jh130007-MD02

さまざまなアーキテクチャからなる計算機システムの性能評価と最適化

深沢圭一郎（九州大学）

概要　本研究では、x86 系、SPARC 系、POWER 系、ベクトル系、GPU 系、MIC 系といった異なるアーキテクチャのコンピュータシステムを利用し、それぞれのコンピュータシステムの専門家と共同研究により、複数のアプリケーションを利用し、システムの実性能評価、最適化を行い、アーキテクチャ毎に共通のチューニング手法・共通の並列化手法をまとめることを目的としている。現在までに MHD コードを用いて異なる CPU アーキテクチャを持つ計算機システム上で性能評価を実施した。今後は JetFDTD を用いた性能評価を行っていく。

1. 研究の目的と意義

現在いわゆるスーパーコンピュータと呼ばれる大規模計算機システムは x86 系、SPARC 系、POWER 系、ベクトル系、GPU 系などさまざまなアーキテクチャから構成されている。さらに近年では ARM 系や MIC などから構成される新しいスーパーコンピュータシステムも出てきている。これらのコンピュータシステムでは計算コアのアーキテクチャが異なるため、アプリケーションによってはそのシステムに対して向き不向きがあり、また性能チューニングも各アーキテクチャにより基本的には異なる。そのため、今までと異なるコンピュータシステムにアプリケーションの移植を行うことは非常に手間のかかる作業となっている。また、科学アプリケーションを利用して研究を行っている計算機ユーザにとって移植作業は時間がかかるだけで、その作業自体は成果にならないため、移植が行われにくい状況にある。

そこで本研究では、JHPCN で利用可能である x86 系、SPARC 系、POWER 系、ベクトル系、GPU 系といった異なるアーキテクチャのコンピュータシステムを利用し、それぞれのコンピュータシステムの専門家と共同研究により、システムの実性能評価、最適化を行い、アーキテクチャ毎に共通のチューニング手法・共通の並列化手法をまとめることを目標とする。さらに、各アーキテクチャ向けの最適化を異なるアーキテクチャのコンピュータシステムにも施し、実効性能が実際に最適化時とどれほど違うのか明らかにする。また、それら結果から、ベクトル機向けコードは GPU に向いているなど各アーキテクチャ間での最適化手法に類似性があるか明らかにする。

対象とするアプリケーションとしては汎用性の高い、流体系、粒子系、電磁界解析等のアプリケーションを用いて、各スーパーコンピュータシステムに精通している研究者と協力し、アプリケーションを各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並列化・最適化し、一般への還元が可能なレベルでの性能最適化、その評価結果を公開することも目標とする。

2000 年以前では大規模計算機システムとしてはベクトル型 CPU が主流であったが、近年ではスカラ型 CPU によるクラスタ型超並列計算機が主流となっている。しかし、これまでのベクトル型 CPU による並列計算機とは異なり、スカラ型 CPU で高い実効性能を達成することは容易ではなく、また並列数が格段に上がったために高い並列効率を達成することが容易ではないという 2 つの問題が生じている。例えば、これまで地球シミュレータなどのベクトル型 CPU において高い実効効率を誇っていたコードが、スカラ型 CPU においては実効性能が低い例は少なくない。また、スカラ型 CPU であっても、CPU アーキテクチャが異なれば、チューニングの手法も異なるため、同じ最適化で実行効率が常に高くなるとは限らない。一方で、近年開発されたアプリケーションは x86 型の計算機シ

ステムに最適化されていることが多く、ベクトル型 CPU では高メモリバンド幅を使い切れないという問題もある。

並列化の問題に関しては、1000 個以上の CPU コアを用いた計算を日常的に実行できる環境が日本にあまり存在しないために、どのシステムにおいても並列化のスケーラビィティが保障されるコードの開発が困難となっている。

また、GPU からなるコンピュータシステムでは CPU 向けの最適化とは全く異なる最適化が必要となり、単に移植しただけでは GPU の高計算能力を活用することはできない。GPU を用いた並列計算においても、GPU 計算機システムの構成を理解した上で、並列化を施さなければ、高いスケーラビリティは得られない。

特に重要なことは、使用可能な最大の CPU、GPU 数で、さらにメモリをフルに使った条件まで含めて実証実験を行って確信を得ることである。さまざまな CPU 環境、GPU 環境において高並列度におけるベンチマークテストを行い、共通のチューニング手法・共通の並列化手法を見出すことが、さまざまなスーパーコンピュータシステムを使うことのできる学際大規模共同研究の大きな意義であると言える

2. 当拠点公募型共同研究として実施した意義

(1) 共同研究を実施した拠点名および役割分担

・北海道大学　SR16000/VM2 を利用した MHD シミュレーションの性能評価と最適化（大宮 学）
・東北大学　SX-9 を利用した MHD シミュレーション JetFDTD の性能評価（江川 隆輔）
・東京大学　FX10 を利用した MHD シミュレーション、JetFDTD の性能評価（片桐 孝洋）
・京都大学　XE6 を利用した MHD シミュレーション、JetFDTD の性能評価（岩下 武史）
・九州大学　CX400 を利用した MHD シミュレーション、JetFDTD の性能評価（深沢 圭一郎）

(2) 共同研究分野

超大規模数値計算系応用分野、超大規模情報システム関連研究分野

(3) 当公募型共同研究ならではという事項など

ユーザ開発で、実際に研究に利用しているシミュレーションコードを本研究課題のように多数の計算機システム上で、そのシステムを理解している教員が性能評価する共同研究は他では難しい。

3. 研究成果の詳細と当初計画の達成状況

(1) 研究成果の詳細について

3. 1. MHD シミュレーションコードの性能評価

まず、MHD シミュレーションコードの性能評価を様々な計算機システムで行った。評価に利用した MHD シミュレーションコードは惑星磁気圏の構造、ダイナミクスを調べるために利用されているコードである。このコードは電磁流体（MHD）方程式からなり、簡単に言えば、一般の流体方程式に電磁力を考慮した方程式である。この MHD コードは有限差分法を利用している。並列化には MPI を使用した（Flat MPI）。並列化手法としては MHD 方程式で解く 3 次元空間を分割する領域分割法を用いた。

一般的にスカラ機で性能を出す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データアクセス時に、その前後含めて数 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ので、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 $f(i, j, k, m)$（これを Type A とする）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 $f(m, i, j, k)$（これを Type B とする）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。そのため、本性能評価

においてもこの配列定義を使った性能評価も行った。

本性能評価では、北海道大学 SR16000/M1（POWER 系）、東北大学 SX-9（ベクトル系）、東京大学 FX10（SPARC 系）、京都大学 XE6（x86 系 Opteron）、九州大学 CX400（x86 系 Xeon）を利用した。

今回の性能評価では 64 MB/コアの配列を計算するが、MHD 方程式を差分法で解くための作業配列として 192 MB/コアを追加で使用した。惑星磁気圏を解く MHD シミュレーションでは、weak scaling が重要なため、コア当たりのメモリサイズは不変とした。プログラム言語は Fortran を利用している。また流体の差分計算が主であるため、並列化に伴う通信は袖領域の通信が支配的である。

3. 1. 1. 北海道大学 SR16000/M1 の性能評価

北海道大学 SR16000/M1 では、最大 4、096 コア（128 ノード）を用いて性能評価を行った。コンパイラは日立最適化 FORTRAN を利用した。コンパイラオプションは以下の通りである。

-model=M1 -Oss -parallel=0 -divopt -pvfunc=0 -looptiling -nolimit -noscope -loglist

図 1 に MHD コードの SR16000/M1 における実効性能を載せる。横軸が利用コア数、縦軸が実効性能領を示す。POWER7 では SMT（Simultaneous Multithreading）が利用できるため、SMT を利用して、1 コアに 2 プロセス立ち上げた場合も評価した。図 1 より SR16000/M1 ではキャッシュヒット効率を高めた Type B の配列が Type A よりも性能が良い（SMT の有無に関係なく）。このとき、SMT 無しの場合で 14.2 TFlops（実行効率 11.4 %）、SMT 有りで、19.5 TFlops（15.5 %）を達成した。Type A は SMT 無しで、6.4 TFlops と Type B に比べて、半分未満の性能にとどまっている。

また、SR16000/M1 では MPI の隣接通信を非ブロッキング通信に変えることで、1%ほど実効性能が上がることが確認されている。

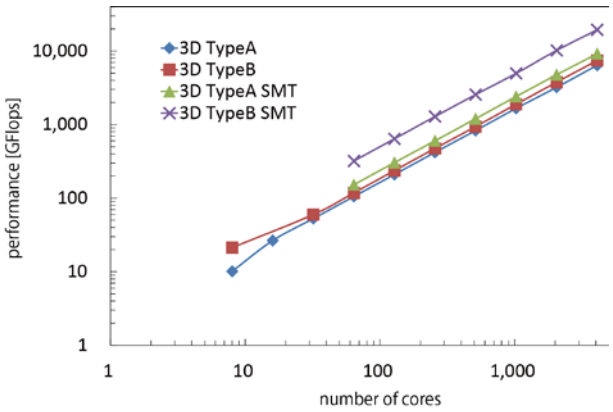

図 1　SR16000/M1 における MHD コードの実効性能

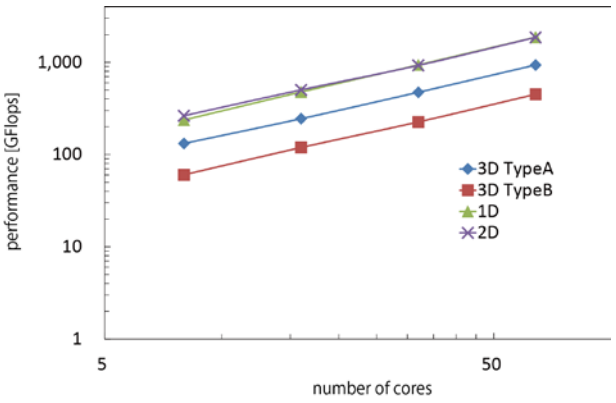

図 2　SX-9 における MHD コードの実効性能

3. 1. 2. 東北大学 SX-9 の性能評価

東北大学 SX-9 では、最大 64 コア（4 ノード）を用いて性能評価を行った。コンパイラは NEC FORTRAN90/SX を利用した。コンパイラオプションは以下の通りである。

-Chopt -size_t64 –Wl，-h，8T_memlayout

図 2 に SX-9 における MHD コードの実効性能を載せる。横軸が利用コア数、縦軸が実効性能領を示す。本研究では 3 次元領域分割 Type A と B の評価だけだが、1 次元、2 次元領域分割の性能が良いので、参考までに結果を図 2 に載せている。図 2 より SX-9 では 3 次元領域分割 Type A の性能（935.2 GFlops）が Type B（449.7 GFlops）よりも明らかに高い。これは SR16000/M1 と逆の結果となっている。ただし、今回の評価ではコア当たりの利用配列が小さいため、3 次元領域分割ではループ長が長くなりにくいため、1 次元、2 次元領

域分割に比べて性能が悪くなった。それぞれ最高で 1.9 TFlops（28 %）の性能を達成している。これも 3 次元領域分割の結果と倍程度差があり、ベクトル機の特徴に合わせた最適化が重要とわかる。

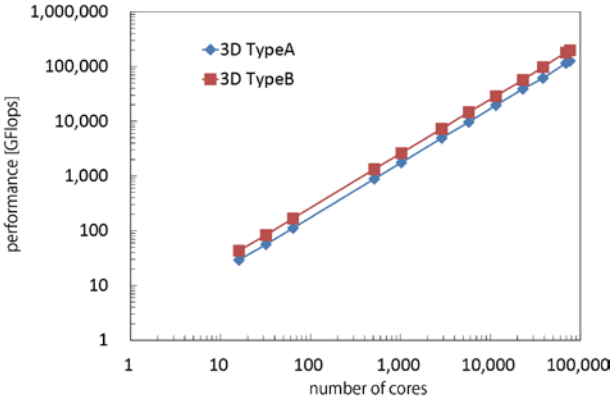

図 3　FX10 における MHD コードの実効性能

3. 1. 3. 東京大学 FX10 の性能評価

東京大学 FX10 では、最大 76,800 コア（4,800 ノード）を用いて性能評価を行った。コンパイラは Fujitsu Technical Computing Suite v1.0 を利用した。コンパイラオプションは以下の通りである。

-x3000 -Kfast, SPARC64IXfx, nomfunc, noalias=s, fsimple, prefetch_indirect, prefetch_strong, noparallel, array_private

図 3 に MHD コードの FX10 における実効性能を載せる。横軸が利用コア数、縦軸が実効性能領を示す。FX10 では、SR16000/M1 と同様に Type B の方が Type A よりも性能が高い。最大で 76,800 コア利用時に 198.2 TFlops（17.5 %）を達成している。一方で、Type B では半分近くの性能である 125.9 TFlops となっている。ここでは詳細は述べないが、OpenMP と非同期通信を利用することで、5 %ほどの実行効率向上が可能であった。

3. 1. 4. 京都大学 XE6 の性能評価

京都大学 XE6 では、最大 8,192 コア（256 ノード）を用いて性能評価を行った。コンパイラは CRAY コンパイラを利用した。コンパイラオプションは以下の通りである。

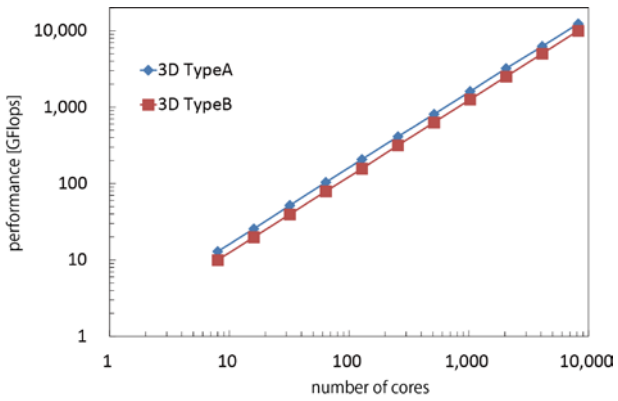

図 4　XE6 における MHD コードの実効性能

-O3,ipa5,aggress -em -h pic -dynamic -h noomp -h msgs -h negmsg

図 4 に MHD コードの XE6 における実効性能を載せる。横軸が利用コア数、縦軸が実効性能領を示す。XE6 では今までの結果より Type A と Type B に性能の差がないが、Type A の性能（12.3 TFlops）が Type B（9.9 TFlops）より良かった。さらに詳細は示さないが、2 次元領域分割で性能が最も良くなり、最大で 14.3 TFlops を達成している。このことから XE6、Opteron 6000 シリーズはベクトル機向け最適化が効果的と言える。

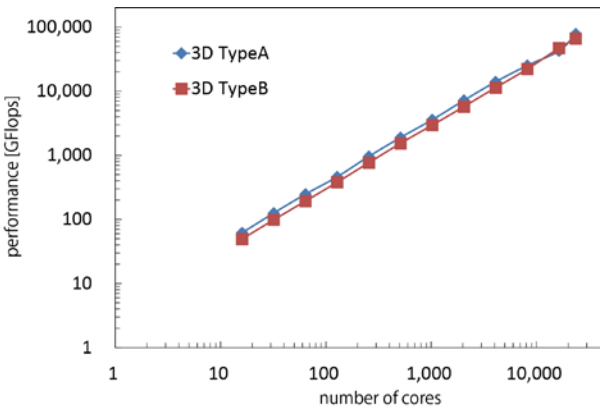

図 5　CX400 における MHD コードの実効性能

3. 1. 5. 九州大学 CX400 の性能評価

九州大学 CX400 では、最大 23,616 コア（1,476 ノード）を用いて性能評価を行った。コンパイラは Fujitsu Technical Computing Suite v1.0 を利用した。コンパイラオプションは以下の通りである。

表 1　様々な計算機システムにおける性能の傾向

	Core/CPU	Rpeak [TFlops]	Rmax [TFlops]	Rpeak /CPU [Gflops]	Efficiency [%]	Suitable domain decomposition	CPU architecture
SX-8R	8/8	0.08	0.28	10.0	28	1D	Vector
SX-9	64/64	2.19	6.55	34.2	33	2D	Vector
HA8000	8,192/1,024	10.04	75.37	9.8	13	3D_A	Opteron (Barcelona)
HX600	1,024/256	2.17	10.24	8.5	21	3D_A	Opteron (Shanghai)
XE6	8,192/512	14.16	81.92	27.7	17	1D or 2D	Opteron (Interlagos)
RX200S6	864/144	3.51	10.13	24.4	35	3D_A	Xeon (Westmere)
RX200S3	1,536/768	2.54	18.43	3.3	14	3D_A	Xeon (Woodcrest)
CX400	23,616/2,952	104.23	510.11	35.3	20	3D_A	Xeon (Sandy Bridge)
FX1	1,024/256	2.08	10.24	8.1	21	3D_B	SPARC64VII
FX10	76,800/4,800	234.59	1135.41	48.9	21	3D_B	SPARC64 IXfx
K	262,144/32,768	914.12	4,194.30	27.9	22	3D_B	SPARC64 VIIIfx
SR16000/L2	1,344/672	5.38	25.27	8.0	21	3D_B	POWER6
SR16000/M1	4,096/512	19.49	125.50	38.1	16	3D_B	POWER7
Xeon Phi 5110P	60/1	0.049	1.01	49.0	5	3D_A	Knights Corner

-x3000 -Kfast, nomfunc, noalias=s, fsimple 、*noparallel*

図 5 に MHD コードの CX400 における実効性能を載せる。横軸がコア数、縦軸が実効性能領を示す。この計算機システムにおいても Type A と Type B の性能差が少ない。図ではほぼ重なって

見えるが、Type A で最高性能を達成し、78.0 TFlops となった。Type B では 66.3 TFlops になっている。この計算機システムは 256 ノードを一グループにし、グループ内と外で通信帯域に差があるため、利用コア数が延びるほど、スケーラビリティが下がっている様子が見える。

3. 1. 6. 他システムとの比較

今回行った 5 つの計算機システムの性能評価結果と今までに性能評価を行ってきた近年の計算機システムを比較するために、最大実行性能、CPU 当たりの実効性能や最適な領域分割手法を表 2 にまとめた。この表では FX10 は Hybrid MPI＋非同期通信の結果、XE6 は 2 次元領域分割の結果、CX400 は非同期通信の結果から値を持ってきている。

表を見ると、ベクトル機はベクトル長が長く取れる 2 次元領域分割で性能が出ており、RISC プロセッサである POWER 系と SPARC 系ではキャッシュヒットを考慮した 3 次元領域分割 Type B で性能が出ている。x86 系である Xeon 系、Opteron 系ではあまりキャッシュチューニングは効果が無く、ベクトル的な領域分割が最適という結果になっている。さらに近年発表された x86 系アーキテクチャを持つ MIC である Xeon Phi でもベクトル的な最適化が効果的という結果になっている。

一般的にベクトル機はメモリバンド幅が大きく、B/F（Bandwidth / Flops）値が大きいため、実行効率が高いという傾向だったが、SX-9 では実効効率が下がっており、33 %となっている。近年スカラチップは実行効率が上がってきており、Westmere 世代の Xeon ではベクトルチップと同様の実行効率を達成している。また、AVX により SIMD が倍になり理論性能が大きく上がった Sandy Bridge 世代 Xeon においても 20 %程度の実行効率を達成している。SPARC 系では、FX10 や京で FX1 と同様の実効効率を達成しており、

CPU あたりでは Xeon 系に勝る性能を達成している。

CPU 当たりの性能を比べてみると、今回性能を評価した FX10 と Xeon Phi ではほぼ同じ性能を持っていることが分かる。この Xeon Phi は Native で利用している。また CX400 は K より CPU 当たりの性能は高く、SX-9 の 1CPU よりも高い性能となっている。

3. 2. JetFDTD の性能評価

JetFDTD はマクスウェルの偏微分方程式を中心差分により離散化し、時間領域で電磁界の変化を解析するアプリケーションである。解析アルゴリズムが並列計算処理向きであることから、これら特徴を活かした大規模計算が高周波デバイスや光学デバイスの設計などにおいて期待されている。しかし、空間をセルと呼ばれる微小要素で離散化し，時間領域において定常状態に達するまでの解析を行うことから、計算機シミュレーションには大規模な主記憶容量と多数のコアなど計算リソースと長い時間にわたる解析が必要である。すでに、SR16000/M1 向けのチューニングを行い、2560 コアおよび主記憶容量 6.1TB を利用した大規模解析においても合理的な時間内に安定した解が得られることを確認している。そのため、本課題では、SX-9、FX10、XE6、CX400 において JetFDTD の性能評価を行い、SR16000/M1 と性能を比べた。性能評価において、計算サイズを 147x85x858 に設定し、strong scaling で評価を行った。また並列化手法は MPI と自動並列を用いており、SMP 並列とハイブリッド MPI 並列（自動並列＋MPI 並列）の 2 種類を評価した。JetFDTD に関しては、実行時間だけで、実効性能、実行効率は測定できていないので、TFlop という実行時間×計算機の性能を用いて評価する。TFlop は数値が小さいほど性能が良いことをここでは表している。

・東北大学 SX-9（1CPU：102.4GFlops、1node: 1638.4GFlops）
SMP 並列をシリアル実行すると 24 分、16 スレッドで 2 分の実行時間だった。MPI 並列は動作に問題があったためここでは計測していない。16 スレッド時の実行時間を利用して、TFlop は 192 TFlop（Elapse × GFlops）となる。

・東京大学 FX10（1CPU：102.4GFlops、1node: 1638.4GFlops）
SMP 並列を 16 スレッドで実行すると 2 時間 23 分かかった。MPI 並列で 16 プロセス×1 スレッドで実行すると 18 分の実行時間となった。性能の良い MPI 並列実行結果から TFlop を計算すると、250.4 TFlop となる。

・京都大学 XE6（1core：10GFlops、1node：320GFlops）
SMP 並列においてシリアル実行で 4 時間 10 分、16 スレッドで 1 時間 15 分の実行時間となった。MPI 並列で 4 プロセス×8 スレッドで実行したところ 38 分の実行時間となった。MPI 並列結果から TFlop を計算すると、729.6 TFlop となる。

・九州大学 CX400（1core：21.6GFlops、1node:345.6GFlops）
SMP 並列においてシリアル実行で 2 時間 41 分、8 スレッドで 57 分、16 スレッドで 29 分の実行時間であった。MPI 並列では、4 プロセス×8 スレッドで 14 分の実行時間となった。MPI 並列の結果から TFlop を計算すると、580.6 TFlop となる。

・北海道大学 SR16000/M1（1core：30.64GFlops、1node: 980.48GFlops）
SMP 並列において 32 スレッド実行すると 12 分の実行時間となり、MPI 並列で 5 プロセス×32 スレッドで実行すると 4 分の実行時間となった。MPI 並列から TFlop を計算すると 705.9 TFlop となる。

上記の結果から相対的に JetFDTD の各システムでの性能を見るために、TFlop の値を見てみると、

SX-9 が 192TFlop、FX10 が 250.4TFlop、CX400 が 580.6TFlop、SR16000/M1 が 705.9TFlop、XE6 が 729.6TFlop となる。これらはノードあたりの JetFDTD をどれだけ早く、少ない性能で実行したかを表しており、圧倒的にベクトル機である SX-9 での性能が高いことがわかる。一方でスカラ機の FX10 も SX-9 についで良い性能を見せている。他のシステムは約 580～730TFlop とそれほど差が無く、SX-9 と FX10 に比べて性能が出ていないことがわかる。この結果から特に X86 系の計算機システムでは今までとは異なった性能最適化が必要と言うことがわかる。

また、計算機システムによって、SMP 並列の性能が悪いものもあり、各システムに合わせた実行方法が必要なことがわかる。一方で MPI によるプロセス並列の並列化性能はどのシステムでも良いことがわかった。

(2) 当初計画の達成状況について

計画では格子系電磁流体アプリケーション（MHD）、粒子系アプリケーション、JetFDTD という 3 種類のアプリケーションについて性能評価を行う予定であり、そのうち 2 つのアプリケーションに対して性能評価を行ったが、粒子系アプリケーションについては評価を行えなかった。

また、性能評価を実施する計算機システムが格子系電磁流体アプリケーションでは CPU 系で SR16000/M1、SX-9、FX10、XE6、CX400、GPU 系で TSUBAME2.0 の計画であったが、アプリケーションの GPU 化に困難があったため、GPU を用いた性能評価は行えなかった。その他のシステムでは問題なく計画通り性能評価を実施することができた。

4. 今後の展望

本研究課題で蓄積できた情報をもとに、MHD シミュレーションコードと JetFDTD の様々なシステムに対する最適化を実施し、共通の最適化手法、各アプリケーションに効果的な最適化手法などをまとめていく。

5. 研究成果リスト

(1) 学術論文（投稿中のものは「投稿中」と明記）

特になし

(2) 国際会議プロシーディングス

1) 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)

2) Fukazawa, K., T. Nanri, and T. Umeda, Performance evaluation of magnetohydrodynamics simulation for magnetosphere on K computer, In: AsiaSim 2013, Communications in Computer and Information Science, Vol.402, edited by G. Tan, G. K. Yeo, S. J. Turner, and Y. M. Teo, pp.570-576, Springer-Verlag Berlin Heidelberg, 2013. (ISBN: 978-3-642-45036-5) (DOI: 10.1007/978-3-642-45037-2_61)

(3) 国際会議発表

1) K. Fukazawa, T. Nanri and T. Umeda, "Performance Measurements of MHD Simulation for Planetary Magnetosphere on Peta-Scale Computer FX10", International Conference on Parallel Computing 2013, 10 - 13 Sep. 2013, Munich, Germany.

2) T. Umeda, K. Fukazawa, “Performance Measurement of Parallel Vlasov Code for Space Plasma on Scalar-Type Supercomputer Systems with Large

Number of Cores", AsiaSim2013, 6 – 8 Nov. 2013, Singapore.

3) K. Fukazawa, T. Nanri, T. Umeda, "Performance Evaluation of Magnetohydrodynamics Simulation for Magnetosphere on K computer", AsiaSim2013, 6 – 8 Nov. 2013, Singapore.

(4) 国内会議発表

1) 梅田隆行、深沢圭一郎、"京、FX10 及び CX400 におけるブラソフコードの性能チューニング"、第 141 回ハイパフォーマンスコンピューティング研究発表会、沖縄、2013 年 9 月 30 日 - 10 月 1 日.

2) 深沢圭一郎、岡慶太郎、"電磁流体コードを用いた Xeon Phi の性能評価"、第 141 回ハイパフォーマンスコンピューティング研究発表会、沖縄、2013 年 9 月 30 日 - 10 月 1 日.

3) 深沢圭一郎," Performance Measurement of MHD Simulation Code for Planetary Magnetosphere on Xeon Phi"、 地球電磁気・地球惑星圏学会　第１３４回総会及び講演会、高知、2013 年 11 月 2 日 –5 日.

(5) その他（特許，プレス発表，著書等）

特になし