平成 26 年度先端的計算科学研究プロジェクト

運動論的第一原理宇宙プラズマシミュレーションコードの性能評価

梅田隆行（名古屋大学太陽地球環境研究所）
深沢圭一郎（京都大学学術情報メディアセンター）

1. 研究の目的と意義

太陽から地球に至るジオスペース環境の変動を理解することは、人類の活動が宇宙へと拡大しつつある今日、極めて重要な課題である。人類の活動に影響を与えるジオスペースの変動現象としては、突発的な磁気嵐やオーロラの爆発現象、放射線高エネルギー粒子生成、高エネルギー粒子線による人工衛星の誤作動などが挙げられる。これらの現象は、電磁気圏プラズマのグローバルな対流循環、境界層で生起するメソスケールでの突発的な不安定性（平衡状態の破れ）及び、電子・イオンが粒子として振舞うミクロスケール現象（粒子加速・加熱）が複雑に結びついており、マルチスケール結合過程であり、宇宙天気と呼ぶ。本研究の大きな目的は、ジオスペースで生起する非線形プラズマ現象を解明し、宇宙環境変動の因果関係を理解すると共に、数値宇宙天気予報に適用することである。

多様な時空間スケールの非線形プラズマ現象を解明するために、グローバル現象を扱う磁気流体力学（MHD）／多流体モデル、ミクロ現象を扱う運動論（粒子／ブラソフ）モデル及び、両者の中間（メソ）スケール現象を扱う流体と運動論のハイブリッドモデルなどの様々なコードが発達してきた。本研究プロジェクトではその中でも特に、Particle-In-Cell 型の粒子モデルと流体型のブラソフモデルの２つの運動論モデルについて着目しており、最新のスーパーコンピュータの能力を最大限に活用できるように、コードの性能評価及びチューニングを行うことを目的とする。その意義は、メソ・マクロスケールから粒子運動論的なミクロスケールまでをシームレスに扱う大規模シミュレーションにより、無衝突宇宙プラズマのマルチスケール結合過程の解明を目指すことにある。

本研究課題は、平成 24 年度より先端的計算科学研究プロジェクトのベンチマーク課題として採択されており[1,2]、得られた成果の概要については公開されている報告書を参照して頂けると幸いである。また、粒子モデルの性能評価については、先駆的科学計算に関するフォーラム 2014 ～数値シミュレーションと並列化技術～において紹介したためにここでは割愛し、そちらの講演資料を参照している頂けると幸いである。平成 26 年度は、新たに全ノードベンチマーク用のシステムに加わった HA8000 において、これまでに得られたチューニングの成果の有効性を確認すること及び、HA8000 と CX400 の結合環境（通称 Quartetto）において全ノード測定に挑戦することが目的である。

2. ブラソフコードの概要

ブラソフコードは、無衝突宇宙プラズマの運動論シミュレーション手法の１つであり、位置－速度位相空間に定義されたプラズマ粒子の分布関数の時間発展を、以下のブラソフ（無衝突ボルツマン）方程式により直接解き進めている。

$$\frac{\partial f_s}{\partial t} + \vec{v}\frac{\partial f_s}{\partial \vec{r}} + \frac{q_s}{m_s}(\vec{E} + \vec{v} \times \vec{B})\frac{\partial f_s}{\partial \vec{v}} = 0 \qquad (1)$$

ここで $\vec{E}$、$\vec{B}$、$\vec{r}$ 及び $\vec{v}$ はそれぞれ電場、磁場、位置、速度を表す。また、$f_s(\vec{r},\vec{v},t)$ は位置－速度位相空間におけるプラズマ粒子の分布関数であり、s はイオンや電子など粒子種を示す。q_s と m_s はそれぞれ電荷と質量を表す。

プラズマ粒子の分布関数は、式(1)で示される通り電磁場によって変形する。電磁場の時空間発展は以下のマックスウェル方程式によって記述される。

$$\nabla \times \vec{B} = \mu_0 \vec{J} + \frac{1}{c^2}\frac{\partial \vec{E}}{\partial t} \tag{2.1}$$

$$\nabla \times \vec{E} = -\frac{\partial \vec{B}}{\partial t} \tag{2.2}$$

$$\nabla \cdot \vec{E} = \frac{\rho}{\varepsilon_0} \tag{2.3}$$

$$\nabla \cdot \vec{B} = 0 \tag{2.4}$$

ここで $\vec{J}$ は電流密度、ρ は電荷密度、μ_0 は真空中の透磁率、ε_0 は真空中の誘電率、c は光速を示す。ブラソフ方程式(1)を速度空間で積分すると、以下の電荷保存則が得られる。

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \vec{J} = 0 \tag{3}$$

電磁場を変化させるための電流密度 $\vec{J}$ はプラズマの運動によって生じるが、これはブラソフ方程式(1)の第二項にある実空間の流束 $\vec{v}f_s$ を速度空間で積分することによって求まり、電流密度 $\vec{J}$ が電荷保存則(3)を満足する限り、ポアソン方程式(2.3)は自動的に満たされる。以上の方程式は、ブラソフコードにおいて解いているプラズマ粒子の運動論方程式であり、無衝突プラズマの第一原理と呼ぶ。

ブラソフ方程式は最大で実空間 3 次元及び速度空間 3 次元の「超多次元」を扱う方程式であるため、そのままの形で多次元数値積分を行うのは非常に困難であり、またコンピュータで解くには膨大なリソースを必要とする。本研究グループは長年にわたりブラソフシミュレーション手法の開発を行ってきており[4—7]、磁気リコネクション（テアリング不安定性）やケルビンーヘルムホルツ不安定性などのメソスケール現象のみならず[8—10]、イオンジャイロ半径スケールの半径を持つ非磁化小天体と太陽風との相互作用などのグローバルシミュレーションにも世界で唯一成功している[11—13]。ここでは手法の詳しい解説については省略する。

ブラソフシミュレーションでは非常に多くのメモリを必要とするため、並列計算が必須となる。ブラソフコードで使用する物理量は全て格子点上で与えられており、並列化においては領域分割法が有効である。図 1 は実空間 2 次元及び速度空間 3 次元を使用する 5 次元ブラソフコードにおける並列化の概念を示す。我々の目は 4 次元以上の空間を認識できないが、2 次元実空間の各格子上に 3 次元速度空間（速度分布関数）が定義されていると考えると分かりやすい。本研究では図 1 のように実空間（x－y 平面）においてのみ領域分割を行い、速度空間の領域分割は行わない[6]。これは、電荷密度や電流密度などのモーメント量を計算する際に必要な速度空間の積分において、各実空間でのリダクション処理を行わないようにするためである。

5 次元ブラソフコードでは、OpenMP による多重ループのスレッド並列も併用している。スレッド並列はそのオーバーヘッドの大きさから、より外側のループで行うのが効率的である。また、本研究グループにおける京コンピュータ 6000 ノードの実利用経験より、IO 処理や分散ファイルのデータ解析などの観点からプロセス数をできるだけ減らしたほうが利点は大きい。さらに、ブラソフモデルは 4 次元以上の超次元を扱い、メモリ使用量が非常に多いため、速度空間の格子点を 30^3－60^3 に固定してコアあたりのメモリ使用量 1-4GB に設定しつつ、使用ノード数を増やして計算領域（実空間の格子数）を拡張していくのが実際の超並列計算機の利用方法である。しかし、近年の計算機においては、ノード内の共有メモリの容量は増えずにコア数のみが増加していく傾向にあるため、単一のループのみをスレッド化する単純な方法には限界がある。昨年度の成果より、OMP DO ディレクティブの COLLAPSE オプションにより多重ループをスレッド並列化することにより、x86 系 CPU においても、FX10 や京コンピュータなどの近年の計算においても、ハイブリッド並列のほうが効率的になることが分かっている[2]。

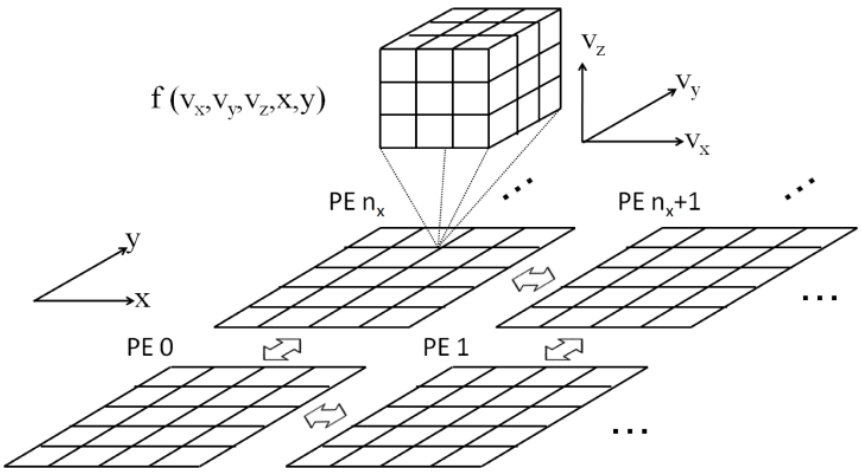

図 1：5 次元ブラソフコードにおける空間領域分割による MPI プロセス並列化[6]。

3. 性能測定

Hitachi HA8000 システムは、ノードあたり Xeon E5-2697 v2 (IvyBridge 2.70GHz 12 cores) を 2 つ搭載し、ノードあたりのメモリは 256GB である。またノード数は 965 であり、全ノードがフルバイセクションバンド幅で接続されている。また比較のために性能データを示す Fujitsu CX400 システムは、ノードあたり Xeon E5-2680 (SandyBridge 2.70GHz 8 cores)を 2 つ搭載し、ノードあたりのメモリは 128GB である。またノード数は 1476 であり、256 ノードがフルバイセクションバンド幅で接続されたグループが 5 つ、196 ノードがフルバイセクションバンド幅で接続されたグループが 1 つあり、それぞれのグループが FDR Infiniband で接続されている。

本プロジェクトの性能測定では、“コアあたり”の格子数を実空間では 40×20、速度空間では 30×30×30 に設定している。これはコアあたりのメモリ使用量を約 1GB に固定してノード数を増やしていく弱いスケーリングに該当する。しかし全ノード数の約数より、x 方向及び y 方向の分割数は制限される。表 1 及び 2 に、HA8000 及び CX400 の全ノード測定において使用したノードの分割数及びノードあたりの実空間の格子点数をそれぞれ示す。HA8000 は 965 = 193 x 5 というノード数より、2 次元分割の組み合わせが大きく制限され、その結果としてシステム長が x 方向に極端に長くなっている。CX400 においても、1476 = 2^2 x 3^2 x 41 というノード数により、2 次元分割の組み合わせは表 2 のようになっており、HA8000 の場合に比べるとシステムのアスペクト比は 1 に近くなっている。

表 1：HA8000 全ノード測定におけるノードあたりのスレッド数と MPI 並列数及び実空間格子点数。

ノードあたりの スレッド数	分割数 (MPI プロセス数)	プロセスあたりの実空間 格子点数(Nx x Ny)	トータルの実空間 格子点数(Nx x Ny)
1	193 x 120	40 x 20	7720 x 2400
4	193 x 30	80 x 40	15540 x 1200
8	193 x 15	80 x 80	15540 x 1200
24	193 x 5	160 x 120	30880 x 600

表 2：CX400 全ノード測定におけるノードあたりのスレッド数と MPI 並列数及び実空間格子点数。

ノードあたりの スレッド数	分割数 (MPI プロセス数)	プロセスあたりの実空間 格子点数(Nx x Ny)	トータルの実空間 格子点数(Nx x Ny)
1	164 x 144	40 x 20	6560 x 2880
2	144 x 82	40 x 40	5760 x 3280
4	82 x 72	80 x 40	6560 x 2880
8	72 x 41	80 x 80	5760 x 3280
16	41 x 36	160 x 80	6560 x 2880

図 2 に、HA8000 及び CX400 における全ノード測定結果を示す。横軸はノードあたりのスレッド数であり、縦軸は時間ステップあたりの経過時間を示す。また経過時間は、メインループの全ての計算及び、速度空間における分布関数データの更新を行う Velocity、実空間における分布関数データの更新を行う Position 及び、電磁場データの更新を行う Maxwell の 4 つを示す。また、青い丸で示したデータは、COLLAPSE オプションを付けない”As-is”コードでの測定結果であり、赤い四角で示したデータは、COLLAPSE オプションにより多重ループのスレッド化を行ったコードでの測定結果である。

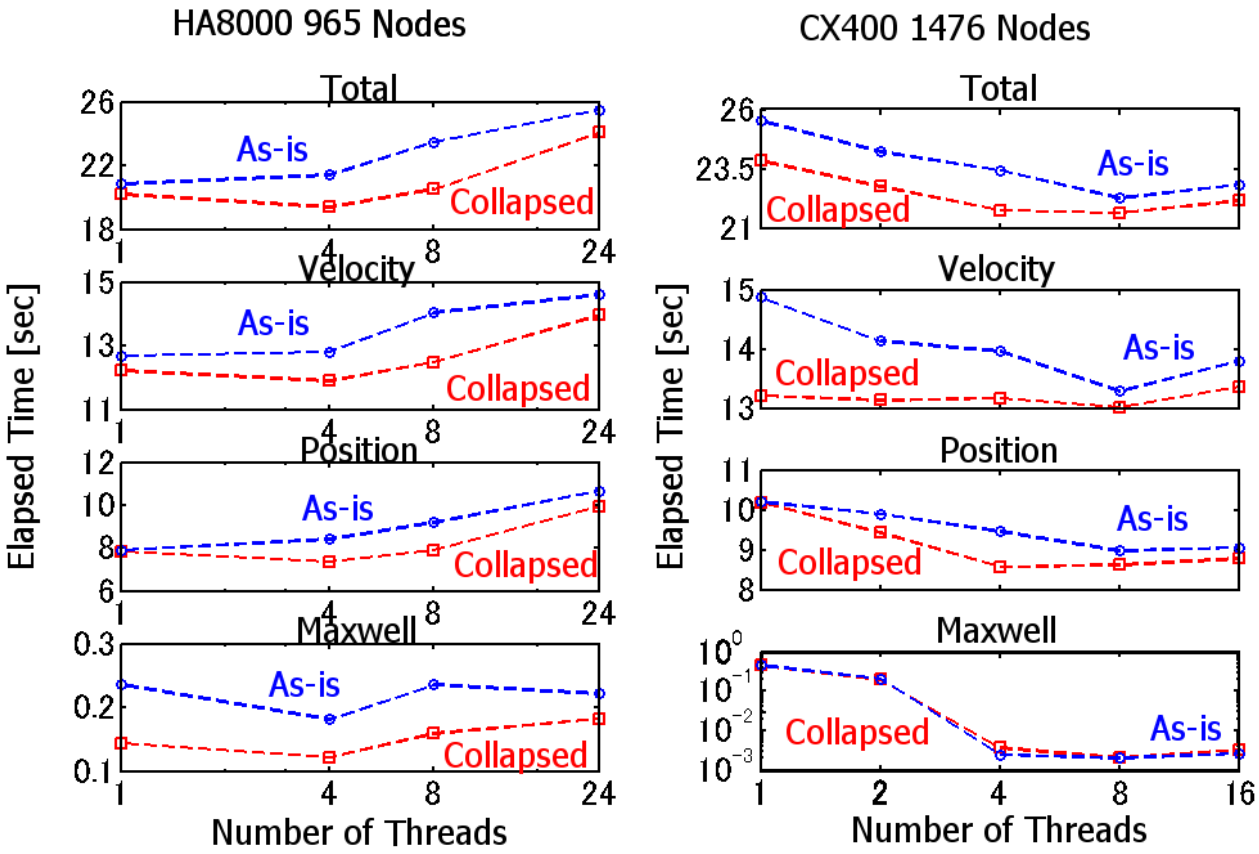

図 2：HA8000（左）及び CX400（右）の全ノード測定における、ノードあたりのスレッド数と時間ステップあたりの経過時間。

HA8000 及び CX400 の両方のシステムにおいて、COLLAPSE オプション有の場合のほうが As-is コードに比べて経過時間が短い、即ち計算速度が若干向上していることが分かる。どちらのコードにおいても、ノードあたりのスレッド数による経過時間の変化は似たようなものになった。HA8000 においてはノードあたり 4 スレッドの場合が最速であり、CX400 においてはノードあたり 4 スレッド及び 8 スレッドが最速であった。これらの結果は、ネットワークトポロジー及び MPI 分割の組み合わせによって性能が大きく依存することを示唆している。

最後に、1 コアあたり 1GB にメモリ使用量を固定したときの、5 次元ブラソフコードの弱いスケーリング性能を図 3 及び図 4 に示す。図 3 はコア数に対する実効速度（GFlops）を表し、図 4 はコア数に対する実効並列効率（スケーラビリティ：N コアの実効効率を 1 コアの実効効率で割った量）を表す。本測定では、COLLAPSE コードを用いてスレッド数 4 のハイブリッド並列で計算した結果を用いている。ネットワークトポロジーの関係で CX400 の 1476 ノードのほうがスケーラビリティがより低下するかと思われたが、HA8000 では 965 ノードを用いた場合のスケーラビリティが最も悪く、約 80%であった。HA8000 ではスケーラビリティが 4 コアから 24 コアにかけて急激に落ちているため、ノード内のメモリバンド幅が原因ではないかと考えられる。

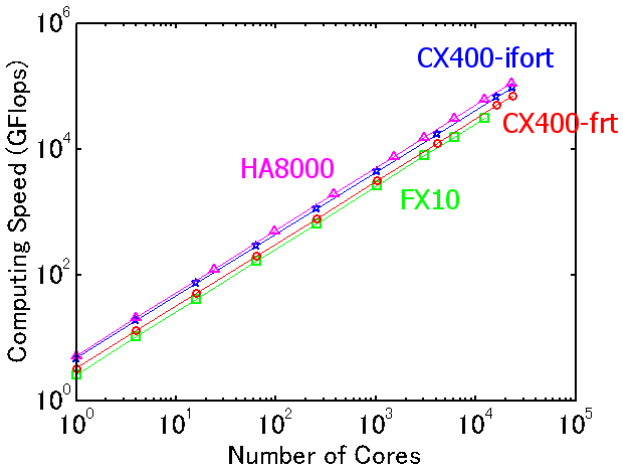

図 3：1GB/core 使用時の 5 次元ブラソフコードの弱いスケーリング性能(1)。コア数に対する実効速度を示している。

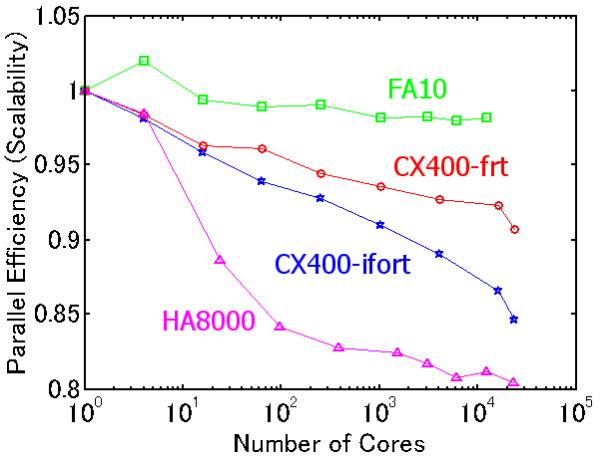

図 4：1GB/core 使用時の 5 次元ブラソフコードの弱いスケーリング性能(2)。コア数に対する実効並列効率を示している。

4. Quartetto 性能測定

結論から言うと、HA8000 と CX400 の結合環境では全ノード測定はうまくいかなかった。HA8000 のみ 512 ノード及び CX400 のみ 512 ノードの環境については、ほぼ正常に動作した。また HA8000 と CX400 をそれぞれ 128 ノード（計 256 ノード）の場合についても動作を確認した。しかし、HA8000 と CX400 をそれぞれ 256 ノード（計 512 ノード）の場合についてはうまく動作しなかった。標準出力が何も出ないまま数時間後に broken pipe で終了していたため、MPI プロセスが全てのノードにおいてうまく起動していなかったと考えている。

まとめに代えて、結合環境での測定において障害となったものについて、以下にまとめる。

①通常運用のジョブ管理機能と異なる。

結合環境では、ログインノードからジョブを投入せず、バックエンドノードの 1 つに直接ログイン

し、そこから mpirun を起動した。そのため、mpirun を起動したノードで ssh プロセスが大量に起動し、そのノードが固まる現状が頻発した。またこれにより、mpirun を起動するノードを測定に含められなかった。

②バックグラウンド実行ができない。

端末に張り付いている必要があり、通勤など長時間端末を離れる必要がある際にジョブを実行できない。

③machinefile の作成を手動で行う。

どのマシンを何ノード用いるかに依って、machinefile を複数用意する必要がある。

④残留プロセスの確認や消去を手動で行う。

残留プロセスの確認や消去を行うコマンドがログインノードに用意されていた。ジョブが終了する度にコマンドを頻繁に実行しており、その過程でログインノードが 2 回落ちた。

最後に、HA8000 と CX400 の結合環境を触れる貴重な機会を与えて下さった情報基盤研究開発センタースタッフ及び両システムの SE に感謝申し上げる。本報告書のまとめが次回の結合環境に役立てば幸いである。

参 考 文 献

[1] 梅田 隆行, 深沢 圭一郎: 第一原理無衝突プラズマシミュレーション用 5 次元ブラソフコードの性能評価, 平成 24 年度先端的計算科学研究プロジェクト報告書 (2013).
http://www2.cc.kyushu-u.ac.jp/scp/users/forum/forum20130426/04_umeda.pdf

[2] 梅田 隆行, 深沢 圭一郎: 流体型宇宙プラズマシミュレーションコードの性能チューニング, 平成 25 年度先端的計算科学研究プロジェクト報告書 (2014).

[3] 梅田 隆行, プラズマシミュレーションにおける粒子法, 先駆的科学計算に関するフォーラム 2014 ～数値シミュレーションと並列化技術～講演資料 (2014).
http://www.cc.kyushu-u.ac.jp/scp/users/forum/forum20140805-06/umeda.pdf

[4] Umeda, T.: A conservative and non-oscillatory scheme for Vlasov code simulations, *Earth Planets Space*, Vol.60, No.7, 773—779 (2008).

[5] Umeda, T., Togano, K., Ogino, T.: Two-dimensional full-electromagnetic Vlasov code with conservative scheme and its application to magnetic reconnection, *Comput. Phys. Commun.*, Vol.180, No.3, 365—374 (2009).

[6] Umeda, T., Fukazawa, K., Nariyuki, Y., Ogino, T.: A scalable full electromagnetic Vlasov solver for cross-scale coupling in space plasma, *IEEE Trans. Plasma Sci.*, Vol.40, No.5, 1421—1428 (2012).

[7] Umeda, T., Nariyuki, Y., Kariya, D.: A non-oscillatory and conservative semi-Lagrangian scheme with fourth-degree polynomial interpolation for solving the Vlasov equation, *Comput. Phys. Commun.*, Vol.183, No.5, 1094—1100 (2012).

[8] Umeda, T., Togano, K., Ogino, T.: Structures of diffusion regions in collisionless magnetic reconnection, *Phys. Plasmas*, Vol.17, No.5, 052103(6pp.) (2010).

[9] Umeda, T., Miwa, J., Matsumoto, Y., Nakamura, T. K. M., Togano, K., Fukazawa, K., Shinohara, I.: Full electromagnetic Vlasov code simulation of the Kelvin-Helmholtz instability, *Phys. Plasmas*, Vol.17, No.5, 052311(10pp.) (2010).

[10] Umeda, T., Ueno, S., Nakamura, T. K. M.: Ion kinetic effects to nonlinear processes of the Kelvin-Helmholtz instability, *Plasma Phys.Contr.Fusion*, Vol.56, No.7, 075006(11pp.) (2014).

[11] Umeda, T., Kimura, T., Togano, K., Fukazawa, K., Matsumoto, Y., Miyoshi, T., Terada, N., Nakamura, T. K. M., Ogino, T.: Vlasov simulation of the interaction between the solar wind and a dielectric body, *Phys. Plasmas*, Vol.18, No.1, 012908(7pp.) (2011).

[12] Umeda, T.: Effect of ion cyclotron motion on the structure of wakes: A Vlasov simulation, *Earth Planets Space*, Vol.64, No.2, 231—236 (2012).

[13] Umeda. T., Ito, Y.: Entry of solar-wind ions into the wake of a small body with a magnetic anomaly: A global Vlasov simulation, *Planet. Space Sci.*, Vol.93-94, 5—40, (2014).

[14] Umeda, T., Fukazawa, K.: A high-resolution global Vlasov simulation of a small dielectric body with a weak intrinsic magnetic field on the K computer, *Earth Planets Space*, in press, (2015).