Leading Computational Methods on the Earth Simulator and IBM Power3

Project Leaders

Horst D. Simon National Energy Research Scientific Computing Center

Leonid Oliker National Energy Research Scientific Computing Center

Shigemune Kitawaki The Earth Simulator Center, Japan Agency for Marine-Earth Science and Technology

Authors

Horst D. Simon *1, Leonid Oliker *1, Andrew Canning *1, Jonathan Carter *1, Michael Wehner *1, Stephane Ethier *2, Bala Govindasamy *3, Art Mirin *3, David Parks *4, 5, Shigemune Kitawaki *5 and Yoshinori Tsuda *5

*1 Lawrence Berkeley National Laboratory

*2 Princeton Plasma Physics Laboratory

*3 Lawrence Livermore National Laboratory

*4 NEC Solutions America

*5 The Earth Simulator Center, Japan Agency for Marine-Earth Science and Technology

This work explores four applications from leading scientific domains in the areas of atmospheric modeling (FVCAM), magnetic fusion (GTC), plasma physics (LBMHD3D), and material science (PARATEC). We compare performance between the vector-based Earth Simulator, and superscalar-based IBM Power3. Overall results show that the ES attains unprecedented aggregate performance across our evaluated application suite, demonstrating the tremendous potential of modern parallel vector systems.

Keywords: Performance evaluation, vectorization, scientific computing

1. Introduction

Applications scientists have observed a frustrating trend of stagnating application performance despite dramatic increases in claimed peak performance of high-performance computing (HPC) systems. This effect has been widely attributed to systems composed of commodity components, whose architectural designs are unbalanced and inefficient for large-scale scientific computations. The recent development of parallel vector systems offers the potential to bridge this performance gap for a significant number of scientific codes, and to increase computational power substantially. In order to quantify what a vector capability entails for scientific communities that rely on modeling and simulation, it is critical to evaluate it in the context demanding computational algorithms. This work build on our previous effort [5] and compares performance of the cacheless vector Earth Simulator (ES) versus the superscalar cache-based IBM Power3 located at NERSC [2]. Performance results are presented from several key scientific computing domains including atmospheric modeling, astrophysics, material science and magnetic fusion.

2. FVCAM

The Community Atmosphere Model (CAM) is the atmospheric component of the flagship Community Climate System Model (CCSM3.0). Developed at the National Center for Atmospheric Research (NCAR), the CCSM3.0 is extensively used to study climate change. The CAM application is an atmospheric general circulation model (AGCM) and can be run either coupled within CCSM3.0 or in a standalone mode driven by prescribed ocean temperatures and sea ice coverages [1]. AGCMs are key tools for weather prediction and climate research. They also require large computing resources: even the largest current supercomputers cannot keep pace with the desired increases in the resolution of these models.

AGCMs generally consist of two distinct sections, the 'dynamical core' and the 'physics package'. The dynamical core approximates a solution to the Navier-Stokes equations suitably expressed to describe the dynamics of the atmosphere. The physics package calculates source terms to these equations of motion that represent unresolved or external physical phenomena. These include turbulence, radiative

transfer, boundary layer effects, clouds, etc. The dynamical core of CAM was constructed with two very different methodologies to solve the equations of motion. The default method, known as the spectral transform method, exploits spherical harmonics to map a solution onto the sphere. An alternate formulation, based on a finite volume methodology is also supplied. This option, referred as FVCAM, is based on a regular latitude-longitude mesh and conserves certain higher order moments.

2.1 Experimental Results

Table [1] shows a direct comparison of FVCAM timing results obtained on the ES and Power3. Processor configurations were identically maintained up to 896 processors using a 0.5°x0.625° horizontal mesh, also know as the D grid. To eliminate the costs associated with initialization, two integrations were performed and the timing results subtracted. By measuring the time spent in the routine STEPON, we can determine the cost of integrating the model in the absence of I/O.

The performance results shown in Table [1] reveal that the balance between communication and computation as achieved by FVCAM is very different on each of these machines. At the low processor configurations in each of the three vertical discretizations 1,4,7, the ES achieves a significantly better percent of the peak performance than does the Power3. However, at the higher processor counts for each vertical discretization, the percent of peak performance is essentially the same for the two machines. There are at least two reasons for this. The first is that vector lengths on the ES shorten as the processor count increases. The second is the difference between the sustained computational rates achieved relative to the communication rates: Although the ES communications are faster than the Power3, the effective computational speed sustained between communication phases is faster yet, causing the communication to eventually become more significant on the ES than on the Power3.

Table 1 FVCAM results on Power3 and ES

D Grid		Power3		ES		
De-comp	P	Gflp/ Proc	% Pk	Gflp/ Proc	% Pk	Spd up
1D	32	0.11	7.2	1.18	13.8	10.9
	64	0.11	7.0	0.93	11.7	8.9
	128	0.10	6.4	0.66	8.3	6.9
	256	0.09	5.7	0.47	5.9	5.6
2D 4-Vert	128	0.09	5.9	0.77	9.6	8.7
	256	0.08	5.5	0.63	7.8	7.8
	512	0.07	4.9	0.40	4.9	5.4
2D 7-Vert	336	0.07	4.8	0.51	6.4	7.0
	448	0.07	4.6	0.51	6.4	7.5
	672	0.06	4.2	0.44	5.5	7.0
	896	0.06	3.8	0.33	4.2	5.9

3. GTC

GTC is a 3D particle-in-cell code used for studying turbulent transport in magnetic fusion plasmas [4]. The simulation geometry is that of a torus, which is the natural configuration of all tokamak fusion devices. As the charged particles forming the plasma move within the externally-imposed magnetic field, they collectively create their own self-consistent electrostatic (and electromagnetic) field that quickly becomes turbulent under driving temperature and density gradients. Waves and particles interact self-consistently with each other, exchanging energy that grows or damps their motion or amplitude. The particle-in-cell (PIC) method describes this complex phenomenon by solving the 5D gyro-averaged kinetic equation coupled to the Poisson equation.

GTC was originally optimized for superscalar SMP-based architectures by utilizing two levels of parallelism: a one-dimensional MPI-based domain decomposition in the toroidal direction, and a loop-level work splitting method implemented with OpenMP. However, the mixed-mode GTC implementation is poorly suited for vector platforms due to memory constraints and the fact that vectorization and thread-based loop-level parallelism compete directly with each other. As a result, previous vector experiments [5] were limited to 64-way parallelism – the optimal number of domains in the 1D toroidal decomposition. Note that the number of domains (64) is not limited by the scaling of the algorithm but rather by the physical properties of the system, which features a quasi two-dimensional electrostatic potential when put on a coordinate system that follows the magnetic field lines. GTC uses such a coordinate system and increasing the number of grid points in the toroidal direction does not change the results of the simulation.

To increase GTC's concurrency in pure MPI mode, a third level of parallelism was recently introduced. Since the computational work directly involving the particles accounts for almost 85% of the overhead, the updated algorithm splits the particles between several processors within each domain of the 1D spatial decomposition. Each processor then works on a subgroup of particles that span the whole volume of a given domain. This allows us to divide the particle-related work between several processor and, if needed, to considerably increase the number of particles in the simulation. The update approach maintains a good load balance due to the uniformity of the particle distribution.

3.1 Experimental Results

For this performance study, we keep the grid size constant but increase the total number of particles so as to maintain the same number of particles per processor, where each processor follows about 3.2 million particles. Table [2] shows the performance for the Power3 and ES. The first striking difference from the previous GTC vector study [5],

Table 2 GTC results on Power3 and ES

P	Part/ Cell	Power3		ES		
		Gflp/ Proc	% Pk	Gflp/ Proc	% Pk	Spd up
64	100	0.14	9.3	1.60	20.0	11.4
128	200	0.14	9.3	1.56	19.5	11.1
256	400	0.14	9.3	1.55	19.4	11.1
512	800	0.14	9.4	1.53	19.1	11.0
1024	1600	0.14	8.7	1.88	23.5	13.4
2048	3200	0.13	8.4	1.82	22.7	14.0

Table 3 LBMHD3D results on Power3 and ES

P	Grid Size	Power3		ES		
		Gflp/ Proc	% Pk	Gflp/ Proc	% Pk	Spd up
16	256	0.14	9.3	5.50	68.7	39.2
64	256	0.15	9.7	5.25	65.6	35.1
256	512	0.14	9.1	5.45	68.2	38.9
512	512	0.14	9.4	5.21	65.1	37.2
1024	1024			5.44	68.0	
2048	2048			5.41	67.6	

is the considerable increase in concurrency. The new particle decomposition algorithm allowed GTC to efficiently utilize 2,048 processors (comp red with only 64 using the previous approach), although this is not the limit of its scalability. With this new algorithm in place, GTC fulfilled the very strict scaling requirements of the ES and achieved an unprecedented 3.7 Tflop/s on 2,048 processors. Additionally, the Earth Simulator sustains a significantly higher percentage of peak (24%) compared with other platforms. The Power3, on the other hand, achieves only 8.4% of peak, running about 14X slower than the ES. The relatively poor scalar performance is due to the irregularity of the data access patterns, obviating effective cache utilization.

4. LBMHD-3D

Lattice Boltzmann methods (LBM) have proved a good alternative to conventional numerical approaches for simulating fluid flows and modeling physics in fluids. The basic idea of the LBM is to develop a simplified kinetic model that incorporates the essential physics, and reproduces correct macroscopic averaged properties. Recently, several groups have applied the LBM to the problem of magneto-hydrodynamics with promising results [6]. As a development of previous 2D codes, LBMHD3D simulates the behavior of a three-dimensional conducting fluid evolving from simple initial conditions through the onset of turbulence. The 3D spatial grid is coupled to via a 3DQ27 streaming lattice and block distributed over a 3D Cartesian processor grid. Each grid point is associated with a set of mesoscopic variables, whose values are stored in vectors proportional to the number of streaming directions – in this case 27 (26 plus the null vector).

4.1 Experimental Results

Table [3] presents LBMHD performance on the Power3 and ES. Observe that the vector architecture clearly outperform the scalar systems by a significant factor. The ES, sustains the highest fraction of peak across all architectures to date — an amazing 68% even at the highest 2048-processors concurrencies. Further experiments on the ES on 4800 processors attained an unprecedented aggregate performance of over 26 Tflop/s. However, the Power3 only achieves 140 Mflop/s (8.4% of peak), about 39X slower than the ES. The low performance of the superscalar system is mostly due to limited memory bandwidth. LBMHD has a low computational intensity – about 1.5 FP operations per data word of access – making it extremely difficult for the memory subsystem to keep up with the arithmetic units. Vector systems are able to address this discrepancy through a superior memory system and support for deeply pipelined memory fetches.

5. PARATEC

PARATEC (PARAllel Total Energy Code[3] performs ab-initio quantum-mechanical total energy calculations using pseudopotentials and a plane wave basis set. The pseudopotentials are of the standard norm-conserving variety. Forces can be easily calculated and used to relax the atoms into their equilibrium positions. PARATEC uses an all-band conjugate gradient (CG) approach to solve the Kohn-Sham equations of Density Functional Theory (DFT) and obtain the ground-state electron wavefunctions. DFT is the most commonly used technique in materials science, having a quantum mechanical treatment of the electrons, to calculate the structural and electronic properties of materials. Codes based on DFT are widely used to study properties such as strength, cohesion, growth, magnetic, optical, and transport for materials like nanostructures, complex surfaces, and doped semiconductors.

5.1 Experimental Results

Table [4] presents performance data for 3~CG steps of a 488 atom CdSe (Cadmium Selenide) quantum dot and a standard LDA run of PARATEC with a 35 Ry cut-off using norm-conserving pseudopotentials. CdSe quantum dots are luminescent in the optical range at different frequencies depending on their size and can be used as electronic dye tags by attaching them to organic molecules. They represent a nanosystem with important technological applications and the understanding of their properties and synthesis through first principles simulations represents a challenge for large-scale parallel computing in terms of computer resources and code development. This 488-atom system is, to the best of

Table 4 PARATEC results on Power3 and ES

P	Power3		ES		
	Gflp/ Proc	% Pk	Gflp/ Proc	% Pk	Spd up
128	0.93	62.2	5.12	64.0	64.0
256	0.85	56.7	4.97	62.1	62.1
512	0.73	48.8	4.36	54.5	54.5
1024	0.60	39.8	3.64	45.5	45.5
2048			2.67	33.4	33.4

our knowledge, the largest physical system (number of real space grid points) ever run with this code. Previous vector results for PARATEC [5] examined smaller physical systems at lower concurrencies.

PARATEC runs at a high percentage of peak on both superscalar and vector-based architectures due to the heavy use of the computationally intensive FFTs and BLAS3 routines, which allow high cache reuse and efficient vector utilization. The main limitation to scaling PARATEC to large numbers of processors, is the distributed transformation during the parallel 3D~FFTs which requires global interprocessor communications. Table [4] shows that PARATEC achieves unprecedented performance on the ES system, sustaining 5.5~Tflop/s for 2048 processors. The declining performance at higher concurrencies is caused by the increased communication overhead of the °due to the decreasing vector length of this fixed-size problem. Note that the Power3 system obtains a high fraction of peak (40% at 1024 processors), and is therefore only 6X slower than the ES platform.

6. Summary

This study examined four diverse scientific applications on the vector-based ES and superscalar Power3 platforms. Our work makes several significant contributions. We are the first to present vector results of the Community Atmosphere Model using the finite-volume solver in the dynamics phase of the calculation. Results on a 0.5°x0.625 (D) grid, show that the ES performance at high concurrency is sufficiently fast to practically conduct this high-fidelity simulation. We also presented a new parallel decomposition parallelization for the GTC magnetic fusion simulation. This new approach allowed scalability to 2048 processors on the ES (compared to only 64 using the previous code version), opening the door to a new set of high-phase space-resolution simulations, that to date have not been possible. Next we presented, LBMHD3D: a 3D version of a lattice Bolzmann magneto-hydrodynamics application used to study the onset evolution of plasma turbulence. The ES showed unprecedented LBMHD3D performance, achieving over 68% of peak for a total of 26Tflop/s on 4800 processors. Finally, we investigated performance of the PARATEC application, using the largest cell size atomistic simulation ever run with this material science code. Results on 2048 processors of the ES show the highest aggregate performance to date, allowing for high-fidelity simulations that hitherto have not been possible due to computational limitations.

Overall results show that the ES achieved the highest aggregate performance on any architecture tested to date across our full application suite, demonstrating the tremendous potential of modern parallel vector systems.

Acknowledgements

The authors would like to gratefully thank: the staff of the Earth Simulator Center, especially Dr. T. Sato, S. Kitawaki and Y. Tsuda, for their assistance during our visit. All authors from LBNL were supported by the Office of Advanced Scientific Computing Research in the Department of Energy Office of Science under contract number DE-AC03-76SF00098. Dr. Ethier was supported by the Department of Energy under contract number DE-AC020-76-CH03073.

References

1) CAM: Community Atmosphere Model http://www.cgd.ucar.edu/csm/models.atm-cam
2) National Energy Scientific Computing Center. http://www.nersc.gov
3) PARAllel Total Energy Code. http://www.nersc.gov/projects/paratec
4) Z. Lin, S. Ethier, T. S. Hamh, and W. M. Tang, “Size Scaling of turbulent transport in magnetically confined plasmas”, Phys. Rev. Lett., vol. 88, pp. 195004 (2002).
5) L. Oliker, A. Canning, J. Carter, J. Shalf. And S. Ethier, “Scientific Computations on Modern Parallel Vector Systems”, Proc. SC2004: High performance computing, networking, and storage conference, 2004.
6) P. Pavlo, G. Vahala, and L. Vahala, “Higher Order Isotropic Velocity Grids in Lattice Methods”, Phys. Rev. Lett., vol. 80, pp. 3960, (1998).

地球シミュレータ及びIBMPower3上の先導的計算手法

プロジェクトリーダー

Horst D. Simon　国立エネルギー研究科学コンピューティングセンター(NERSC)

Leonid Oliker　国立エネルギー研究科学コンピューティングセンター(NERSC)

北脇　重宗　海洋研究開発機構　地球シミュレータセンター

著者

Horst Simon*1，Leonid Oliker*1，Andrew Canning*1，Jonathan Carter*1，Michael Wehner*1，Stephane Ethier*2，Bala Govindasamy*3，Art Mirin*3，David Parks*4，北脇　重宗*5，津田　義典*5

＊1　ローレンス バクレー国立研究所(米国)

＊2　プリンストンプラズマ物理研究所(米国)

＊3　ローレンス リバモア国立研究所(米国)

＊4　NECソリューションズ(米国)

＊5　海洋研究開発機構　地球シミュレータセンター

このプロジェクトでは4つの先端的な科学研究分野、大気大循環(FVCAMコード)、磁気核融合(GTCコード)、核融合、プラズマ物理(LBMHD3Dコード)、材用科学(PARATECコード)を対象としてベクトルプロセッサベースの地球シミュレータとスーパースカラプロセッサベースのIBM Power3の性能評価を実施した結果、以下のような重要な成果を得た。

(1) 有限体積法に基づくCAMコードのダイナミカルコアが、高解像度の下ではじめてベクトル化された;
(2) 新たに導入されたデータ分割手法により、GTCコードが粒子コードとして初めて1テラフロップスの壁を破った;
(3) 格子ボルツマン手法を応用した新しい磁気流体力学コード(LBMHD3D)によって、プラズマ乱流の発生と成長過程が調べられた。なお、この計算では地球シミュレータの4800プロセッサを用いて26TFlops以上の性能が得られた;
(4) PARATECコードで、過去最大規模の第一原理計算が実行された。

総合的な結果として地球シミュレータは評価した全てのアプリケーションで今までにない統合性能を達成し、最新の並列ベクトルプロセッサシステムの素晴らしい潜在能力を示した。

キーワード:性能評価，ベクトル化，科学技術計算