Scalability of Semi-Implicit Time Integrators for Nonhydrostatic Galerkin-based Atmospheric Models on Large Scale Cluster

James F. Kelly and Francis X. Giraldo
Department of Applied Mathematics
Naval Postgraduate School
Monterey, CA, USA
Email: jfkelly@nps.edu and fxgirald@nps.edu

Gabriele Jost
Texas Advanced Computing Center
University of Texas at Austin
Austin, TX, USA
Email: gjost@tacc.utexas.edu

***Abstract*—In this paper we describe the current status of our ongoing effort to optimize the efficiency of a novel application package for Nonhydrostatic Unified Model of the Atmosphere (NUMA) when running on large cluster architectures. The linear solver within a distributed memory paradigm is critical for overall model efficiency. The goal of this work-in-progress is to investigate the scalability of the model for different solvers and determine a set of optimal solvers to be used for different situations. We will describe our novel approach and demonstrate its scalability on a variety of clusters of multicore node clusters. We also present performance statistics to explain the scalability behavior.**

***Keywords*-atmospheric models, time intergrators, MPI, scalability, performance;**

I. Introduction

Current limited-area, or mesoscale, atmospheric models require modeling nonhydrostatic effects, while next-generation global models are moving toward the nonhydrostatic regime. The nonhydrostatic atmospheric models, which run at resolutions finer than 10 km, possess fast-moving acoustic and gravity waves. These fast moving waves require a stringent CFL condition if the equations are discretized explicitly. To mitigate this problem, semi-implicit (SI) time-integrators, based on a Schur complement technique, have been developed for our Nonhydrostatic Unified Model for the Atmosphere (NUMA). SI time-integrators, which discretize the linear, fast moving waves implicitly, while discretize the slower dynamical processes explicitly, require a linear solve at each time-step. These time-integrators, while only conditionally stable, require less computational effort than fully-implicit time integrators, which are unconditionally stable. However, engineering efficient semi-implicit time-integrators which scale to tens of thousands of cores remains a significant challenge.

The efficiency of this linear solve within a distributed memory paradigm is critical for overall model efficiency. Therefore, we analyze the performance of two different semi-implicit time-integrators within distributed memory architectures. The goal of our work-in-progress is to compare the scalability of these solvers, thus determining the optimal time-integrator as a function of both the test problem and the parallel architecture. The rest of the paper is structured as follows: In Section 2 we describe the NUMA application and our approach to parallelization. In Section 3 we discuss in more detail the semi-implicit time-integrators which we employ and how they effect the efficiency of NUMA model. In Section 4 we describe the systems and test cases used in our evaluation. Section 5 presents presents scaling results and corresponding performance statistics. We draw our conclusions in Section 6 and also describe our future plans.

II. The Nonhydrostatic Unified Model of the Atmosphere (NUMA) Model

A. Governing Equations

NUMA utilizes the fully compressible, non-hydrostatic Euler equations in non-conservative form in a global setting. These equations have previously been considered within a semi-implicit framework for 2D flows [2] and more recently for 3D flows [5] in Cartesian geometries. In the present study, we consider three-dimensional flow (x-y-z) subject to gravitational and Coriolis forces

$$\frac{\partial \rho'}{\partial t} + \mathbf{u}\cdot\nabla\rho' + \nabla_V \rho_0 \cdot \mathbf{u} + (\rho' + \rho_0)\,\nabla\cdot\mathbf{u} = 0 \quad (1a)$$

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u}\cdot\nabla\mathbf{u} + \frac{1}{\rho' + \rho_0}\nabla P' + \frac{\rho'}{\rho' + \rho_0} g\hat{\mathbf{r}} + \mathbf{f}\times\mathbf{u} = 0 \quad (1b)$$

$$\frac{\partial \theta'}{\partial t} + \mathbf{u}\cdot\nabla\theta' + \nabla_V \theta_0 \cdot \mathbf{u} = 0 \quad (1c)$$

where the unknown variables are $(\rho', \mathbf{u}^{\mathcal{T}}, \theta')$, where $\rho' = \rho - \rho_0$ is a density perturbation, $\mathbf{u} = (u, v, w)$ is velocity, and $\theta' = \theta - \theta_0$ is potential temperature perturbation. In Eq. (1), ∇_V denotes a vertical gradient and $\hat{\mathbf{r}}$ is a unit radial vector. The reference states ρ_0 and θ_0 are hydrostatic and time-independent. Defining a solution vector $\mathbf{q} = (\rho', \mathbf{u}^{\mathcal{T}}, \theta')$, Eq. (1) is written in condensed form as

$$\frac{\partial \mathbf{q}}{\partial t} = S(\mathbf{q}) \quad (2)$$

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2011		**00-00-2011 to 00-00-2011**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Scalability of Semi-Implicit Time Integrators for Nonhydrostatic Galerkin-based Atmospheric Models on Large Scale Cluster	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Naval Postgraduate School,Department of Applied Mathematics,Monterey,CA,93943**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
2011 IEEE International Conference on Cluster Computing (CLUSTER), 26-30 Sep, Austin, TX.

14. ABSTRACT
In this paper we describe the current status of our ongoing effort to optimize the efficiency of a novel application package for Nonhydrostatic Unified Model of the Atmosphere (NUMA) when running on large cluster architectures. The linear solver within a distributed memory paradigm is critical for overall model efficiency. The goal of this work-in-progress is to investigate the scalability of the model for different solvers and determine a set of optimal solvers to be used for different situations. We will describe our novel approach and demonstrate its scalability on a variety of clusters of multicore node clusters. We also present performance statistics to explain the scalability behavior.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**4**	

where the source term $S(\mathbf{q})$ is a right-hand side which is discretized by the spectral element method.

B. Spectral Element Discretization

The spectral element method [1] decomposes the spatial domain $\Omega \subset \mathcal{R}^3$ into N_e disjoint elements Ω_e via

$$\Omega = \bigcup_{e=1}^{N_e} \Omega_e \tag{3}$$

where we choose Ω_e as hexahedra, which provides 1) simple grid generation and 2) efficient (fast) evaluation of the necessary differentiation and integration operators. Within each element Ω_e, a finite-dimensional approximation $\mathbf{q}_N$ is formed by expanding $\mathbf{q}(\mathbf{x}, t)$ in basis functions $\psi_j(\mathbf{x})$

$$\mathbf{q}_N(\mathbf{x}, t) = \sum_{j=1}^{M_N} q_j(t)\psi_j(\mathbf{x}) \tag{4}$$

where $M_N = (N+1)^3$ is the number of nodes per element and N is the order of the basis functions. The discrete solution $\mathbf{q}_N$ is assumed continuous across inter-element boundaries. Basis functions are constructed as tensor products of Lagrange polynomials $\psi_i(\mathbf{x}) = h_\alpha(\xi) \otimes h_\beta(\eta) \otimes h_\gamma(\zeta)$., where $h_\alpha(\xi)$ is the Lagrange polynomial associated with the Legendre-Gauss-Lobatto (LGL) points ξ_i and (ξ, η, ζ) are functions of the physical variable $\mathbf{x}$. Approximating the prognostic vector $\mathbf{q}(\mathbf{x}, t)$ by a finite-dimensional approximation $\mathbf{q}_N$ in Eq. (4), multiplying by a basis function ψ_i and integrating over the domain Ω yields the weak form

$$\int_\Omega \psi_I \frac{\partial \mathbf{q}_N}{\partial t}\, d\Omega = \int_\Omega \psi_I S(\mathbf{q}_N)\, d\Omega \tag{5}$$

where we have replaced the index i with I to emphasize that Eq. (5) is now a global representation. Applying the Galerkin expansion given by Eq. (4) to Eq. (5) yields the matrix-vector equation

$$\frac{\partial q_I}{\partial t} = M_{IJ}^{-1} S(q_I) \tag{6}$$

where the mass-matrix $M_{IJ} = \int_\Omega \psi_I \psi_J\, d\Omega$ is diagonal if the interpolation and integration points are co-located. This approximation is valid for $N \geq 4$ while incurring a small error in integration [4]. Denoting the right-hand size (RHS) of Eq. (6) by $R_I(q_I)$, Eq. (6) is expressed as

$$\frac{\partial q_I}{\partial t} = \bigwedge_{e=1}^{N_e} R_i^{(e)}\left(q_i^{(e)}\right). \tag{7}$$

where $\bigwedge_{e=1}^{N_e}$ denotes the global assembly, or direct stiffness summation (DSS) operator that maps local, element-wise coordinates i to global coordinates I. Eq. (7) forms the core of the spectral element method, allowing local, element-wise information $q_i^{(e)}$ to propagate to adjacent elements via the DSS operator.

C. Parallelization

III. Impact of Time Interators on Model Efficiency

The fast-moving acoustic and gravity waves in Eq. (1) make explicit time-integration unfeasible do to the stringent CFL restriction. Therefore, we have developed both 3D and 1D semi-implicit time-integrators [3] which allow much larger time-steps. Both time-integrators utilize a psuedo-Helmholtz decomposition, which reduces the poorly conditioned monolithic $5N_p \times 5N_p$ system matrix to a well-conditioned $N_p \times N_p$ matrix.

A. 3D Semi-Implicit Time-Integrator

In the 3D Semi-Implicit (3D-SI) time-integrator, both the horizantally and vertically propagating acoustic and gravity waves are discretized implicitly, resulting in following linear system which must be solved at each time-step:

$$\mathcal{H}_{3D} P_{tt} = R_{tt} \tag{8}$$

where $\mathcal{H}_{3D} = \mathcal{H}\left(\mathbf{D}, \mathbf{D}^{\mathcal{T}}\right)$ is a linear, pseudo-Helmholtz operator consisting of gradient $\mathbf{D}$ and divergence $\mathbf{D}^{\mathcal{T}}$ operators operating on the discretized pressure P_{tt} and R_{tt} is an effective source term. Eq. (8) is solved iteratively using a Krylov sub-space method, which requires global reductions.

B. 1D Semi-Implicit Time-Integrator

In contrast, the 1D Semi-Implicit (1D-SI) time-integrator, only the vertically propagating acoustic and gravity waves are discretized implicitly. Hence, the global $N_p \times N_p$ problem is reduced into N_H, independent linear systems of size $N_V \times N_V$ which are solved at each time-step

$$\mathcal{H}_{1D} P_{tt}^{(i)} = R_{tt}^{(i)} 1 \leq i \leq N_H. \tag{9}$$

where N_H is the number of horizantal grid points, N_V is the number of vertical grid points, and $N_p = N_H * N_V$. In Eq. (9), $\mathcal{H}_{1D} = \mathcal{H}\left(\mathbf{D}_{1D}, \mathbf{D}_{1D}^{\mathcal{T}}\right)$ is a linear, pseudo-Helmholtz operator consisting of vertical gradient $\mathbf{D}$ and divergence $\mathbf{D}^{\mathcal{T}}$ operators. When combined with a horizantol domain-decomposition, no inter-processor communication is required. In typical global NWP problems, $N_V \ll N_H$.

IV. Scalability and Performance Performance Analysis

In this section we present timings and performance analysis statistics for the 1D and 3D integrators.

A. Test Systems

We conducted scalability experiments of the following systems:

- The Constellation Linux Cluster Ranger is located at the Texas Advanced Computing Center (TACC). The Ranger system consists of of 3,936 16-way SMP compute nodes. Each nodes has 4 2.3 GHz AMD Quadcore Opteron™providing 15,744 AMD Opteron processors

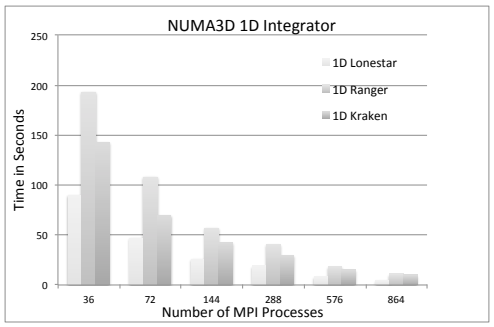

Figure 1. Scalability of the 1D Integrator

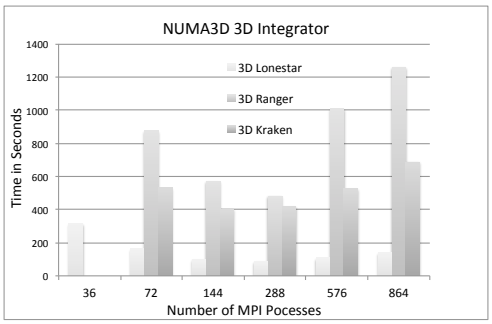

Figure 2. Scalability of the 3D Integrator expressed

for a total of 62,976 compute cores, 123 TB of total memory and 1.7 PB of raw global disk space. It has a theoretical peak performance of 579 TFLOPS. All Ranger nodes are interconnected using InfiniBand technology in a full-CLOS topology providing a 1GB/sec point-to-point bandwidth. For further details see [6] .

- The Lonestar Linux Cluster is also located at TACC. It consists of 1,888 12way nodes, with two 6-Core Intel Xeon Intel Hexa-Core 64-bit 3.33GHz Westmere^{TM}processors on a single as an SMP unit. This provides for a total of 22,656 cores. It is configured with 44 TB of total memory and 276TB of local disk space. The peak performance is 302 TFLOPS. Nodes are interconnected with InfiniBand technology in a fat-tree topology with a 40Gbit/sec point-to-point bandwidth. For further details see [7].
- The Cray XT5 system Kraken is located at the National Institute for Computational Sciences (NICS) at the University of Tennesee in Oakridge. The system runs the Cray Linux Environment (CLE) 2.2. It consists of 12way 9408 compute nodes, each node containing two 2.6GHz 6 core AMD Opteron Istambul^{TM}prcessors, providing a total of 112,896 compute cores. The nodes are connected by the Cray SeaStar2+ router. For more information see [8].

B. Scalability Results and Analysis

We ran our test case for about 1400 time steps. Figures 1 and 2 show the scalability of the 1D and 3D integrators expressed as decrease in elapsed execution time with increasing number of MPI processes. The timings indicate that the 1D Integrator scales well on all systems and yields shorter execution time than the 3D integrator.

The timings indicate that the 3D integrator does not scale beyond 144 processes on all systems. Using more than 288 MPI processes actually yields negative scalability: that is an increase in execution time when increasing the number of MPI processes. The reason for the poor scalability is the excessive time spent in global reductions, which is required by the GMRES iterative solver. We employed the TAU [9] performance analysis tool on the Lonestar system for a run with 864 MPI processes. The timing statistics displayed in Figure 3 show, that 77% of the time is spent in global communication caused by a huge number of calls to MPI_Allreduce short message size. The test case under consideration is therefore a good candidate for the usage of the 1D integrator, yielding better scalability as well as a shorter execution time than the 3D integrator.

V. Conclusions and Future Work

From the experiments and that we have conducted so far we conclude that our implementation of the 3D integrator excessive MPI communication time due to global reductions, at least for certain sets of input data. In our future work we will focus on issues: We will investigate possible optimizations for the 3D integrator, including the usage of other Krylov subspace methods. Second, we will conduct many more experiments on different types of input data sets. Our goal is to derive characteristics of the application runtime behavior in order to choose the most efficient solver.

Acknowledgment

The experiments on the Kraken Cray XT5 system were conducted via the TeraGrid. We would like to thank the Texas Advanced Computing Center (TACC) of the University of Texas at Austin which supported us by providing compute time and valuable support for our experiments on the Ranger and the Lonestar Clusters.

References

[1] M. O. Deville, P. F. Fischer, and E. H. Mund. *High-Order Methods for Incompressible Fluid Flow*. Cambridge University Press, 2002.

[2] F. X. Giraldo and M. Restelli and M. Läuter. Semi-Implicit formulations of the Navier-Stokes equations: Appliciations to nonhydrostatic atmospheric modeling. *SIAM Journal on Scientific Computing*, submitted, 2009.

[3] F. X. Giraldo and J. F. Kelly. Development of the non-hydrostatic unified model for the atmosphere: Combined mesoscale/global implicit-explicit formulations. *SIAM J. Sci. Comp.*, in preparation:1–20, 2011.

%Time	Exclusive msec	Inclusive total msec	#Call	#Subrs	Inclusive usec/call	Name
100.0	9	2:16.913	1	19	136913239	NUMA3D
93.9	32	2:08.629	1	4322	128629276	NUMA3D => TIME_LOOP
93.9	32	2:08.629	1	4322	128629276	TIME_LOOP
93.5	54	2:07.988	1439	11516	88943	TIME_LOOP => TI_BDF2
93.5	54	2:07.988	1439	11516	88943	TI_BDF2
92.9	1,046	2:07.243	1442	1.97725E+06	88241	SOLVE_GMRES_SCHUR
92.8	1,045	2:06.991	1439	1.97163E+06	88250	TI_BDF2 => SOLVE_GMRES_SCHUR
77.0	1:45.438	1:45.438	1.67138E+06	0	63	MPI_Allreduce()
75.4	2,010	1:43.249	1.60276E+06	1.60276E+06	64	GLSC2_MPI
75.4	2,010	1:43.249	1.60276E+06	1.60276E+06	64	SOLVE_GMRES_SCHUR => GLSC2_M
73.9	1:41.238	1:41.238	1.60276E+06	0	63	GLSC2_MPI => MPI_Allreduce()

USER EVENTS Profile :NODE 1, CONTEXT 0, THREAD 0

NumSamples	MaxValue	MinValue	MeanValue	Std. Dev.	Event Name
1.671E+06	8	8	8	0	Message size for all-reduce
5	4	4	4	0	Message size for broadcast
3	40	32	37.33	3.771	Message size for reduce
8	4	4	4	0	Message size for scatter

Figure 3. Timing Statistics for a run on 288 MPI Processes on Lonestar using the 3D Integrator

[4] Francis Xavier Giraldo. The Lagrange-Galerkin spectral element method on unstructured quadrilateral grids. *J. Comp. Phys.*, 147(1):114–146, NOV 20 1998.

[5] J. F. Kelly and F. X. Giraldo. Development of the nonhydrostatic unified model for the atmosphere (NUMA): Limited-area mode. *J. Comp. Phys.*, submitted:1–20, 2011.

[6] Texas Advanced Computing Center, The University of Texas at Austin, *Ranger User Guide*, http://services.tacc.utexas.edu/index.php/ranger-user-guide, Austin, TX, 2011.

[7] Texas Advanced Computing Center, The University of Texas at Austin, *Lonestar User Guide*, http://services.tacc.utexas.edu/index.php/lonestar-user-guide, Austin, TX, 2011.

[8] National Institute of Computational Science, http://www.nics.tennessee.edu/computing-resources/kraken, *Computing Resources*, Oak Ridge, TN, 2011.

[9] S. Shende and A. D. Malony *The TAU Parallel Performance System* , International Journal of High Performance Computing Applications, Volume 20 Number 2 Summer 2006. Pages 287-311