

5. GeoFEM-Cube の性能

中 島 研 吾

東京大学情報基盤センター

1. 概 要

HITACHI SR16000/M1 (Yayoi) 1 ノード (4 プロセッサ, 32 コア) (図 1 参照) [1] を使用して, GeoFEM プロジェクト [2] で開発された並列有限要素法アプリケーションを元に整備した性能評価のためのベンチマークプログラム GeoFEM-Cube [3] による評価を実施した.

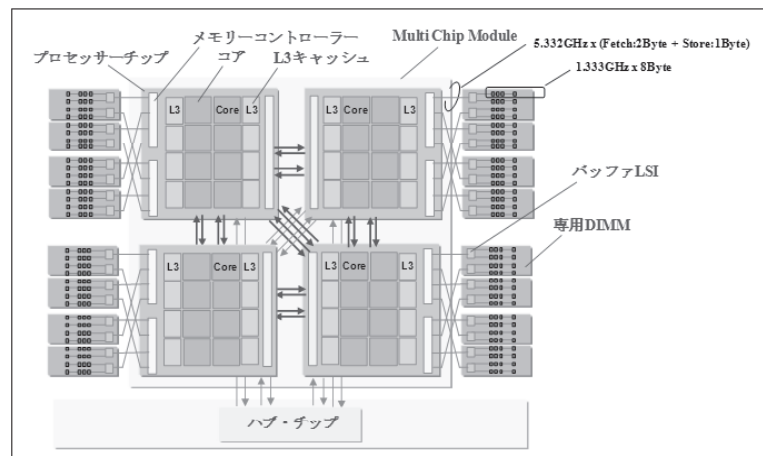


図 1 HITACHI SR16000/M1 (Yayoi) のノード構成図 [1]

オリジナルの GeoFEM ベンチマーク [4] は,

- ① 三次元弾性静解析問題 (Cube 型モデル, PGA モデル)
- ② 三次元接触問題
- ③ 二重球殻間領域三次元ポアソン方程式

に関する並列前処理付き反復法ソルバーの実行時性能 (GFLOPS 値) を様々な条件下で計測するものである。係数行列は全て疎行列であり, memory-bound なベンチマークである。プログラムは全て OpenMP ディレクティブを含む FORTRAN90 および MPI で記述されている。各ベンチマークプログラムでは, GeoFEM で採用されている局所分散データ構造 [2] を使用しており, マルチカラー法等に基づくリオーダーリング手法によりベクトルプロセッサ, SMP, マルチコアプロセッサにおいて高い性能が発揮できるように最適化されている。また, MPI, OpenMP, Hybrid (OpenMP+MPI) の全ての環境で稼動する。

著者等は [4] において, 3 種類の GeoFEM ベンチマークのうち図 2 に示すような一様な物性を有する単純形状 (Cube 型) を対象とした三次元弾性静解析問題について cc-NUMA アーキテ

クチャを有する HITACHI HA8000 クラスシステム (T2K 東大) [1] に対して様々な最適化を試みた。この成果を性能評価用のベンチマークプログラムとして整備したものが GeoFEM-Cube である。

GeoFEM-Cube では、係数行列が対称正定な疎行列となることから、SGS (Symmetric Gauss-Seidel) [4] を前処理手法とし共役勾配法 (Conjugate Gradient, CG) 法によって連立一次方程式を解いている (以下 SGS/CG 法と呼ぶ)。三次元弾性問題では 1 節点あたり 3 つの自由度があるため、これらを 1 つのブロックとして取り扱っている。

連立一次方程式の係数マトリクスの格納法としてオリジナルの GeoFEM ベンチマークでは (a) CRS (Compressed Row Storage), (b) DJDS (Descending order Jagged Diagonal Storage) の 2 種類の方法が準備されているが、GeoFEM-Cube ではスカラプロセッサ向けの CRS 法を使用している。

SGS 前処理では、係数行列 A そのものが前処理行列として利用されるため ILU 分解は実施しないが、前処理における前進後退代入はグローバルなデータ依存性を有するプロセスのため、並列性を抽出するためのリオーダーリングが必要である [4]。GeoFEM ベンチマークでは、マルチカラー法 (Multicoloring, MC) 法, Reverse Cuthill-McKee (RCM) 法, 更に RCM 法にサイクリックに再番号付けする Cyclic マルチカラー法 (cyclic multicoloring, CM) を適用する手法 (CM-RCM) の 3 種類が利用可能となっている。

並列プログラミングモデルとしては各コアを独立に扱う Flat MPI と Hybrid 並列プログラミングモデルの両者を扱うことができる。Hybrid については「Hybrid $a \times b$ (HB $a \times b$)」(a : MPI プロセス当りの OpenMP スレッド数, b : ノード内 MPI プロセス数) という形で、ノード構成に応じてスレッド数, MPI プロセス数を自由に決められるようになっている。

GeoFEM の局所分散データ構造に基づき、局所的なデータは各ローカルメモリに格納されている。HITACHI SR16000/M1 は、T2K (東大) と同様に NUMA (Non Uniform Memory Access) アーキテクチャを有しており、実行時制御コマンド (NUMA control) 使用して、コア (またはプロセッサ) とメモリの関係を明示的に指定することによって、性能が向上することは広く知られている [4]。T2K (東大) では、NUMA control を陽に指定する必要があったが、HITACHI SR16000/M1 では：

- ローカルメモリにデータを確保するための環境変数 (export MEMORY_AFFINITY=MCM)
- CPU を固定的に割り当てるための実行時コマンド (mpibind)

が準備されている。

GeoFEM-Cube では、Hybrid 並列プログラミングモデルにおける性能改善のため、

① First Touch Data Placement の適用 [5]

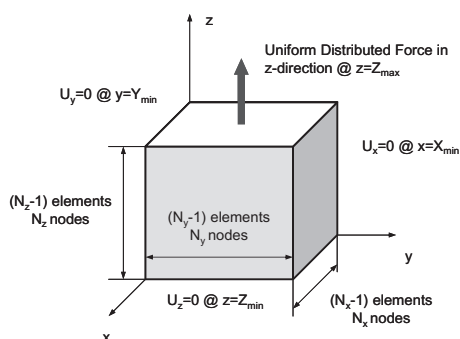


図2 Cube型ベンチマークの境界条件

- ② 連続データアクセス，キャッシュヒット率向上のためのデータ再配置（Sequential リオーダリング，図 3）

を適用している．各最適化手法の詳細については，[4] を参照されたい．

本研究では以下の 3 ケースについて評価を実施した：

- CASE-1：MC，RCM，CM-RCM によるリオーダリングを適用
- CASE-2：更に First Touch Data Placement を適用（Flat MPI は除く）
- CASE-3：更に図 3 に示すデータ再配置を適用（Flat MPI は除く）

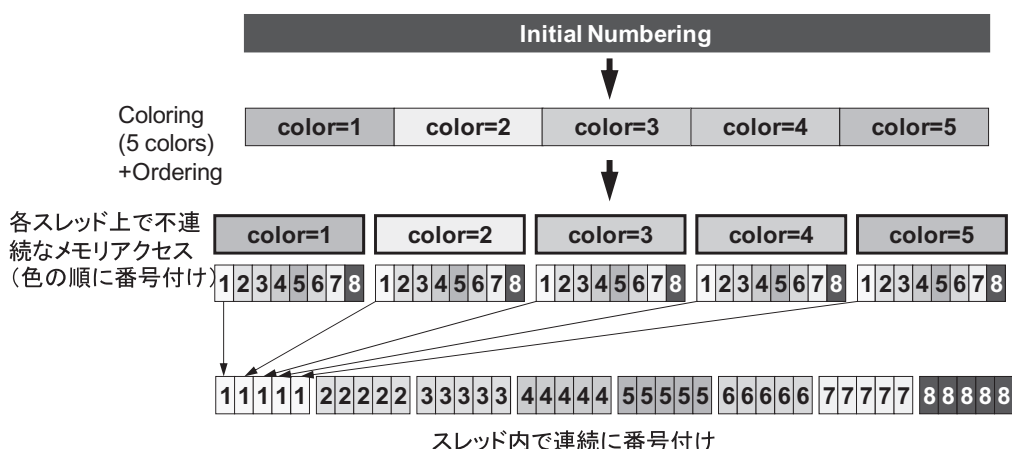


図 3 連続データアクセスのためのデータ再配置（Sequential Reordering）
（5 色，8 スレッドの場合）

2. 性能評価結果

図 4 に 2,097,152 節点（6,291,456 自由度），CM-RCM（色数 10），最適化 CASE-3 における計算結果を示す．1 ノード（4 プロセッサ，32 コア）を使用，スレッド数×コア数=32 であり SMT は適用していない．Flat MPI の他，HB 2×16，4×8，8×4，16×2，32×1（ノード内 OpenMP のみ）を実施した．図 4 で示したのは SGS/CG 法部分の計算性能（GFLOPS 値）である．収束までの反復回数は各並列プログラミングモデルによって若干異なるが数%以内であり，この計算性能が実際の計算時間に直接反映する．Flat MPI の性能が最も良く，以下，OpenMP のスレッド数を増加するに従って性能が低下し，HB 32×1 では約 50%にまで低下している．図 5 は，同じ問題に対する HITACHI SR11000/J2，T2K（東大）における CASE-3 の例である [4]．いずれの場合も Flat MPI と他の手法はほぼ同じ性能を示しており，ノード数を増加させると全般的に Hybrid 並列プログラミングモデルの方が全体的に性能が良いことも既に示されている [6]．

図 6 は HITACHI SR16000/M1 において問題サイズ（節点数）を 100^3 （1,000,000 節点，3,000,000 自由度）から 256^3 （16,777,216 節点，50,331,648 自由度）まで変化させた場合の性能である．リオーダリングは RCM 法によっている．問題サイズを大きくすることによって，全体的な性能は低下するものの，OpenMP のオーバーヘッドが隠蔽されるため Flat MPI と Hybrid 並列プログラミングモデルの差は少なくなり，節点数=256³では，Flat MPI に対して HB 32×1 の性能は約 84%である．

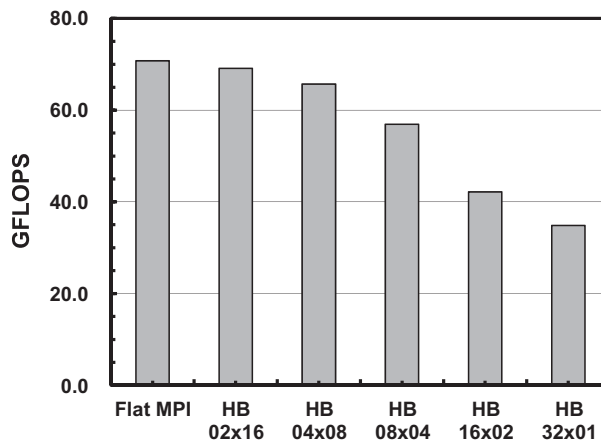
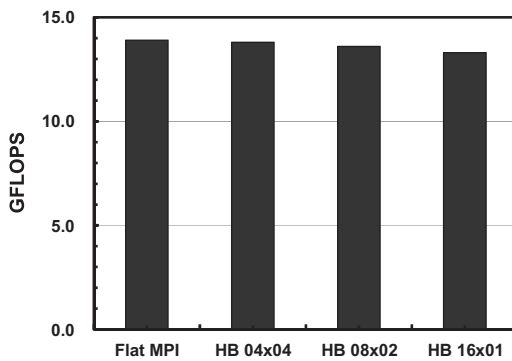
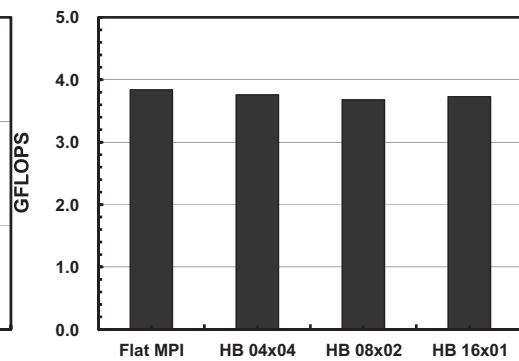


図4 GeoFEM-Cube 性能評価結果 (SGS/CG 法の性能 (GFLOPS 値)), HITACHI SR16000/M1 1 ノード (4 プロセッサ, 32 コア), 2,097,152 節点 ($=128^3$, 6,291,456 自由度), CM-RCM (色数 10), 最適化: CASE-3



(a) HITACHI SR11000/J2 [4]



(b) T2K (東大)

図5 GeoFEM-Cube 性能評価結果 (SGS/CG 法の性能 (GFLOPS 値)), 2,097,152 節点 ($=128^3$, 6,291,456 自由度), CM-RCM (色数 10), 最適化: CASE-3, (a) HITACHI SR11000/J2, (b) T2K (東大)

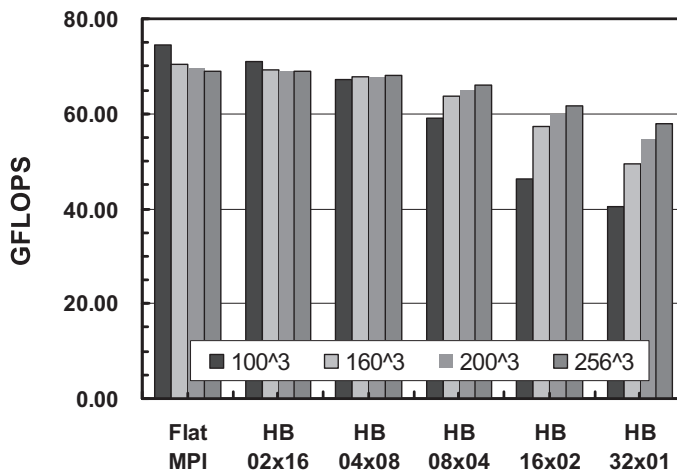


図6 GeoFEM-Cube 性能評価結果 (SGS/CG 法の性能 (GFLOPS 値)), HITACHI SR16000/M1 1 ノード (4 プロセッサ, 32 コア), 節点数: 100^3 (1,000,000 節点, 3,000,000 自由度) ~ 256^3 (16,777,216 節点, 50,331,648 自由度), RCM, 最適化: CASE-3

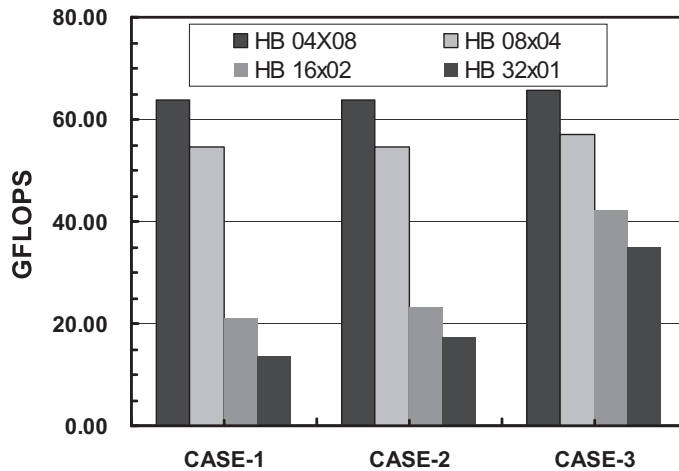


図7 GeoFEM-Cube 性能評価結果 (SGS/CG 法の性能 (GFLOPS 値)), 最適化手法 (CASE-1～3) の効果, HITACHI SR16000/M1 1 ノード (4 プロセッサ, 32 コア), 2,097,152 節点 ($=128^3$, 6,291,456 自由度), CM-RCM (色数 10)

図7 は前項で示した最適化手法 (CASE-1～CASE-3) の効果を各並列プログラミングモデルについて示したものである. 特に MPI プロセス当りのスレッド数が多い HB 16×2, 32×1 においては特に CASE-3 による最適化の効果が顕著であることがわかる.

表1 は, コア当りの問題サイズを 40^3 節点 $=3 \times 64,000 = 192,000$ 自由度とした場合の 1 ノードの性能を東大情報基盤センターの新旧様々な計算機 (HITACHI SR11000/J2, HITACHI SR16000/M1 (Yayoi), HITACHI HA8000 クラスタシステム (T2K 東大), Fujitsu PRIMEHPC FX10 (Oakleaf-FX)) で比較したものである. 並列プログラミングモデルとしてはいずれも Flat MPI を使用している. SR11000/J2 (Power 5+) と SR16K/M1 (Power 7) を比較すると, ノード当りのピーク演算性能は 6.66 倍にもなっているが, メモリバンド幅はわずか 2.62 倍にしかならず, GeoFEM-Cube の性能も 3.83 倍の増加に留まっている. 対ピーク性能比は 12.9% から 7.41% に低下している. SR16000/M1 (Power 7) を Oakleaf-FX (SPARC64 IXfx) と比較すると, Byte/Flop 値はほぼ同じであるが, コア当りキャッシュサイズの大きい SR16000/M1 の方が対ピーク性能比は若干高い.

表1 各計算機環境における GeoFEM-Cube 性能評価結果, 1 ノード, Flat MPI, コア当り問題サイズ: 40^3 節点 $=3 \times 64,000 = 192,000$ 自由度, HITACHI SR11000/J2, HITACHI SR16000/M1 (Yayoi), HITACHI HA8000 クラスタシステム (T2K 東大), Fujitsu PRIMEHPC FX10 (Oakleaf-FX)

	HITACHI SR11000/J2	HITACHI SR16000/M1	T2K 東大	Fujitsu FX10 Oakleaf-FX
Processor	IBM Power 5+ 2.3 GHz	IBM Power 7 3.83 GHz	AMD Opteron 8356 2.3 GHz	SPARC64 IXfx 1.848 GHz
Core #/Node	16	32	16	16
Peak Performance (GFLOPS)	147.2	980.5	147.2	236.5
STREAM Triad (GB/s)	101.0	264.2	20.0	64.7
Byte/Flop	0.686	0.269	0.136	0.274
GeoFEM-Cube (GFLOPS)	19.0	72.7	4.69	16.0
% to Peak	12.9	7.41	3.18	6.77
Last Level Cache/core (MB)	18.0	4.00	2.00	0.75

3. まとめ

HITACHI SR16000/M1 (Yayoi) 1 ノード (4 プロセッサ, 32 コア) を使用して, GeoFEM プロジェクトで開発された並列有限要素法アプリケーションを元に整備した性能評価のためのベンチマークプログラム GeoFEM-Cube による評価を実施した. IBM Power 5+アーキテクチャに基づく HITACHI SR11000/J2 と比較すると, IBM Power 7 は NUMA (Non Uniform Memory Access) アーキテクチャとなったため, 特にスレッド並列を用いた場合に性能を引き出すためのプログラミングが格段に困難となった. また OpenMP によるスレッド並列の性能は Flat MPI と比較して非常に低い. ポストペタスケール時代のシステムでは通信によるオーバーヘッドを削減するため, メッセージパッシングとノード内スレッド並列を組み合わせたハイブリッド並列プログラミングモデルが必須であるが, 当センターの HITACHI SR16000/M1 (Yayoi) は 8 ノードまでの利用を限度として運用しているため, GeoFEM-Cube のような memory-bound なアプリケーションでは, 1 ノード当り性能の高い Flat MPI を利用するのがお勧めである.

HITACHI SR11000/J2 と HITACHI SR16000/M1 では 1 ノードのピーク演算性能の増加と比較してメモリバンド幅の増加が少なく, GeoFEM-Cube のような memory-bound なアプリケーションでの性能が低い. 対ピーク性能比に関しては Fujitsu PRIMEHPC FX10 (Oakleaf-FX) とほぼ同じである. Memory-bound なアプリケーションによって大規模な問題を解きたい利用者は, Oakleaf-FX を使用することを強くお勧めしたい.

参 考 文 献

- [1] 東京大学情報基盤センター : <http://www.cc.u-tokyo.ac.jp/>
- [2] GeoFEM : <http://geofem.tokyo.rist.or.jp/>
- [3] UT-HPC benchmark : <http://www.cspp.cc.u-tokyo.ac.jp/ut-hpc-benchmark/>
- [4] 中島研吾, 片桐孝洋 : マルチコアプロセッサにおけるリオーダーリング付き非構造格子向け前処理付反復法の性能, 情報処理学会研究報告 (HPC-120-6) (2009)
- [5] Mattson, T.G., Sanders, B.A., Massingill, B.L.: Patterns for Parallel Programming, Software Patterns Series (SPS), Addison-Wesley (2005)
- [6] Nakajima, K.: New Strategy for Coarse Grid Solvers in Parallel Multigrid Methods using OpenMP/MPI Hybrid Programming Models, ACM Proceedings of PPOPP/PMAM 2012, New
- [7] Nakajima, K.: New strategy for coarse grid solvers in parallel multigrid methods using OpenMP/MPI hybrid programming models, ACM Proceedings of the 2012 International Workshop on Programming Models and Applications for Multicores and Manycores, ACM Digital Library (DOI: 10.1145/2141702.2141713) (2012)