,* White paper, SGI, Mountain View(CA), USA, 2004.