Hitachi SR8000/MPP フルノードによる並列有限要素法コード「GeoFEM」におけるMPI/OpenMPハイブリッド並列ソルバーの性能評価

中島 研吾（財団法人 高度情報科学技術研究機構，nakajima@tokyo.rist.or.jp）
奥田 洋司（東京大学大学院工学系研究科 システム量子工学専攻，okuda@q.t.u-tokyo.ac.jp）

1. はじめに

近年，超並列計算機の世界ではHitachi SR8000,「地球シミュレータ」[1]などに代表されるような，複数のプロセッサが共有メモリユニットを構成するSMP（Symmetric Multiprocessor）クラスタ型のアーキテクチュアが一般的である。アメリカのASCI（Accelerated Strategic Computing Initiative）[2]のハードウェア群もすべてこのSMPクラスタ型を採用している。このようなアーキテクチュアにおいてはLoop DirectiveとMessage Passingを組み合わせたいわゆるHybrid型の並列プログラミングが適していると考えられる。すなわち，SMPノード（共有メモリユニット）間の通信にはMPI[3]などのMessage Passing，SMPノード内の並列化にはOpenMP[4]などのLoop Directiveを使用するというものである。このようなプログラミングモデルに関する研究[5][6]はここ数年数多く実施されているが，多くはNAS Parallel Benchmarks（NPB）[7]に代表される構造格子を対象としたもので非構造格子を使用した例はほとんどない。

本稿では有限要素法,すなわち非構造格子における前処理付き並列反復法ソルバーに対してHybrid型並列プログラミングモデルを適用し，東京大学情報基盤センターHitachi SR8000/MPPフルノード（128SMPノード，1024プロセッサ）を使用して最大805,306,368自由度までの計算を実施した例を示す。使用した有限要素法コードおよび線形ソルバーは財団法人高度情報科学技術研究機構によって開発されている「GeoFEM」[8]の一部である。GeoFEMは文部科学省科学技術振興調整費「高精度の地球変動予測のための並列ソフトウエア開発に関する研究」の固体地球分野において，固体地球変動予測シミュレーションのための並列有限要素法プラットフォームとして開発されている。

以下，（1）反復法を使用した有限要素法の並列化，（2）並列計算用局所前処理手法について，「GeoFEM」における実装例を述べた後，（3）SMPクラスタ型並列計算機向けに最適化されたソルバーの概要と計算結果について述べる。

2. 反復法を使用した有限要素法の並列化[8][9]

並列計算で扱うデータのサイズ（メッシュ数）は非常に大きいため，全体領域を一括して取り扱うことは困難で，全体データを部分領域（局所データ）に分割する必要がある。有限要素法は差分法などと比較して並列化が困難であると考えられてきた。間接データ参照があるため，1プロセッサ（Processing Element，PE）あたりの計算効率は差分法と比較して低いが，有限要素法の処理は基本的に要素単位の局所的な処理であり,領域間の通信は図1に示すように線形ソルバーの部分だけで生じる。この特性を最大限利用し，適切なデータ構造を設定，並列計算に適した反復法を採用することによって後述するように95%を越えるような並列化効率を達成することも可能である。

GeoFEMでは領域間の通信の記述にはMPIを使用している。差分法などに使用されている構造格子（Structured Grids）に関してはMPI固有の領域間通信用のサブルーチンが準備されているが，有限

要素法に代表される非構造格子（Unstructured Grids）では，プログラム開発者が独自にデータ構造と領域間通信を設計しなくてはならない。

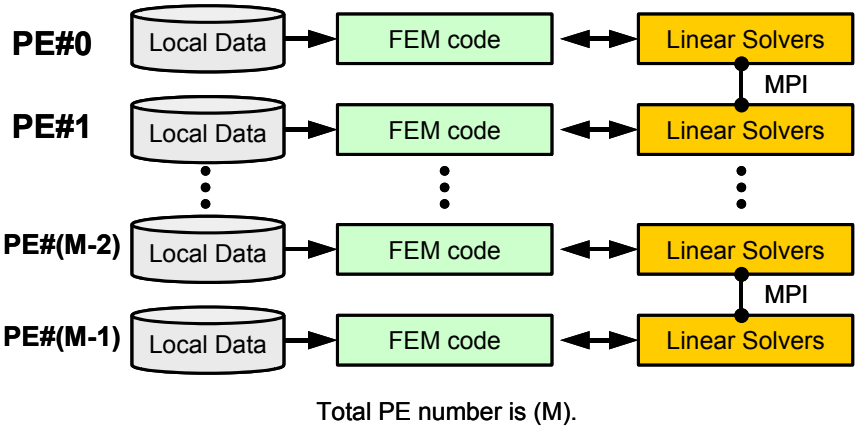


図１．並列有限要素法の処理
有限要素法の処理は要素単位の局所的な処理であり並列化が容易。通信は線形ソルバーでのみ発生。

GeoFEM では，領域間の負荷バランスを考慮して全体領域を「節点ベース（Node-Based）」に分割している。すなわち，各部分領域，各プロセッサ（PE）で扱う節点数が均等になるように領域を分割している。（図 2（a）参照）。

節点ベースの領域分割では各局所データは以下の情報を含んでいる：

（1） 本来その領域に割り当てられた節点
（2） （1）の節点を含む要素
（3） （2）の要素に含まれる節点のうち（1）に含まれないもの
（4） 領域間の通信テーブル（Communication Table）
（5） その他，節点／要素／面グループ等

上記のうち（2），（3）の情報は各領域において，要素単位のマトリクス生成を実施するために必要な情報である。そのため図 2（b）に示すように各領域境界では各領域によって共有されるオーバーラップ要素が生じる。

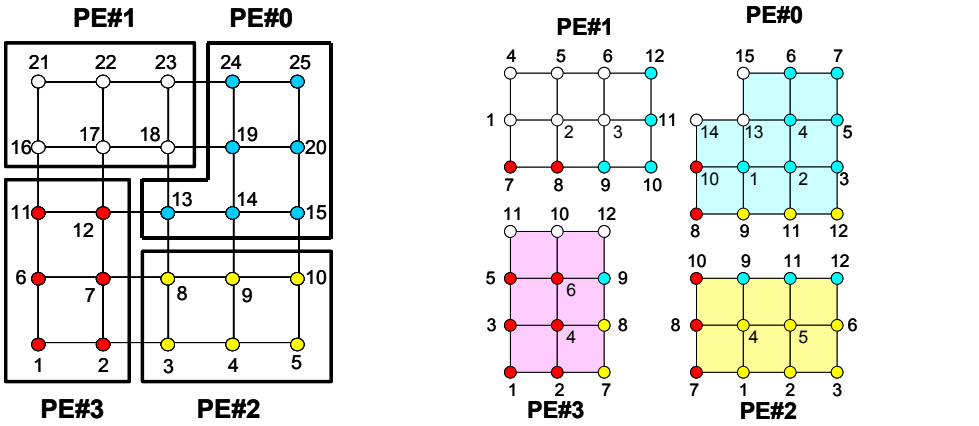
部分領域内の節点は領域間通信の見地から以下の 3 種類に分類される：

- 内　点　（Internal Nodes，上記の（1））
- 外　点　（External Nodes，上記の（3））
- 境界点　（Boundary Nodes，「内点」のうちで他領域の「外点」となっている点）

例えば図 2（b）の「PE#2」の領域に注目すると，内点：{1，2，3，4，5，6}，外点：{7，8，9，10，11，12}，境界点：{1，4，5，6} となる。

隣接する領域間の通信に関して記述しているデータが「通信テーブル（Communication Tables）」

である。「境界点」のデータが各隣接領域に「送信（send)」され，送信先の各領域では「外点」データとして「受信（receive)」される（図 3 参照)。



（a）全体メッシュと内点

（b）各領域に属する内点，要素，外点
領域間でオーバーラップする要素が生じる

図2．GeoFEMの局所データ：節点単位の領域分割

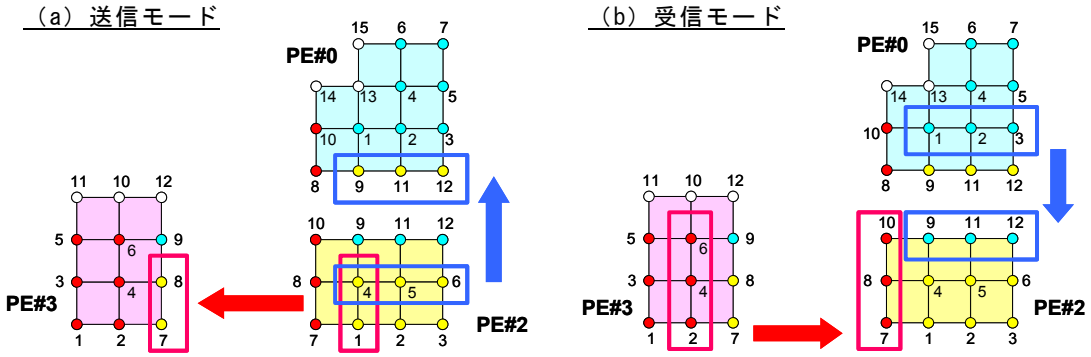


図3．GeoFEMの局所データ：通信テーブル

3．並列計算用局所前処理手法[8][9]

図 1 にも記述されているように，GeoFEM では領域間の通信は線形ソルバーの部分でのみ生じる。前処理つき反復法における計算プロセスは以下の 4 種類に分類される：

（1） 列ベクトル積
（2） ベクトル～ベクトル内積
（3） ベクトル（およびその実数倍）の加減
（4） 前処理

このうち（3）を除く各プロセスでは領域間の通信が発生する。（1）は計算前に 2．で述べた通信を

実施すれば局所的な処理が可能である。(2)は MPI_ALLREDUCE などのサブルーチンを使用して容易に達成可能である。

(4)については前処理手法によって異なる。例えば代表的な前処理手法である不完全 LU 分解（Incomplete LU Factorization, ILU），不完全コレスキー分解（Incomplete Cholesky Factorization, IC）などの手法は前進／後退代入により大域的な変数の依存性が生じるため，並列化が困難である。単独プロセッサを使用した計算の場合，Fill-in のない ILU(0)法を前処理として使用すると，前処理計算部分が全体の 50%程度を占めるため[9]，前進／後退代入部分の並列化は計算効率の向上のために不可欠である。

GeoFEM では局所前処理法（局所 ILU(0)法，Localized ILU(0)）[4]を使用している。局所 ILU(0)法は一種の「擬似」ILU(0)法である。局所 ILU(0)法では前進／後退代入計算時に領域外からの影響(すなわち外点の影響）を 0 とすることによって，前処理の局所化を行い，並列性の高いアルゴリズムを実現している（図 4 参照)。

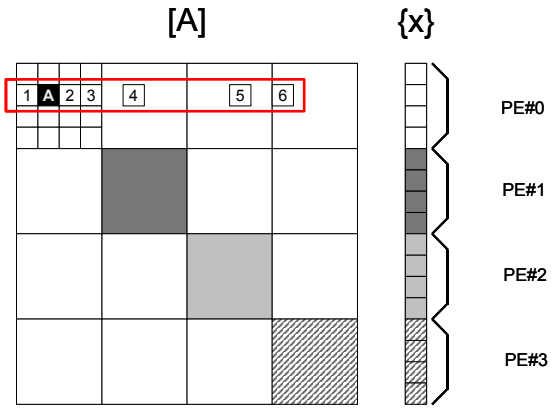


図 4. 局所 ILU（0）法における前進後退代入

たとえば PE#0 の要素「A」に注目すると 6 個の非対角成分があるが，このうち 4，5，6，は他領域に属する「外点」であるため，局所 ILU（0）法における前進後退代入においては 0 とみなされる。

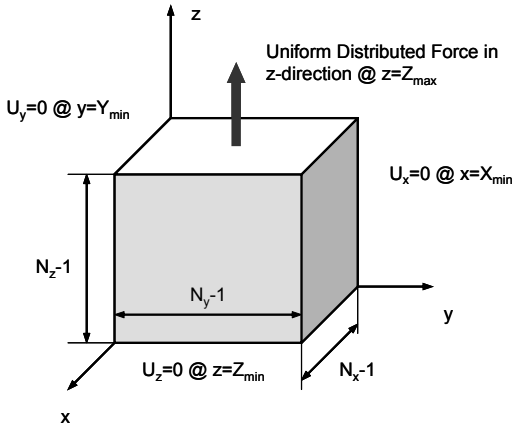
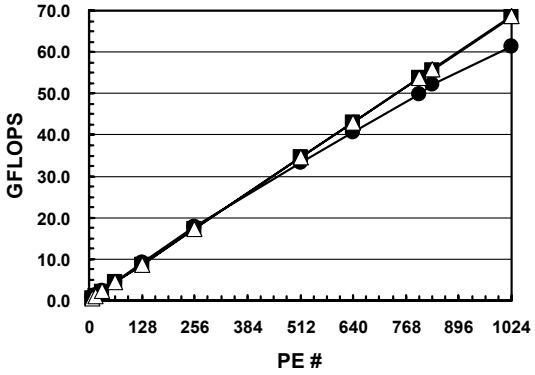


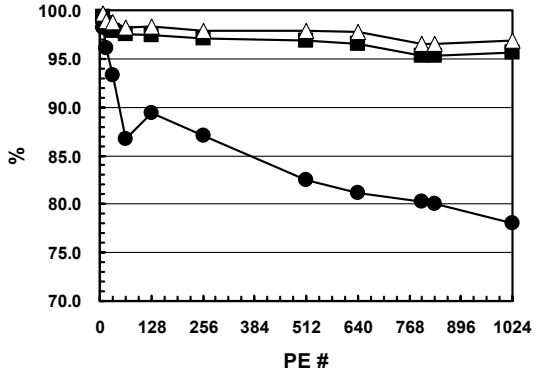
図 5. 三次元固体力学例題（弾性静解析）

図 5 に示す三次元固体力学の例題（弾性静解析，一様単純引張）を Hitachi SR2201（東京大学情報基盤センター，1024 PE，ピーク性能 300GFLOPS）を使用した計算結果を図 6 に示す。各 PE における自由度数が一定となるような問題設定となっている。この問題では一節点あたり 3 成分の自由度（x,y,z 方向の変位）が存在しているため，これらをブロック化して扱うブロック IC(0)前処理[10]を使用している。

最大 1.97×10^8 自由度の問題を 1024 PE を使用して，68.7 GFLOPS の処理性能を達成している。これはピーク性能の約 22.4 %にあたる。図 6（b）は並列化効率（(通信およびそれに要する並替計算等を除いた演算時間／全演算時間)）に関するデータである。PE 数が増加しても並列化効率 95%以上を維持しており，2. で述べたデータ構造と局所 IC(0)法によって高い並列化効率が得られている。



(a) PE数とGFLOPS値の関係：ほぼ正比例



(b) PE数と並列化効率の関係：充分大きい問題規模の場合，並列化効率は95%以上

図 5. Hitachi SR2201 を使用した GeoFEM ソルバー計算例
三次元弾性静解析：Block 局所 IC(0)CG 法
最大 1024 PE，1.97×10^8 自由度， 68.7 GFLOPS，ピーク性能の約 22.4%
各PEの自由度数を固定 ●：1.23×10^4，■：9.83×10^4，△：1.92×10^5

4. SMPクラスタ型アーキテクチュア向けソルバーの開発

「GeoFEM」の SMP クラスタ用並列ソルバーは，もともとベクトル計算機である「地球シミュレータ」を主要なターゲットとして開発されている。「地球シミュレータ」は，8 つのベクトルプロセッサから構成される SMP ノードが 640 ノード，合計 5,120 プロセッサであり，Hitachi SR8000 とは非常によく似た構成である。

「GeoFEM」の SMP クラスタ型用並列ソルバーでは以下の 3 レベルの並列性が考慮されている：

- SMP ノード間：MPI
- SMP ノード内：OpenMP
- 各プロセッサ：ベクトル化

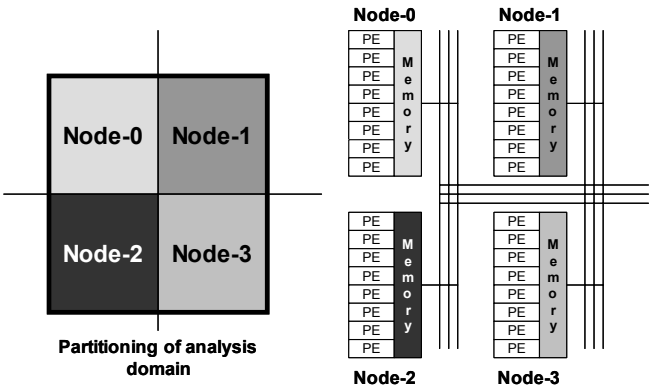


図 7. SMP クラスタ型アーキテクチュアにおける領域分割
1 領域が 1SMP ノードに対応している

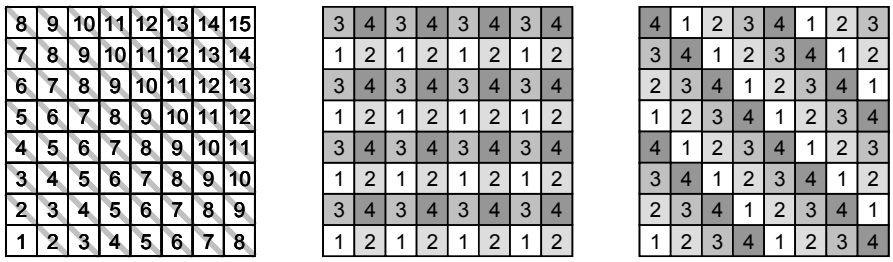
Hitachi SR8000 上で使用する場合は各プロセッサに関しては「疑似ベクトル」オプション[11]を適用している。並列計算にあたっては分割された各領域は各 SMP ノードに対応している（図 7 参照）。非構造格子を使用した計算において高いベクトル／並列性能を得るためには以下の 3 点が重要である：

- 局所的な処理

- 連続メモリアクセス
- 充分なループ長

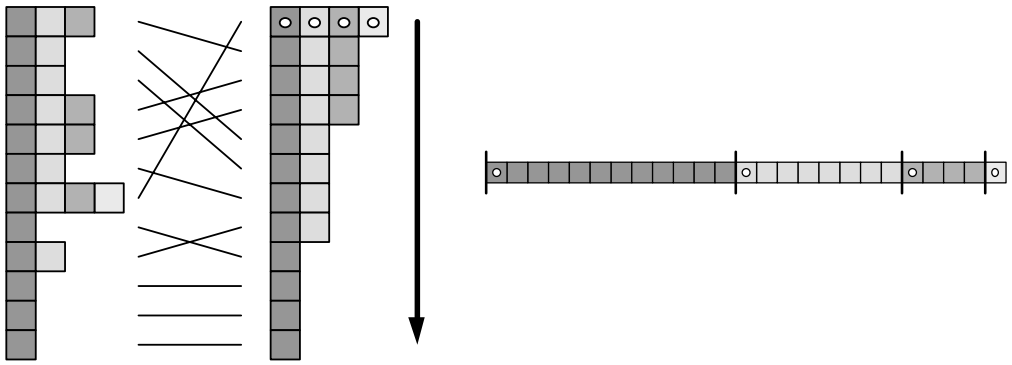
これらの要件を満たすように，鷲尾らによって考案されたベクトルプロセッサおよび SMP ノード向けのオーダリング手法[12][13]（図 8，9 参照）と GeoFEM で使用されている局所前処理手法を統合した手法を適用し，特に ILU，IC 前処理における性能を向上させている。今回使用しているオーダリング（PDJDS/CM-RCM，Parallel Descending Order Jagged Diagonal Storage/Cyclic Multicolor-RCM）の詳細については文献[14]を参照されたいが，以下の 4 段階から成っている：

① 各領域に RCM（Reverse Cuthil-McKee）オーダリング[10]を適用し，依存性のない Hyperplane を生成する（図 8（a））。
② ループ長が均等となるように CM（Cyclic Multicolor）オーダリングを適用する（図 8（c））。
③ DJDS（Descending Order Jagged Diagonal Storage）オーダリングによって各カラー内で非対角成分の多い順に並び替え，係数行列に関してメモリの連続した一次元配列を生成する（図 9）。
④ SMP ノード内での負荷が均等となるようサイクリックなオーダリングを実施する。



(a) Hyperplane/RCM　(b) Multicoloring: 4 colors　(c) CM-RCM: 4 colors

図 8. CM-RCM(Cyclic Multicolor+RCM)オーダリング



(a) 非対角成分の多い順番に並び替え　(b) 一次元圧縮配列

図 9. DJDS（Descending Order Jagged Diagonal Storage）オーダリング

上記①～④のうち，③はベクトル化向け，④は SMP ノード内並列化向けのオーダリングであり，①，

②は両者に効果的なオーダリングである。

図 5 に示した三次元弾性静解析について，最大 8.05×10^9 自由度の問題を 128 ノード（1024 プロセッサ）を使用して，335 GFLOPS の処理性能を達成している（1-SMP ノードあたり 128^3 節点，6,291,456 自由度）。これはピーク性能の約 18.6 %にあたる。また 95%以上の高い並列性能を発揮している（図 10，11 参照）。

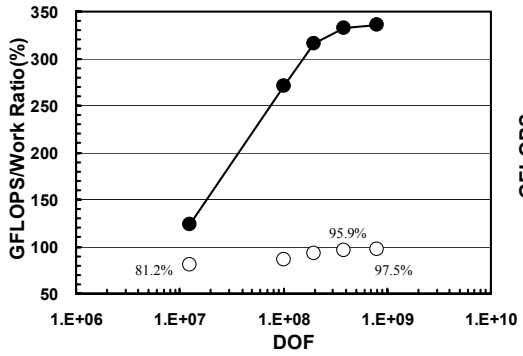


図 10. 三次元弾性静解析結果(SR8000/MPP 128 ノード使用)
●：GFLOPS 値，○：並列性能
最大 805,306,368 自由度，335.2G FLOPS
（ピーク性能 1.8 TFLOPS の約 18.6%）

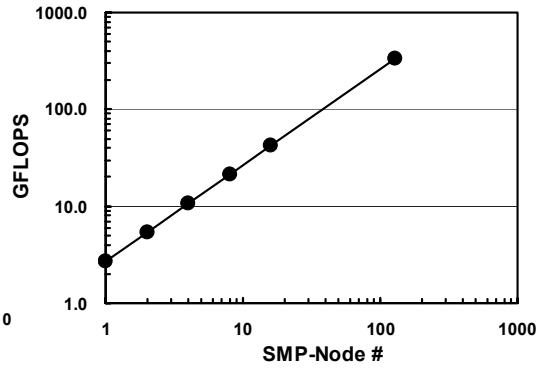


図 11. 三次元弾性静解析結果（SR8000/MPP）
SMP ノード数と GFLOPS 値の関係
1-SMP ノードあたり 128^3 節点，6,291,456 自由度
最大 805,306,368 自由度，335.2G FLOPS
（ピーク性能 1.8 TFLOPS の約 18.6%）

図 12 は 1-SMP ノードを使用して問題規模を変化させた場合（最小：16^3 節点，12,288 自由度，最大：128^3 節点，6,291,456 自由度）の結果で Hitachi FORTRAN の XCLOCK 関数[11]を使用して，SMP ノード内での通信，同期オーバーヘッドを測定し，並列化効率を算出したものである。この図によると最小問題では通信，同期オーバーヘッドは 20%以上であるが，問題規模が大きくなるにつれて減少し，40^3 節点，192,000 自由度（1 プロセッサあたり 24,000 自由度）で 10%以下，64^3 節点，786,432 自由度（1 プロセッサあたり 98,304 自由度）で 5%以下となっている。

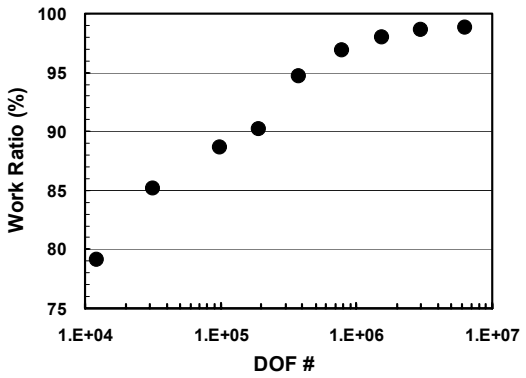


図 12. SR8000/MPP 1-SMP ノード内
並列性能（XCLOCK サブルーチン使用）
40^3 節点，192,000DOF（24,000 自由度/PE）で 90%以上

続いて，オーダリングの影響を検討するため，図 13 に示すように，①PDJDS/CM-RCM（本稿で提案している手法），②PDCRS/CM-RCM（オーダリングは①と同様だが，最内ループ長が短い），③オーダリング無しの 3 つの場合について，1-SMP ノードを使用し，問題規模を変更した計算を実施した。なお，この比較計算には東京大学情報基盤センターの Hitachi SR8000/128（ノードあたりピーク性能 8GFLOPS，SR8000/MPP は 14.4GFLOPS）を使用した。計算結果を図 14 に示す。

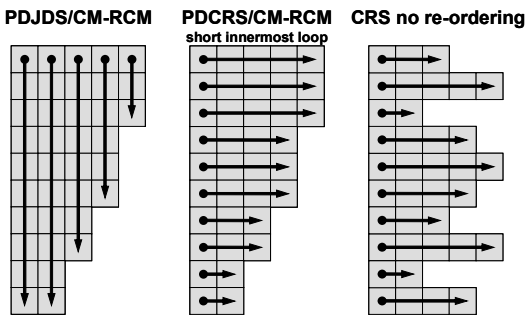


図 13. 各種のオーダリング，係数格納方法

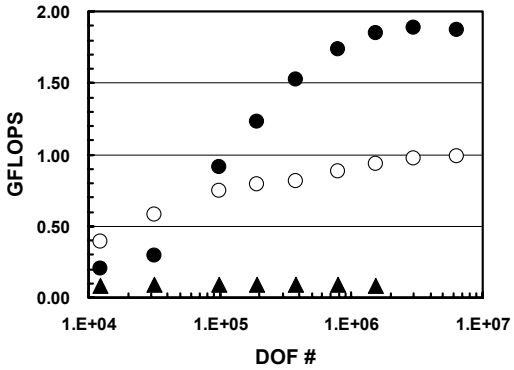


図 14. SR8000/128 1-SMP ノード内並列性能
（オーダリング，係数格納方法の効果）
●：PDJDS/CM-RCM，○：PDCRS/CM-RCM，▲：No-Ordering

図 14 に示すように，問題規模が小さい場合は最内ループ長の短い PDCRS/CM-RCM が有利であるが問題規模が大きくなるにしたがって PDJDS/CM-RCM の方が性能が良くなり，もっとも規模が大きいケース（128^3 節点，6,291,456 自由度）では 1：2 程度の性能比になっている。図 14 における PDJDS/CM-RCM の挙動は，ベクトルプロセッサの場合とよく似た傾向である。オーダリングをしない場合の性能は非常に低く，他のケースの 10 分の 1 程度である。SMP ノード内並列化，ベクトル化の効果が共に得られていないためであると考えられる。

5. 結論および将来への展望

並列有限要素法プラットフォーム GeoFEM の前処理付き並列反復法ソルバーに対して Hybrid 型並列プログラミングモデルを適用し，東京大学情報基盤センターHitachi SR8000/MPP フルノード（128SMP ノード，1024 プロセッサ）を使用して最大 805,306,368 自由度までの三次元弾性静解析を実施し，335GFLOPS（ピーク性能 1.8TFLOPS の 18.6%）の性能を得た。4 段階のオーダリングを適用する PDJDS/CM-RCM 法の効果は大きく，オーダリングを全く実施しない場合と比較して，10 倍以上の効率が得られている。本稿の手法は本来ベクトル計算機向けに開発された手法であるが，SR8000 の疑似ベクトルオプションを使用して，高いピーク性能比が得られた。

今回は単純な形状に関する線形問題の例を紹介した。「GeoFEM」の本来対象としている固体地球シミュレーションの代表的なものとして，地震発生サイクル予測のための断層接触シミュレーションがある。このようなシミュレーションは断層の複雑形状を扱う非線形問題であり，接触面の拘束条件に起因するペナルティ項のために条件数が増加する[15]。このような問題に対しては，通常のILU(0), IC(0)などの前処理手法では収束解を得ることができず，特殊な前処理手法が開発されている[15]。今後はこれらの前処理手法に関してもSMPクラスタ型アーキテクチュア向けの改良を進め，より大規模な問題にチャレンジしていく予定である。

謝 辞

本研究は文部科学省科学技術振興調整費「高精度の地球変動予測のための並列ソフトウェア開発に関する研究」の一環として実施中の「GeoFEM」プロジェクトの成果の一部である。前処理手法に関

して鷲尾 隆氏（NEC）に貴重な助言を頂いた。この場を借りて深甚なる謝意を表するものである。今回の研究紹介の機会を与えてくださった金田康正教授（東京大学情報基盤センター）に篤く御礼申し上げます。

参考文献

[1] http://www.es.jamstec.go.jp/
[2] http://www.llnl.gov/asci/
[3] http://www.mpi.org/
[4] http//www.openmp.org/
[5] Falgout, R. and Jones, J. : "Multigrid on Massively Parallel Architectures", *Sixth European Multigrid Conference*, Ghent, Belgium, September 27-30, 1999.
[6] Cappelo, F. and Etiemble, D. :"MPI versus MPI+OpenMP on the IBM SP for the NAS Benchmarks", *SC2000 Technical Paper*, Dallas, Texas, 2000.
[7] http://www.nas.nasa.gov/Research/Software/swdescription.html#NPB/
[8] http://geofem.tokyo.rist.or.jp/
[9] K.Nakajima and H.Okuda, IJCFD, Vol.12, pp.315-322, 1999.
[10] Dongarra, J.J., Duff, I.S., Sorensen, D.C. and van der Vorst, H.A. : "Numerical Linear Algebra for High-Performance Computers", SIAM, 1998.
[11] http://www.hitachi.co.jp/Prod/comp/hpc/foruser/sr8000/
[12] Washio, T., Maruyama, K., Osoda, T., Shimizu, F. and Doi, S. : "Blocking and reordering to achieve highly parallel robust ILU preconditioners", *RIKEN Symposium on Linear Algebra and its Applications*, The Institute of Physical and Chemical Research, 1999, pp.42-49.
[13] Washio, T., Maruyama, K., Osoda, T., Shimizu, F. and Doi, S. : "Efficient implementations of block sparse matrix operations on shared memory vector machines", *SNA2000 : The Fourth International Conference on Supercomputing in Nuclear Applications*, 2000.
[14] Nakajima, K. and Okuda, H. : " Parallel Iterative Solvers for Unstructured Grids using an OpenMP/MPI Hybrid Programming Model for the GeoFEM Platform on SMP Cluster Architectures", *International Workshop on OpenMP: Experiences and Implementations (WOMPEI 2002)*, Lecture Notes in Computer Science 2327 (in press), May 2002.
[15] Nakajima, K. and Okuda, H, : "Parallel Iterative Solvers with the Selective Blocking Preconditioning for Simulations of Fault-Zone Contact", GeoFEM 2001-010, RIST/Tokyo, 2001.