Vol.2018-HPC-166 No.4
2018/9/27

スーパーコンピュータシステム ITO における MHD シミュレーションコードの計算性能・消費電力評価

深沢圭一郎[†1]　南里豪志[†2]　本田宏明[†3]

概要：スーパーコンピュータシステム ITO は 2017 年度に九州大学に導入された大規模計算機システムであり，総理論性能が約 10PFlops という性能を有している．本研究では，Skylake 世代の Xeon を搭載した ITO のサブシステム A の 2,000 ノード（理論性能 6.91 PFLOPS）をすべて利用し，実アプリケーションである MHD シミュレーションの性能評価を行った．単純な SoA や AoS といった配列の形状だけで無く，それらを変形させた配列形状での性能を比較すると，計算性能の違いが確認できた．実効性能は 2,000 ノードを利用し，約 470TFlops となった．また，性能評価時に消費電力の計測を行なった結果，計算ノード間での CPU 周波数のばらつき，測定毎による周波数のばらつきからにより，消費電力のばらつきが大きいことが分かった．本研究ではこれらの詳細な性能，計測データを示し，その結果を議論する．

キーワード：性能評価，MHD シミュレーション，消費電力

Performance and Power Consumption Evaluation of MHD Simulation Code on Supercomputer System ITO

KEIICHIRO FUKAZAWA[†1]　TAKESHI NANRI[†2]　HIROAKI HONDA[†3]

Abstract: The supercomputer system ITO has introduced to Kyushu University in 2017 and its peak performance reaches around 10 PFlops. In this study the performance of Magnetohydrodynamic (MHD) simulation code is evaluated using the all 2,000 nodes of subsystem A of ITO which has Skylake Xeon and peak performance 6.91 PFlops. In the evaluation several array configurations are used, and the results show the different calculation performance among those configurations. From those results the best performance with 2,000 nodes is about 470TFlops. In this evaluation the large variations of power consumption and CPU frequency are appeared on not only the different calculation node but the same calculation node. This study shows the detail of performance evaluation and power consumption.

Keywords: Performance evaluation, MHD simulation, Power consumption

1. はじめに

初期のスーパーコンピュータは CRAY-1 から始まるベクトル型計算機が主であり，スカラ型 CPU の性能向上とノード間接続・並列計算技術の向上伴い，2000 年代頃から，徐々にスカラ型 CPU を搭載した大並列スーパーコンピュータが増えてきた．近年では，ほぼすべてのスカラ型計算機が x86 型の CPU を利用しているが[1]，プロセス微細化技術やリーク電流の問題もあり，周波数の向上が難しくなった結果，コア数の増加や同時演算数の増加により，CPU 性能の向上を達成している．例えば，Xeon Phi KNL では，60 を超える CPU コアを持ち，倍精度での同時演算数は 32 であり，一般のパソコンなどに利用されている CPU とは大きく異なる．このような中，九州大学のスーパーコンピュータシステム ITO では x86 型 CPU である Skylake 世代の Xeon を採用し，同時演算数が 32，ノード当たりのコア数が 36 となっている．このような計算機システムの理論性能はカタログスペックにより分かるが，実際にアプリケーションを動かした場合どのような性能になるのかは，予測が難しい．

そこで，本研究ではこれまでに様々なスーパーコンピュータで性能評価を行ってきた電磁流体（MHD）シミュレーションコード[2]を用いて，ITO の性能評価を行う．MHD シミュレーションは通常の流体シミュレーションに電磁場の効果を考慮したシミュレーションになっており，本性能評価の結果は流体系のアプリケーションに広く応用でき，また，これまでに評価してきた計算機システムの性能と比較することで，現実的な計算性能を見積もることも可能と考えられる．

また，エクサフロップス級計算機システムを作る上で消費電力が問題になっているように，近年は計算機システムの消費電力に注目が集まっている．一般に，アプリケーション毎に消費電力特性が異なることから，自分の利用するアプリケーションがどのような消費電力特性を持つかを知っておくことが今後の計算機システムを利用する上で重要

†1 京都大学・学術情報メディアセンター
Academic Center for Computing and Media Studies, Kyoto University
†2 九州大学 情報基盤研究開発センター
Research Institute for Information Technology Kyushu University
†3 株式会社ハイドロ総合技術研究所
Hydro Technology Institute Co., Ltd.

となると考えられる．そこで，本性能評価では，計算性能を測るだけでは無く，消費電力についても測定を行う．

本研究報告の構成は以下の通りである．第 2 章では，スーパーコンピュータシステム ITO について説明し，第 3 章では MHD シミュレーションコードについて説明をする．第 4 章で性能評価の結果を述べ，その結果と他システムとの比較を第 5 章で行い，最後に研究のまとめをする．

2. スーパーコンピュータシステム ITO

スーパーコンピュータシステム ITO は，多数の CPU 計算ノードが接続されたサブシステム A（2,000 ノード）と 1 ノード当たり 4GPU が搭載されたサブシステム B（128 ノード）により構成される．九州大学情報基盤研究開発センターによると，スーパーコンピュータシステム ITO は Intel の最新 CPU（Xeon Gold, Skylake-SP）と，NVIDIA 社の GPU（Tesla P100, Pascal）を搭載し，総理論演算性能約 10PFlops を有する国内トップクラスの能力をもつシステムである[3]．本性能評価ではこの内，サブシステム A を利用し，性能評価を行う．サブシステム A の構成を表 1 に示す．

Skylake 世代の Xeon では Xeon Phi KNL と同様に AVX-512 に対応し，Xeon Gold より上のクラスであれば FMA ユニットがコア当たり二つあり，同時演算数は 32 となっている．また，メモリチャンネルが Broadwell 世代の 4 から 6 と増えており，メモリバンド幅が増加している．

3. MHD シミュレーションコード

宇宙空間は真空と思われているが，その 99 %はプラズマで満たされている．プラズマとは電離した気体のことであり，帯電している電子とイオンが分かれて存在する状態である．宇宙空間，特に我々の暮らす太陽系においては太陽から太陽風と呼ばれるプラズマの風が常時吹き出しており，太陽系全体にそのプラズマが充満している．このようなプラズマの振る舞いを記述する方程式として Vlasov-Maxwell 方程式がある．これは，無衝突 Boltzmann 方程式と Maxwell 方程式から成る．Vlasov（無衝突 Boltzmann）方程式は以下の形をとる．

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

ここで$\vec{E}$，$\vec{B}$，$\vec{r}$ と $\vec{v}$ はそれぞれ電場，磁場，距離，速度を表す．また，$f_s(\vec{r},\vec{v}_s,t)$ は位置－速度位相空間における分布関数であり，sはイオンや電子など種類を示す．q_s は電荷をm_s は質量を表す．

しかしながら，Vlasov 方程式は多くの成分からなる非線形方程式であり，計算機システムを用いても解くことが非常に難しい．そこで，Vlasov 方程式のモーメントをとることで求められる電磁流体力学（MHD）方程式が，グローバルなプラズマ構造を調べるときには使用されている．MHD 方程式は以下のようになる．

表 1　ITO サブシステム A の諸元
Table 1　Subsystem A of ITO

機種名	Fujitsu PRIMERGY CX2550/CX2560 M4	
計算ノード	CPU	Intel Xeon Gold 6154 (Skylake-SP) × 2 /node
	コア数	18 cores /CPU
	周波数	3.0 GHz (Turbo 3.7 GHz)
	理論性能	3,5 TFlops /node (**倍精度**)
	メモリ	DDR4 192 GB /node
	Bandwidth	255. 9 GB/s /node
	B/F	0.074
総ノード数	2,000 nodes	
総理論性能	6.91 PFlops	
ノード間接続	InfiniBand EDR 4x (100Gbps)	

$$\begin{aligned}
\frac{\partial \rho}{\partial t}&=-\nabla\cdot(\mathbf{v}\rho)\\
\frac{\partial \mathbf{v}}{\partial t}&=-(\mathbf{v}\cdot\nabla)\mathbf{v}-\frac{1}{\rho}\nabla p+\frac{1}{\rho}\mathbf{J}\times\mathbf{B}\\
\frac{\partial p}{\partial t}&=-(\mathbf{v}\cdot\nabla)p-\gamma p\nabla\cdot\mathbf{v}\\
\frac{\partial \mathbf{B}}{\partial t}&=\nabla\times(\mathbf{v}\times\mathbf{B})
\end{aligned} \qquad (2)$$

上から，連続の式，運動方程式，圧力変化の式（エネルギーの式），最後が磁場の誘導方程式となる．簡単に言えば，電磁場を考慮した流体力学方程式と呼べる．詳しい導出方法は参考文献を参照されたい[4].

MHD 方程式を解く数値計算法としては，Modified Leap Frog（MLF）法[2，5]という計算法を使用する．これは最初の 1 回を two step Lax-Wendroff 法で解き，続く （l － 1）回を Leap Frog 法で解き，その一連の手続きを繰り返す．l の値は数値的に安定の範囲で大きい方が望ましいので，本手法で採用する 2 次精度の中心空間差分では，数値精度の線形計算と予備的シミュレーションから l = 8 に選んでいる．

並列化にはプロセス並列に MPI を使用する．プロセス並列化手法としては 3 次元空間を分割する領域分割法を用いる．領域分割には，1 次元，2 次元，3 次元分割が考えられ，本性能評価ではこれらすべての評価を行う．領域分割の次元数により，計算ループのベクトル長が変わるため，それぞれの性能評価を行う．

一般的にスカラ CPU で性能を出すためにはキャッシュの有効活用が重要である．基本的な動作としてはメモリアクセス時に，その周辺数 KB のデータをキャッシュに格納

する．キャッシュの量や，一度にキャッシュに格納するデータ量は CPU アーキテクチャ毎に変わるため，最高のパフォーマンスを出すにはそれぞれの調整が必要である．MHD シミュレーションにおいては，物理変数がプラズマ密度，速度 3 成分，圧力，磁場 3 成分の計 8 変数となる．そのため，配列を $f(x, y, z, m)$ と定義し，$m = 8$ としている．数値計算時に同じ場所の物理変数を何度も使うことになるため，一般に Fortran では，$f(m, x, y, z)$ と定義した方がキャッシュヒット率は上がることがわかっている[6]．しかしながら，近年の Xeon CPU ではベクトル化が性能向上にとって重要な機構であるため，更に配列を $f(x, m, y, z)$と $f(x, y, m, z)$と定義した場合の性能評価も行う．

4. 性能評価

4.1 計算性能

スーパーコンピュータシステム ITO では，富士通コンパイラと Intel コンパイラ，PGI コンパイラが利用できる．PGI コンパイラは九州大学構成員のみ利用可能なため，ここでは富士通コンパイラと Intel コンパイラを利用し，性能評価を行った．計算サイズはプロセス当たり，512MB（200^3 グリッド）を利用した．計測は 5 回行い，その平均値を取った．

富士通コンパイラはバージョン 1.2.0（Nov 27 2017）を利用し，コンパイラオプションは下記を利用した．

`-Kfast,fsimple,prefetch_indirect,prefetch_sequential,nomfunc,noparallel,simd=2`

オプションの詳細は Fujitsu Technical Computing Suite のマニュアルを参照されたい．

図 1 にサブシステム A を 2,000 ノード利用し，MHD シミュレーションコードの計算性能を測定した結果を示す．前述のように，本研究では 3 種類の領域分割と 3 次元領域分割時に配列の要素を入れ替えた場合（3 種類）の計 6 種類の場合において性能を測定した．これまでの Xeon やベクトル型 CPU と同様にベクトル長が長くなる 1 次元，2 次元領域分割の性能が高くなっている．最大で 2,000 ノード利用時に 1 次元領域分割で 421 TFlops の性能となっている．また，3 次元領域分割における配列順序を変えた結果は，いわゆる AoS（Array of Structure）形式（図 1 内の 3D mxyz, $f(m, x, y, z)$を示す）の場合にのみ明らかに性能が下がる．他の場合ではそれほど顕著な差は出ていないが，xyzm（$f(x, y, z, m)$）か xmyz（$f(x, m, y, z)$）の性能が良く，xymz（$f(x, y, m, z)$）はわずかに性能が下がる結果となっている．このことから，ベクトル化とキャッシュ利用の両方を狙うには，xmyz が一つの候補と考えられる．

次に Intel コンパイラを利用し，性能測定を行う．Intel

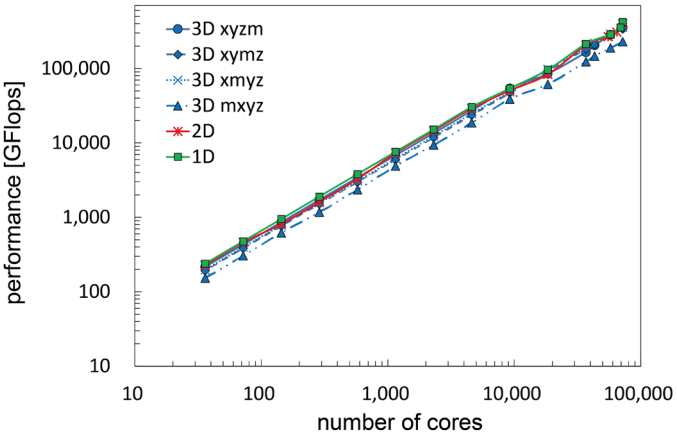

図 1　富士通コンパイラによる MHD コードの性能
Figure 1　Performance of MHD code with Fujitsu compiler

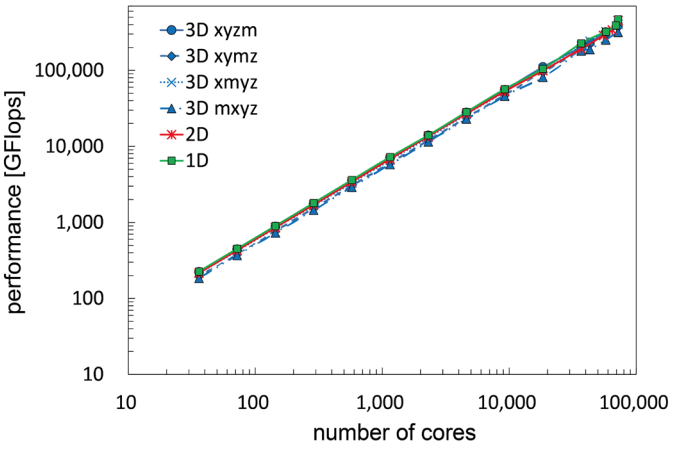

図 2　Intel コンパイラ 2017 による MHD コードの性能
Figure 2　Performance of MHD code with Intel compiler 2017

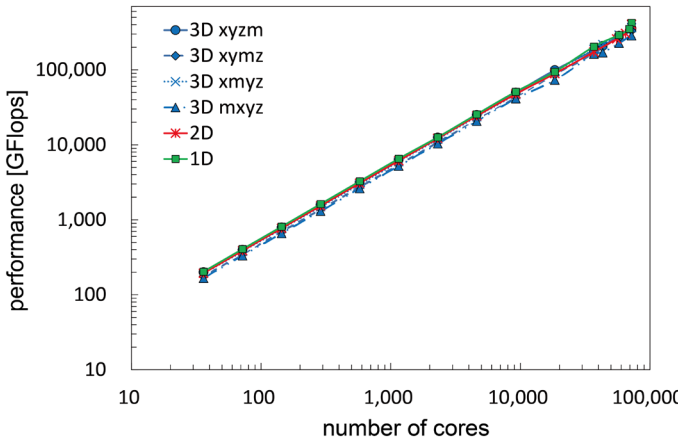

図 3　Intel コンパイラ 2018 による MHD コードの性能
Figure 3　Performance of MHD code with Intel compiler 2018

表 2　ITO における MHD シミュレーションコード実行時の消費電力特性
Table 2　Power consumption characters of MHD simulation code on ITO

	経過時間 [秒]	CPU 消費電力 [W]	DRAM 消費電力 [W]	CPU 周波数 [W]
平均	0.088	135.118	41.754	3.236
最大	0.154	148.252	51.606	3.700
最小	0.056	122.104	34.940	2.320

コンパイラはバージョン 2017 と 2018 が利用できたため，両バージョンを利用し，性能測定を行った．利用したオプションを下記に示す．積極的にベクトル化を行うオプションとなっている．

```
-ipo -O3 -fp-model fast=2 -xCOMMON-AVX512 -no-prec-div -no-prec-sqrt -unroll -static -align array64byte -vec-threshold0 -qopt-zmm-usage=high
```

図 2 と 3 にその計測結果を示す. 図 2 が 2017, 図 3 が 2018 を利用した結果となっている．2017 と 2018 の結果では，どの領域分割の性能が良いなどの傾向は変わらないが，ベクトル最適化のアグレッシブさに違いがあるため, 2017 の性能が約 10%程度高い結果となっていた. 最も高い性能は，1 次元領域分割時に 2,000 ノードを利用し，約 470TFlops となっている．これは富士通コンパイラより 12%近く高い性能となっている．一方で 1 ノードを利用した場合は，富士通コンパイラが 5%程度高い性能となっており, 256 ノード以上利用時に性能の逆転が見える．このことから Intel コンパイラ利用時の MPI の性能が富士通コンパイラ利用時に比べて良いと考えられる．また，Intel コンパイラ利用時の結果では，AoS タイプの配列構造である 3D mxyz の性能がそれほど悪くない．富士通コンパイラ利用時と比べ 38%程度の性能向上が見える．配列の順序による性能の違いは富士通コンパイラと同じ傾向だが，その差は少なくなっている．Intel コンパイラはより強い最適化を行っている結果と言える．

4.2 消費電力特性

Sandy Bridge 世代以降の Xeon には RAPL（Running Average Power Limit）と呼ばれる電力測定，電力制限機構が備わっており，この RAPL を用いて，ITO の電力測定を行った．また，RAPL 利用に際して，ディレクティブ形式でコードに電力測定や消費電力制限を加えることができる RIC を利用した[7]. 後述するが, 計測結果にブレが大きかったため, 64 回の計測を行った．計測には 8 ノードを利用した．

表 2 に MHD シミュレーションコードの消費電力特性をまとめた．ここでは，経過時間，CPU と DRAM の消費電力，CPU 周波数の平均値，最大最小値を示している．表 1 で示したように ITO の CPU である Xeon Gold 6154 はベース周波数が 3.0 GHz，Turbo boost 時に最大 3.7 GHz まで上昇する．しかし，すべてのコアを利用する際は Turbo boost 時には 3.3 GHz の周波数になり，また，AVX-512 利用時には 2.7 GHz に周波数が下がる．更に Skylake 世代の CPU から，動的に CPU の周波数や動作電圧を調整する機能（Intel Speed Shift Technology）もあるため，コードを実行する毎に，消費電力特性が変わることが予想される．

表 2 にある最大最小消費電力の差はこれらが原因となり, 現れていると考えられる. 計測結果では, 最大周波数が 3.7 GHz となり，Turbo boost 最大値と同じ周波数であり，平均周波数はベース周波数より高く, 全般的に Turbo boost 寄りで動作していることが分かる. また, 最低周波数は AVX-512 利用時よりも低いため，熱的問題が発生し，周波数が調整されていると考えられる．これらの変化に伴い，消費電力は増減するが, この CPU の TDP200W 以下で動作しており, 省電力傾向にあることが分かる．一方で，経過時間にも大きなブレが現れており，ある一定の性能を保証することが難しくなってきている．これは専用機に近い，SX-ACE のようなベクトル機，京コンピュータやその後継機には見られなかった変動であり，ITO 利用時には経過時間制限の注意が必要かもしれない.

ここまでは，統計化したデータを見てきたが，Skylake Xeon での消費電力特性の変動が同一ノードでどのように変わっているのか，複数ノードで同じような変動を取るのかなど詳しく確認するために, 各 8 ノードの 64 回計測時における経過時間と CPU 周波数変動を図 4 に示す. 左側の縦軸が周波数を示し，右側の縦軸が経過時間を示す．横軸は計測回数を表している．1 ノードには 2CPU が搭載されているため，図中では赤色でソケット 0，青色でソケット 1 の変動を表し，実線が周波数，破線が経過時間を示している. 計測回数が 40 回手前に大きな経過時間があるが, 周りと比べて遅すぎるため，何らかの不具合があったとし，ここでは扱わない.

まず，ノード内での変動に注目すると，計測回数毎に CPU 周波数が小刻みに変化していることが分かる，計測は連続的に行っているため，計測回数は時間変化とも見て取れる．高い周波数（3.7 GHz）を維持することはほとんど無いが，高い周波数を複数計測回維持している場合もあり，一様な変化があるとは言えない．同一ノードのソケット 0 と 1 において周波数変動が異なるものもあれば，似た変動を示しているもあり，単ソケット内で独立に電力自動調整

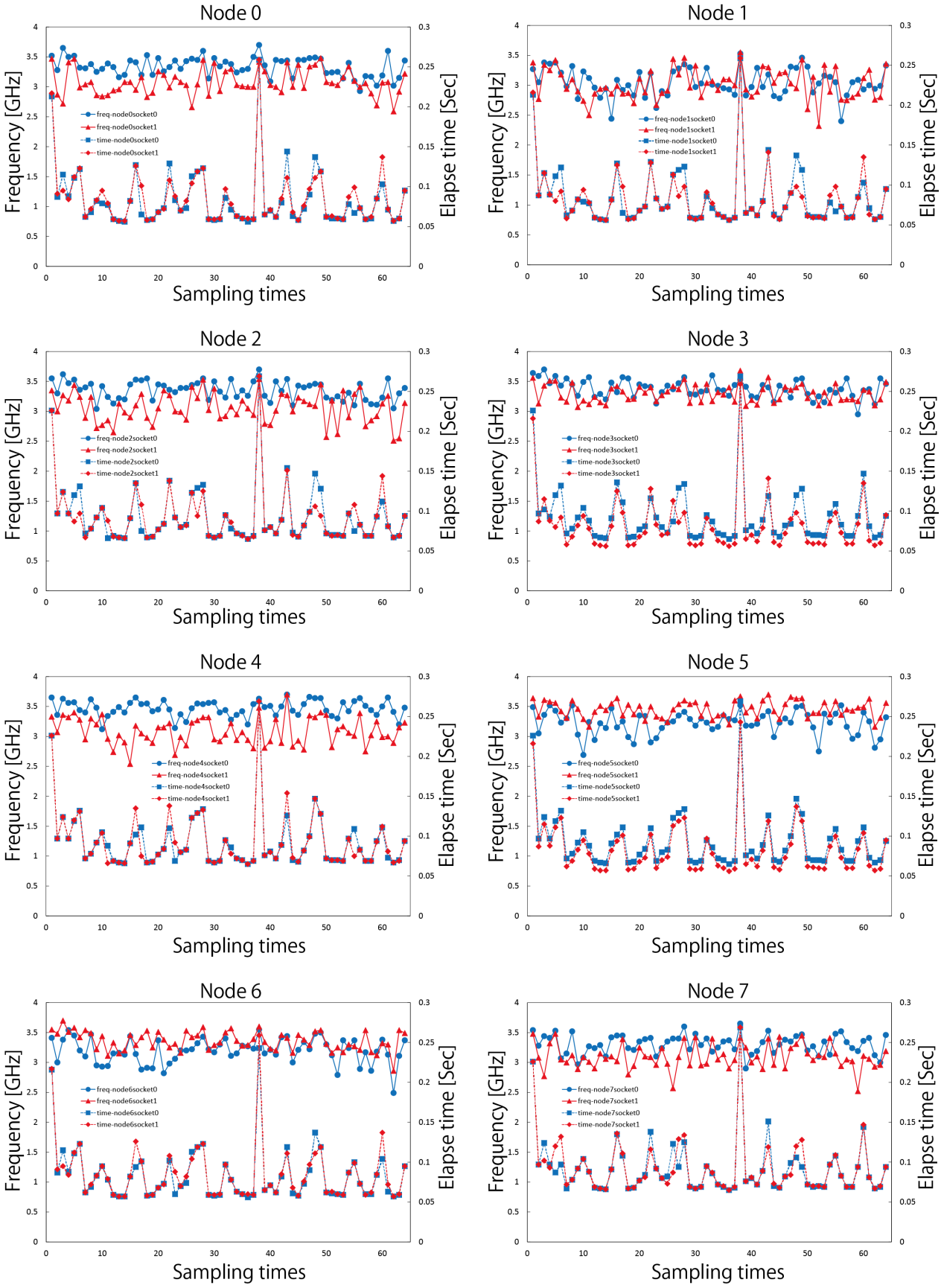

図 4　各ノードにおける CPU 周波数と経過時間の変動

Figure 4　CPU frequency and elapse time on each node

表 3　様々な計算機システムにおける性能の傾向[6，8]
Table 3　Performance trend of various computer systems [6，8]

	Core/CPU	Rpeak [TFlops]	Rmax [TFlops]	Rmax /CPU [GFlops]	Efficiency [%]	Suitable domain decomposition	CPU architecture
SX-ACE	1024/256	65.50	29.20	114.0	45	3D xyzm	Vector
K	262144/32768	4194.30	914.12	27.9	22	3D mxyz	SPARC64 VIIIfx
FX100	16384/512	576.72	91.49	178.7	17	3D xyzm	SPARC64 XIfx
CX400	23616/2952	510.11	104.23	35.3	20	3D xyzm	Xeon (SandyBridge)
HA8000	23160/1930	500.26	83.42	43.2	17	2D	Xeon (IvyBridge)
XC30	448/32	16.49	1.37	42.8	8	2D	Xeon (Haswell)
XC40	1088/16	48.86	4.32	273.3	9	3D xyzm	Xeon Phi KNL
Xeon Phi 5120	60/1	1.00	0.08	84.0	8	3D xyzm	Xeon Phi KNC
Tesla K20X	896/1	1.31	0.15	153.3	12	3D xyzm	Kepler
ITO-A	72000/4000	6912.00	470.10	117.5	7	1D	Xeon (Skylake)

が行われていると考えられる．一方，経過時間と周波数を比べてみると，逆相関ではなく，むしろ似た傾向の変動を示しているように見える．周波数が高い場合に，単純に早く計算が終わるということではないと分かる．これは B/F 値などこの変動以外に原因があると思われる．複数ノードを比較してみると，周波数の変動にはノード間の関連が無いように見えるが，経過時間はソケット間，ノード間でもおおよそ同じ値を取っている．MHD コードには同期部分が含まれるため，同期に伴い経過時間が一律になっていると考えられる．

5. 他計算機システムとの比較

スーパーコンピュータシステム ITO での MHD シミュレーションコードの性能と他の計算機システムでの MHD シミュレーションコードの性能を比較することで，計算機間の実性能理解に繋がると考えられる. 表 3 にこれまで MHD シミュレーションコードの性能を計測したいくつかの計算機システムの結果と ITO の測定結果を示す[5，6]．今回の ITO の性能評価では 3 次元領域分割において，4 種類の配列構造を利用したが，これまではいわゆる SoA（xyzm）と AoS（mxyz）を利用した測定しかしていない．また，CPU（GPU，コプロセッサー）自体の性能を比較しやすいように，CPU 当たりの性能（Rmax/CPU）を表に加えている．CPU 当たりの性能は Xeon 搭載機である CX400 では，35.3GFlops，HA8000 では 43.2GFlops，そして XC30 では 42.8GFlops となっており，Skylake 世代 Xeon 搭載の ITO は 117.5GFlops と約 3 倍の性能となっていることが分かる. 一方で，最新の Xeon Phi を搭載した XC40 は 273.3GFlops となっており, Xeon サーバが 1 ノード当たり 2CPU 搭載していることを考えても，XC40 の 1 ノード（Xeon Phi KNL を 1 つ搭載）が 16%程度高い性能を示している．Skylake 世代は現時点では最新の Xeon であり，コンパイラの性能向上，またコード最適化が進むことで，差は少なくなると想像される．x86 系以外のアーキテクチャと比べると，京コンピュータで 27.9 GFlops，FX100 が 178.7GFlops となっているが，両システムとも 1 ノードに 1CPU のため，ITO はノード当たりでは FX100 の 30%程度高い性能となっている.

このように Xeon としては，世代が新しくなることに実性能も向上しており，また，既設のシステムと比べてノード当たりでは高い性能を達成している．

6. まとめ

九州大学に新しく導入されたスーパーコンピュータシステム ITO に対して, 宇宙プラズマを解く MHD シミュレーションコードの性能測定を行った．富士通コンパイラと Intel コンパイラを利用した結果, 単純な性能の差だけでは無く，異なる構造を持つ配列に対してそれぞれのコンパイラで最適化に差があった．富士通コンパイラは，少ないノードでの性能が Intel コンパイラよりも高く，512 ノード以上では Intel コンパイラの性能が高くなり，スケーラビリティに明かに差があった．今回 3 次元領域分割において 4 種類の配列構造を用いたが，ベクトル化とキャッシュ最適化の観点から考えると，単純な AoS や SoA ではない構造の配列を用いることで高い性能が期待できることが分かった.

近年スーパーコンピュータで懸念される消費電力について ITO を用いて調べた結果，MHD シミュレーションコードは TDP よりかなり低い消費電力で動作していたが，

ノード間だけでは無く，単一ノード内での CPU 周波数などのばらつきが多くあり，単一 CPU での動的電力調整が強く働いていることが示された．このばらつきにより，計算時間も変わるため，注意が必要と思われる．

謝辞 本研究は，九州大学情報基盤研究開発センター平成 29/30 年度先端的計算科学研究プロジェクトの支援による．

参考文献

[1] Top500 Supercomputing Sites. (http://www.top500.org/)
[2] Ogino. T, R. J. Walker, M. Ashour-Abdalla, "A global magnetohydrodynamic simulation of the magnetopause when the interplanetary magnetic field is northward", IEEE Trans. Plasma Sci. vol. 20, 1992, 817-828.
[3]九州大学情報基盤研究開発センターWeb ページ（https://www.cc.kyushu-u.ac.jp/scp/index.html）
[4] F. F. Chen, 1974. Introduction to Plasma Physics. Plenum Press, NY.
[5] Fukazawa, K., T. Ogino, and R. J. Walker (2012), "A Magnetohydrodynamic Simula-tion Study of Kronian Field-Aligned Currents and Aurora", J. Geophys. Res., 117, A02214, doi:10.1029/2011JA016945.
[6] Fukazawa, K., T. Nanri and T. Umeda, "Performance Measurements of MHD Simulation for Planetary Magnetosphere on Peta-Scale Computer FX10", Parallel Computing: Accelerating Computational Science and Engineering (CSE), Advances in Parallel Computing 25, pp.387-394, IOS Press, 2014. (DOI: 10.3233/978-1-61499-381-0-387)
[7] Inadomi. Y., et al., “Analyzing and Mitigating the Impact of Manufacturing Variability in Power-Constrained Supercomputing”, Technical Paper, SC'15, Austin (USA).
[8] Fukazawa, K., T. Soga, T. Umeda, T. Nanri, Performance Evaluation and Optimization of MagnetoHydroDynamic Simulation for Planetary Magnetosphere with Xeon Phi KNL, Parallel Computing is Everywhere: Accelerating Computational Science and Engineering (CSE), Advances in Parallel Computing, 178 - 187, DOI:10.3233/978-1-61499-843-3-178, 2018.