GeoFEM ベンチマークによる Hitachi SR11000/J2 の性能評価

中 島 研 吾

東京大学大学院理学系研究科地球惑星科学専攻

1. はじめに：Flat MPI vs. Hybrid

近年のハードウェア技術の発展によって，単一のメモリに多くのプロセッサ（Processing Element : PE，最近は「プロセッサコア」，または単に「コア」と呼ばれることも多い）が効率的にアクセスすることが可能となり，SMP（Symmetric Multiprocessors）のクラスタによる並列計算機が数多く開発されている。米国エネルギー省の ASC 計画（Advanced Simulation & Computing）[1]，「地球シミュレータ」[2] 等のテラフロップス級の超並列計算機は，すべてこのアーキテクチャによっている。

このような計算機において最大の性能を発揮するために，多段階ハイブリッド手法（multi-level hybrid，以下「Hybrid」）に基づく並列プログラミングモデルがしばしば使用されている（図 1 参照）。この手法はディレクティブによる「fine-grain parallelism」と，メッセージパッシングによる「coarse-grain parallelism」の融合であり，一般的には OpenMP[3]と MPI（Message Passing Interface）[4]を組み合わせたプログラミングスタイルである。各共有メモリユニット（SMP ノード）に OpenMP，SMP ノード間の通信に MPI が適用される。SMP クラスタ型アーキテクチャにおけるもう一つのアプローチは，個々のプロセッサを独立に扱う単段階の「Flat MPI」である（図 1）。Hybrid と Flat MPI の優劣は，さまざまなハードウェア性能諸元（コア単体性能，通信バンド幅，メモリバンド幅等）とそのバランス，アプリケーションの特性，問題サイズに依存するものである〔1〕。近年，マルチコアプロセッサの普及により，Hybrid と Flat MPI の比較については，ふたたび注目される傾向にある。

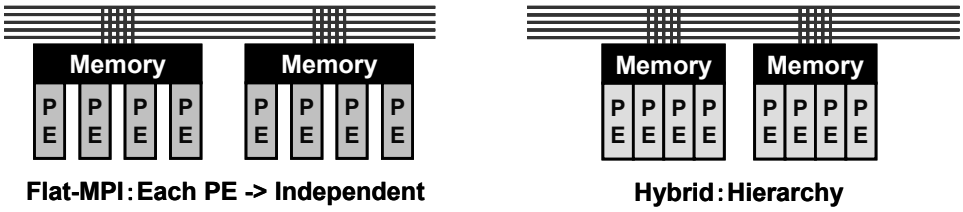


図 1　SMP クラスタにおける並列プログラミングモデル

筆者らは，固体地球シミュレーション用並列有限要素法（Finite Element Method : FEM）プラットフォーム「GeoFEM」[5]の開発の一環として，非構造格子向け前処理付き反復法による線形ソルバーを，「地球シミュレータ（ES）」上で最適化し，Hybrid および Flat MPI の両プログラミングモデルを様々なアプリケーションに適用し，検討を実施してきた〔2,3,4,5,6〕。

---

[1] https://asc.llnl.gov/
[2] http://www.es.jamstec.go.jp/
[3] http://www.openmp.org/
[4] http://www-unix.mcs.anl.gov/mpi/index.htm
[5] http://geofem.tokyo.rist.or.jp/

本学情報基盤センターに導入されている Hitachi SR11000/J2 は，IBM POWER5+プロセッサ（2.30 GHz）に基づく SMP クラスタ型の並列計算機である[6]。本研究では，並列有限要素法コード GeoFEM に基づく様々なベンチマークを，各 SMP ノードの 8 個または 16 個のコアを使って実施し，性能を評価した。

以下，第 2 章では並列有限要素法の概要について「GeoFEM」を例にとって説明する。第 3 章では GeoFEM ベンチマークの概要を紹介し，第 4 章（1 ノード）および第 5 章（複数ノード）で性能評価結果について説明し，第 6 章で本稿をまとめる。第 3 章では，文献〔7〕で紹介した，Hitachi SR11000/J1（IBM POWER5 プロセッサ，1.90 GHz）による結果との比較も実施した。

2. 並列有限要素法について

（1）概　要

有限要素法（FEM）は偏微分方程式の数値解法として，科学技術シミュレーションでは広く使用されている。近年はより詳細な計算のために，並列計算機を使用した大規模シミュレーションが盛んに実施されている。FEM の処理は，線形，非線形いずれの場合も，最終的には疎な全体剛性マトリクスを係数マトリクスとする大規模連立一次方程式を解くことに帰着される。このような方程式の係数マトリクスは，線形化した支配方程式に対して，要素単位で重みつき残差法あるいは変分法を適用して得られる要素剛性マトリクスを足し合わせて得られる。FEM の計算のほとんどの部分は：

- 係数マトリクス生成
- 線形ソルバによる大規模連立一次方程式求解

に費やされる。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，コアあたりの計算効率は差分法と比較して低いが，有限要素法では基本的に要素単位の局所的な処理が中心となるため，並列化には適している。特に，係数マトリクス生成に関しては，要素単位での実行が可能なため，並列化は容易であり，領域間の通信無しに実行することが可能である。すなわち，1CPU の PC 向けに開発されたプログラムをそのまま並列計算機上で実行することが可能である。

科学技術シミュレーションいおける連立一次方程式の解法としてはガウスの消去法などの直接法（Direct Method）が広く使用されてきたが，問題規模と計算量，必要記憶容量の非線形性のため大規模シミュレーションには適していない。大規模シミュレーション，並列計算に適した手法として共役勾配法（Conjugate Gradient Method：CG）などの Krylov 型反復法（Krylov Iterative Method）が利用されている。「GeoFEM」では Krylov 型反復法を使用している。

反復法の収束特性は係数マトリクスの固有値分布に依存するため，実用的な問題に適用するためには前処理（Preconditioning）を施し，固有値分布を変えたマトリクスを解く手法が一般的である。反復法の前処理手法としては不完全 LU 分解（Incomplete LU Factorization：ILU）あるいは対称行列向けの不完全コレスキー分解（Incomplete Cholesky Factorization：IC）などがよく使用される〔8〕。IC／ILU 前処理では，前処理行列を係数行列とする方程式を前進後退代入によって解く必要がある。IC／ILU 前処理，前進後退代入によるプロセスでは大域的な依存性が

[6] http://www.cc.u-tokyo.ac.jp/service/intro/index.html

あるため並列化，ベクトル化が困難とされてきた。筆者らは，「GeoFEM」プロジェクト以来，これらの問題の解決に取り組んできた〔2,3,4,5,6〕。ベクトル計算機，並列計算機で性能を発揮するためには：

① 局所的処理と依存性の排除
② 連続メモリアクセス
③ 充分に長い（最内）ループ長

という条件を満たしていなければならない〔2,3〕。「GeoFEM」では：

- マルチカラーあるいは Reverse Cuthill-Mckee（RCM）法によるオーダリング〔8,9〕
- ブロックヤコビ法のアイディアに基づいた局所 IC／ILU 前処理法
- 最内ループ長を大きくするための係数格納法
- GeoFEM の並列分散データ構造

等によって非常に高いベクトルおよび並列性能が得られている。これらについては既にいくつかの文献で既に紹介しているので，興味のある読者は文献〔2,3〕などを参考にされたい。また OpenMP による並列化については「スーパーコンピューティングニュース」にも連載記事[7]を執筆中であるので，本格的にプログラミングに取り組みたい場合はそちらを参照されたい。

（2）並列分散データ構造と通信

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，全体領域のデータを部分領域（局所データ）に分割する必要がある。それぞれの領域（domain，partition）は，Hybrid 並列プログラミングモデルの場合は各 SMP ノード, Flat MPI の場合は各コアに割り当てられる。並列 FEM の計算においては図 2 に示すように，領域分割された局所データを各部分領域に関して独立に読み込み，係数マトリクスを生成することが可能であり，領域間の通信が発生する可能性があるのは線形ソルバの部分のみである。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，100%に近い並列化効率を達成することも可能である。

FEM に代表される非構造格子（Unstructured Grids）を使用したアプリケーションにおいて，適切な分散データ（局所データ）構造を決定することは，並列計算を効率的に実施する上で重要である。GeoFEM の局所メッシュデータは図 3 に示すような節点ベース（node-based）の領域分割に拠っており，領域間オーバーラップ要素を含んでいる。このようなデータ構造は係数マトリクスを各部分領域で独立に生成し，更に線形ソルバとして前処理つき反復法を適用する場合に有効である〔2,3〕。

FEM において，速度，温度，変位など，線形方程式の解となるような変数は節点において定義される。したがって，並列計算における効率という観点からは領域間の節点数が均等であることが望ましい。これが節点ベースの領域分割法を採用した理由である。節点ベースの領域分

[7] 中島研吾「OpenMP によるプログラミング入門（I）～（III）」スーパーコンピューティングニュース Vol.9 No.5（2007 年 9 月号）より 3 回にわたって連載中

割を使用した場合，剛性マトリクス生成に代表されるような要素単位の処理を各領域において局所的に実施するためには，領域間のオーバーラップ要素が必要である。図 4 はこのようなオーバーラップ要素の例を示したものである。ここで，各節点の色（白，黒，灰色）は所属領域を表す。灰色に塗られた要素は複数の領域によって共有されており，各領域における各節点に対する剛性マトリクスの足し込みなどの処理を，並列に実施するためにはこれらオーバーラップ要素の情報が必要である。

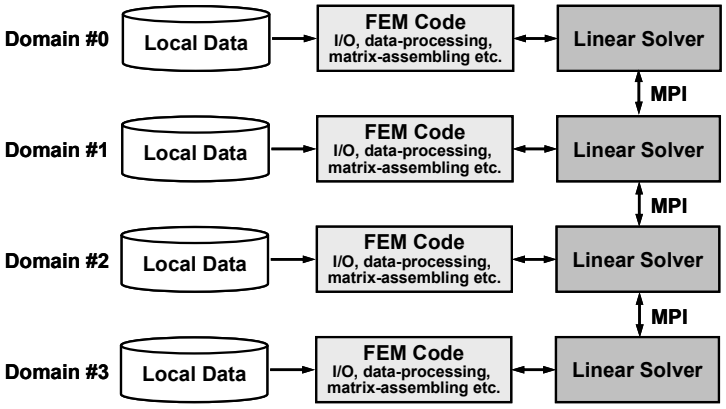


図 2　GeoFEM における分散データ処理，有限要素法処理手順（4 領域の場合）

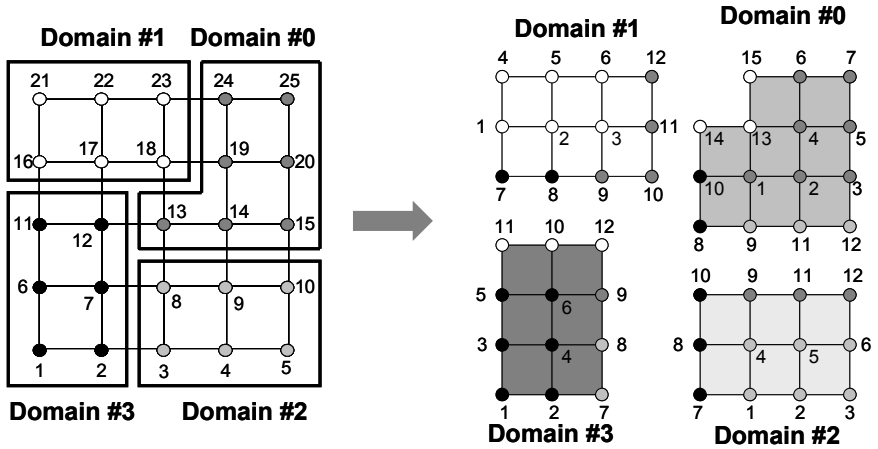


図 3　「節点ベース」領域分割の例（4 領域の場合）〔2,3〕

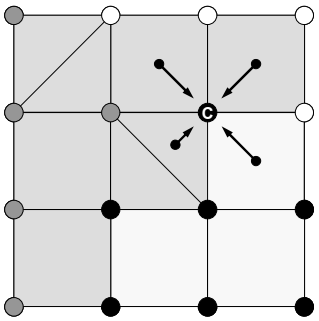


図 4　節点 C 周囲の要素単位オペレーションの概要，灰色に塗られた要素は領域間のオーバーラップ要素である（節点の色（白，黒，灰色）は所属領域を表す）

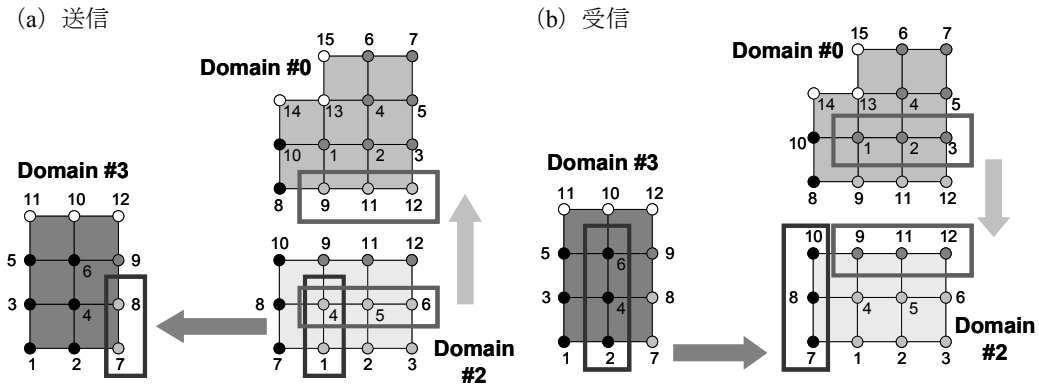


図 5　GeoFEM の局所データ構造と領域間通信

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，非構造格子では，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。

GeoFEM において，各領域は以下の情報を含んでいる：

- 各領域に割り当てられた節点
- 各領域に割り当てられた節点を含む要素
- 他の領域に割り当てられているが上記の要素に含まれている節点
- 領域間の通信テーブル（送信，受信用）
- 節点グループ，要素グループ，面グループ
- 材料物性

節点は，通信という観点から以下の 3 種類に分類される：

- 内点（Internal Nodes）：各領域に割り当てられた節点
- 外点（External Nodes）：他領域に属しているが，各領域の要素に含まれている節点
- 境界点（Boundary Nodes）：他領域の外点となっている節点

図 3 と図 5 における Domain #2 において，節点は以下のように分類される：

- 内点　　{1, 2, 3, 4, 5, 6}
- 外点　　{7, 8, 9,10,11,12}
- 境界点　{1, 4, 5, 6}

局所データには領域間の「通信テーブル」の情報も含まれる。境界点における値は隣接領域へ「送信（send)」され，送信先では外点として「受信（receive)」される。図 3，図 4 に示す局所データ構造と図 5 に示す領域間通信によって非常に高い並列性能が達成されている〔2,3,4,5,6〕。

図 6 は GeoFEM における通信処理部分のサブルーチンである。ここで， EXPORT_INDEX，EXPORT_NODE という配列が送信用の通信テーブルであり， IMPORT_INDEX，IMPORT_NODE という配列が受信用の通信テーブルである。以下のような手順で実施される。

① 従属変数ベクトル X の中身を転送（送信）ベクトル WS に代入する。neib 番目の隣接領域に対して：

```
istart= EXPORT_INDEX(neib-1) + 1
iend  = EXPORT_INDEX(neib)
```

とすると WS(istart)から WS(iend)の連続したデータが neib 番目の部分領域に転送される：

```
do neib= 1, NEIBPETOT
  istart= EXPORT_INDEX(neib-1)
  inum  = EXPORT_INDEX(neib  ) - istart
  do k= istart+1, istart+inum
    WS(k)= X(EXPORT_NODE(k))
  enddo
  call MPI_ISEND (WS(istart+1), inum, etc.)
enddo
```

② ①とは逆の手順で転送（受信）ベクトル WR を受け取る。neib 番目の隣接領域に対して：

```
istart= IMPORT_INDEX(neib-1) + 1
iend  = IMPORT_INDEX(neib)
```

とすると，WR(istart)から WR(iend)までの連続したデータが neib 番目の部分領域から転送される。

```
do neib= 1, NEIBPETOT
  istart= IMPORT_INDEX(neib-1)
  inum  = IMPORT_INDEX(neib  ) - istart
  call MPI_IRECV (WR(istart+1), inum, etc.)
enddo
```

③ 従属変数ベクトル X に転送（受信）ベクトル WR の中身を代入する。

```
do neib= 1, NEIBPETOT
  istart= IMPORT_INDEX(neib-1)
  inum  = IMPORT_INDEX(neib  ) - istart
  do k= istart+1, istart+inum
    X(IMPORT_NODE(k))= WR(k)
  enddo
enddo
```

図 6 に示されるような通信は， Krylov 反復解法の行列ベクトル積（matrix-vector product）の部分で発生する。

(a) 領域間通信サブルーチンの呼び出し（スカラー型，3×3 ブロック型）

1x1 Scalar

```
      allocate (WS(NP), WR(NP), X(NP)
      call SOLVER_SEND_RECV                                          &
     &   ( NP, NEIBPETOT, NEIBPE, IMPORT_INDEX, IMPORT_NODE,         &
     &     EXPORT_INDEX, EXPORT_NODE, WS, WR, X , SOLVER_COMM,       &
     &     my_rank)
```

3x3 Block

```
      allocate (WS(3*NP), WR(3*NP), X(3*NP)
      call SOLVER_SEND_RECV_3                                        &
     &   ( NP, NEIBPETOT, NEIBPE, IMPORT_INDEX, IMPORT_NODE,         &
     &     EXPORT_INDEX, EXPORT_NODE, WS, WR, X , SOLVER_COMM,       &
     &     my_rank)
```

(b) 領域間通信サブルーチンの概要

- 送信フェーズ

```
do neib= 1, NEIBPETOT
  istart= EXPORT_INDEX(neib-1)
  inum  = EXPORT_INDEX(neib  ) - istart
  do k= istart+1, istart+inum
    WS(k)= X(EXPORT_NODE(k))
  enddo
  call MPI_ISEND
       (WS(istart+1), inum, MPI_DOUBLE_PRECISION, &
        NEIBPE(neib), 0, SOLVER_COMM,             &
        req1(neib), ierr)
enddo
```

- 受信フェーズ

```
do neib= 1, NEIBPETOT
  istart= IMPORT_INDEX(neib-1)
  inum  = IMPORT_INDEX(neib  ) - istart
  call MPI_IRECV
       (WR(istart+1), inum, MPI_DOUBLE_PRECISION, &
        NEIBPE(neib), 0, SOLVER_COMM,             &
        req2(neib), ierr)
enddo

call MPI_WAITALL (NEIBPETOT, req2, sta2, ierr)

do neib= 1, NEIBPETOT
  istart= IMPORT_INDEX(neib-1)
  inum  = IMPORT_INDEX(neib  ) - istart
  do k= istart+1, istart+inum
    X(IMPORT_NODE(k))= WR(k)
  enddo
enddo

call MPI_WAITALL (NEIBPETOT, req1, sta1, ierr)
```

図 6　GeoFEM における領域間通信プロセス〔2,3〕

GeoFEM では初期全体メッシュデータから局所分散メッシュデータを自動的に生成するためのツールとして領域分割ツール（Partitioner）が用意されている。利用者は実際には上記の通信テーブルについては意識することなく並列有限要素法コードの開発，利用が可能である。領域分割にあたっては：

- 各領域の負荷が均等となっていること
- 領域間の通信が少ないこと

が重要である。特に前処理付き反復法を使用する場合には収束を速めるために後者が重要なポイントである〔2,3,4,5,6〕。この両条件を満たす手法としては METIS[8] が良く知られている。GeoFEM の領域分割ツールでは文献〔10〕で紹介されている RCB 法（Recursive Coordinate Bisection）等のほか，METIS に関するインターフェースも提供している。

3. GeoFEM ベンチマーク

（1）概　要

本研究では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群〔7〕を使用した。

GeoFEM ベンチマークは，①三次元弾性問題（Cube 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性能（GFLOPS 値）を様々な条件下で計測するものである。プログラムは全て OpenMP ディレクティヴを含む 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（図 3～図 5）を使用しており，マルチカラー法に基づくオーダリング手法によりベクトルプロセッサ，SMP 並列計算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し，SMP クラスタの性能評価に適している。①～③の各ベンチマークは 8 コアから成る SMP ノードの性能評価のためのものであるが，①の Cube モデルは任意の問題サイズで任意のコア数を使用したベンチマークテストが可能である。GeoFEM ベンチマークに基づいた，三次元非定常伝熱解析プログラム「GAPgeofem」は「SPEC MPI 2007」ベンチマーク[9]の一つとして採用されている。

様々なハードウェアに対応可能なように，連立一次方程式の係数マトリクスの格納法として図 7 に示す 2 種類の方法が準備されている。ベクトルプロセッサ向けには，長いループ長が得られるように図 7（a）に示す Descending order Jagged Diagonal Storage（DJDS）法を採用している。スカラープロセッサ向けには非対角成分の走査方向を変えた Descending order Compressed Row Storage（DCRS）（図 7（b））を利用可能である。DCRS では最内ループ長が短くなるが，最内ループにおけるデータの局所性を保つことが可能であり，キャッシュの有効利用に適している〔4,6〕。

以下に各ベンチマーク問題について説明し,「地球シミュレータ（ベクトルプロセッサ）（ES）」および「IBM SP3（スカラープロセッサ）（SP3）（米国国立ローレンスバークレイ研究所）」[10]の 1 ノード，8 コアを使用した場合の計算結果〔2,3,4,5,6〕について紹介する。表 1 は「地球シミュレータ」，「IBM SP3」および 4 章以降で扱う「Hitachi SR11000/J2」のハードウェア諸元について示したものである。「Hitachi SR11000/J2」の MPI latency については公表された値が

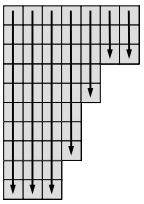
[8] http://glaros.dtc.umn.edu/gkhome/views/metis/
[9] http://www.spec.org/auto/mpi2007/Docs/128.GAPgeofem.html
[10] http://www.nersc.gov/

ないため，文献〔11〕に示されている，ほぼ同じ性能と推定される IBM p5-575（米国国立ローレンスバークレイ研究所）の値を，参考値として記載している。メモリバンド幅は STREAM ベンチマーク[11]による実測値である。

「地球シミュレータ」は NEC SX-6 に基づく並列ベクトル計算機であり，640 の SMP ノード，5,120 のベクトルプロセッサと 10 TB のメモリから構成されている。ピーク性能は 40 TFLOPS である。各 SMP ノードは 8 個のベクトルプロセッサ，16GB のメモリから構成されており，各 PE のピーク性能は 8 GFLOPS である。各 SMP ノードは単段クロスバーにより接続されており，双方向の転送速度は 12.3 GB/sec.である。

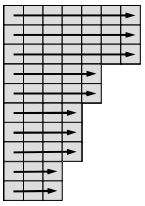
IBM SP-3 は IBM POWER3 に基づいたスカラーシステムであり，380 の SMP ノード，6,080 の PE，7.3TB のメモリから構成されており，ピーク性能は 9.12 TFLOPS である。各 SMP ノードは 16 個の PE，16～64GB のメモリから構成されており，各 PE のピーク性能は 1.50 GFLOPS である。各 PE は 64KB の L1 キャッシュと 8MB の L2 キャッシュをそれぞれ独立に持っている。各 SMP ノードは双方向の転送速度が 2.00 GB/sec.のスイッチにより接続されている。本研究では，ES との比較のため，各 SMP ノード 8 PE を使用した。Hitachi SR11000/J2 については次章で説明する。

(a) DJDS
(Descending-order
Jagged Diagonal Storage)



```
do j= 1, NJmax
  do i= 1, Imax(j)
    k=(j-1)*N+i; kk=IA(k)
    Y(i)= Y(i)+A(k)*X(kk)
    …
  enddo
enddo
```

(b) DCRS
(Descending-order
Compressed Row Storage)



```
do i= 1, N
  do k= IND(i-1)+1, IND(i)
    kk=IA(k)
    Y(i)= Y(i)+A(k)*X(kk)
    …
  enddo
enddo
```

図 7　係数マトリクスの格納方法とループ構造

表 1　地球シミュレータ，IBM SP3，Hitachi SR11000/J2 のハードウェア諸元〔4,6,11〕

	Earth Simulator	IBM SP3	Hitachi SR11000/J2
Core#/node	8	16	16
Architecture of each Core	NEC SX6	IBM POWER3	IBM POWER5+
Clock Tate (MHz)	500	375	2300
Peak Performance/core (GFLOPS)	8.00	1.50	9.20
Memory/node (GB)	16	64	128
Measured Memory BW (GB/sec/core)	26.6	0.623	6.40
Network BW (GB/sec/node)	12.3	2.00	12.0
MPI Latency (μsec)	5.6-7.7	16.3	4.7〔11〕

[11] http://www.streambench.org/

（2）三次元弾性問題

三次元弾性問題の対象は，単純形状（Cube）モデル（図 8）と図 9 に示すような PC のマイクロプロセッサの Pin Grid Array（PGA）を模擬したモデルである。いずれも，三次元弾性問題を局所不完全コレスキー分解付き共役勾配法（局所 ICCG 法）により解く。Cube モデルは任意の問題サイズ，領域数でのベンチマークを実施可能である。PGA モデルは問題規模が固定されており（1,012,354 節点，3,037,062 自由度（Degrees of Freedom : DOF）），マルチカラーの色数の効果，OpenMP と MPI（8 領域）の比較検討に使用される。

図 10 は，Cube 型モデルについて，1 ノード（8 コア）において，色数＝100 または最内ループ長＞256 とした場合の，様々な問題サイズ（自由度数 : degrees of freedom，以下 DOF）における計算結果（GFLOPS 値）である。ES では，ベクトル機の特徴として 3.81 GFLOPS（ピークの 6%）から 22.7GFLOPS（ピークの 35.5%）まで，問題サイズとともに増加する。また，ループ長を長くとれる DJDS（●○）が DCRS（■□）よりも性能が高い。MPI（●■）と OpenMP（○□）の差はほとんど無いが，MPI が若干速い。SP3 では L2 キャッシュの効果により，問題サイズが小さい場合に性能が高い。DJDS と DCRS，MPI と OpenMP の違いは少ないが，特に問題サイズが小さい場合は，各 PE に独立に装着された L2 キャッシュを有効利用できる手法（DCRS，MPI）の性能が高い。表 2 は，ハードウェアの諸元（ピーク性能，メモリバンド幅）を元に，1 コアあたりの性能を予測し，図 10 に示した測定値と比較したものである。MPI（8 コア）を使用して，3,000,000（$=3\times100^3$）DOF の問題を色数 100 で計算した場合の性能を 8 で割ったものである。性能は文献〔12〕に示した推定法に基づき，表 1 に示した各プロセッサのピーク性能とメモリバンド幅（STREAM ベンチマークによる実測値）によって推定した。SP3 においてキャッシュの効果は考慮していない。ES では推定値と実測値は非常によく一致している。SP3 のようなスカラープロセッサではキャッシュの効果を考慮していないこともあり，実測値は概して予測値を上回っている。

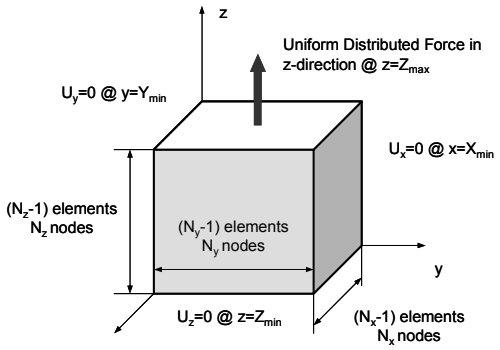


図 8　Cube モデルの概要，境界条件

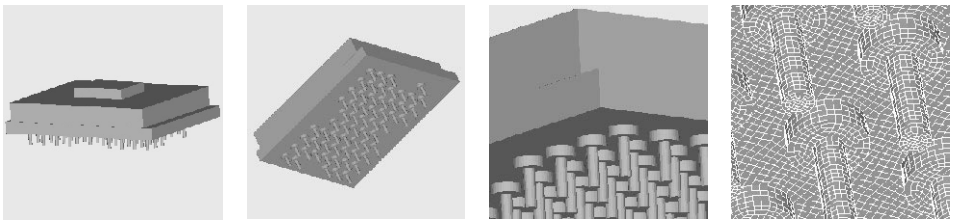
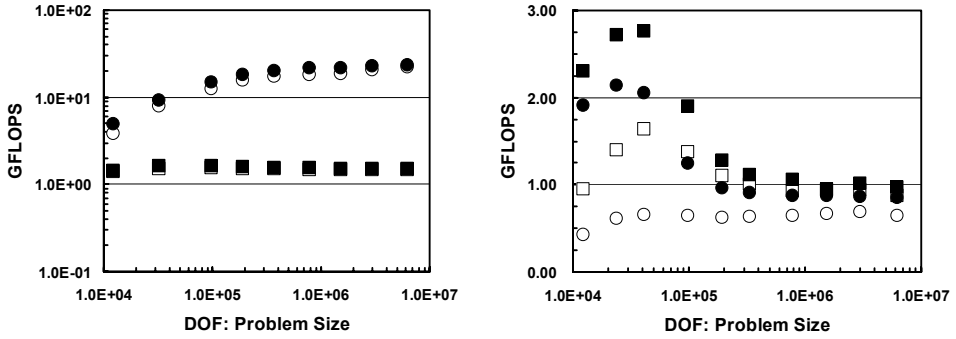


図 9　PGA モデルの概要（956,128 要素，1,012,354 節点，3,037,062 自由度）



（a）Earth Simulator　　（b）IBM SP3

図 10　Cube モデルの計算結果（問題規模と性能の関係），並列プログラミングモデル，行列格納方法の影響（●：MPI/DJDS，○：OpenMP/DJDS，■：MPI/DCRS，□：OpenMP/DCRS）

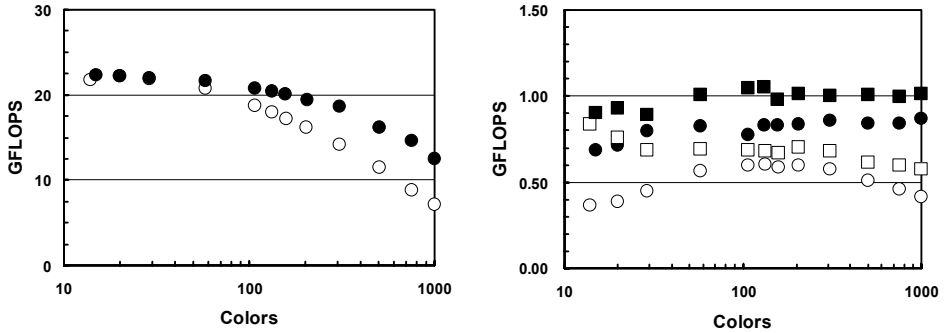
表 2　地球シミュレータ，IBM SP-3 のハードウェア諸元と Cube モデルの性能〔4,6〕

	Earth Simulator	IBM SP-3
Peak Performance/core (GFLOPS)	8.00	1.50
Measured Memory BW (GB/sec/core)	26.6	0.623
Estimated Performance (GFLOPS (% of peak))	2.31-3.24 (28.8-40.5)	0.072-0.076 (4.80-5.05)
Measured Performance (GFLOPS (% of peak))	2.93 (36.6) (DJDS)	0.122 (8.11) (DCRS)

図 11 は PGA モデルの実行例（色数の効果）である。マルチカラーオーダリングに基づく反復解法では，色数を増加させることによって反復回数は減少する〔2,3,4,5,6〕。しかしながら，各色内での要素数が減少するため，ループ長が短くなり，ベクトルプロセッサにおける性能（GFLOPS 値）は低下することが知られている〔2,4,6〕。ES では MPI，OpenMP のいずれも色数が増加すると GFLOPS 値は低下している。この傾向は OpenMP において特に顕著である。これは主として，不完全コレスキー分解（IC）による前処理の前進後退代入処理における OpenMP のオーバーヘッドによるものと考えられる（図 12）。

SP3 ではキャッシュを有効利用できる手法（DCRS，MPI）の性能が高い（■＞●＞□＞○）。スカラープロセッサにおいては色数の性能に対する影響は顕著ではないが，DJDS（●○）を採用すると，図 11 に示すように色数が増加するほど性能は向上する。これは，図 7（a）における最内ループが色数増加とともに短くなり，データの局所性が増大し，キャッシュがより有効に利用されるためと考えられる。この結果からベクトル計算機向けに開発されたマルチカラーオーダリングを使用したコードは，色数を多くすることにより，スカラープロセッサでも高い性能を達成することが可能になる。

OpenMP の場合（○□）は DJDS，DCRS に関わらず色数が増加すると，図 8 に示した前進後退代入処理における OpenMP のオーバーヘッドにより性能が低下する。OpenMP/DJDS（○）の場合は 10 色～100 色程度までは色数の増加によって性能が向上する傾向が見られるが，色数が 100 から 1000 まで増加すると，OpenMP のオーバーヘッドによる性能の低下が見られる。



（a）Earth Simulator　　（b）IBM SP3

図 11　PGA モデルの計算結果（色数と性能の関係），並列プログラミングモデル，行列格納方法の影響（●：MPI/DJDS，○：OpenMP/DJDS，■：MPI/DCRS，□：OpenMP/DCRS）

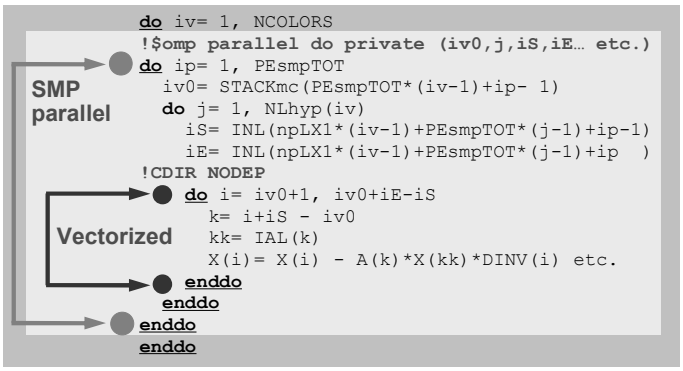


図 12　OpenMP による IC 前処理における前進後退代入の並列化例〔2,3,4,5〕

（3）三次元接触問題

プレート境界の断層接触面（図 13 参照）における応力蓄積と地震発生サイクルのシミュレーションは，地震発生メカニズムに関する知見を得るために重要なアプリケーションであり，「GeoFEM」のターゲットアプリケーションの一つである〔2,3〕。「選択的ブロッキング（Selective Blocking：SB）」前処理はこのような問題を効率良く計算するために著者によって開発された手法である〔2〕。ペナルティ数を導入することによって断層周辺の拘束条件を表現している場合に特に有効である。

三次元固体力学においては 1 節点に 3 方向の変位成分が自由度として存在するため，これらの 3 自由度をブロック化して取り扱っている。IC／ILU 型前処理では，この 3×3 行列に完全 LU 分解を適用することによって，より安定な収束性を得ている。選択的ブロッキングでは，「接触グループ」に属する節点群を並べ替えによって，1 つのグループとして扱い，このグループ（選択的ブロック）における（3×NB）×（3×NB）行列（但し NB は選択的ブロック内の節点数）に対して完全 LU 分解を適用するものである（図 14 参照）。対象とする行列が対称正定の場合には，選択的ブロッキング（SB）を，Fill-in なしのブロック ICCG 法と組み合わせた SB-BIC(0)-CG 法が非常に有効であり，広範囲のペナルティ数に関して安定である〔2〕。図 13 に示す西南日本領域を対象とした固定サイズのモデル（784,000 要素，2,471,439 DOF）を解き，マルチカラーの色数の効果，OpenMP と MPI（8 領域）の比較検討を実施する。マトリク

ス格納法は DJDS のみである。図 15 に示すように，色数と性能の関係については三次元弾性解析（PGA モデル）と同様である。

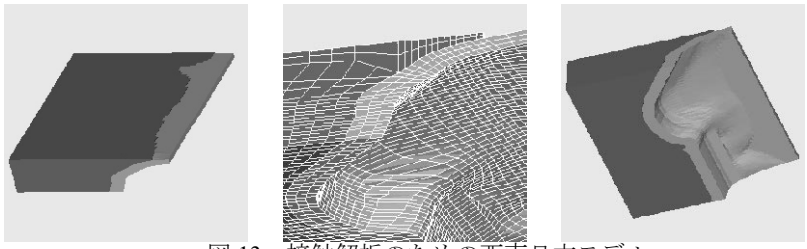


図 13　接触解析のための西南日本モデル

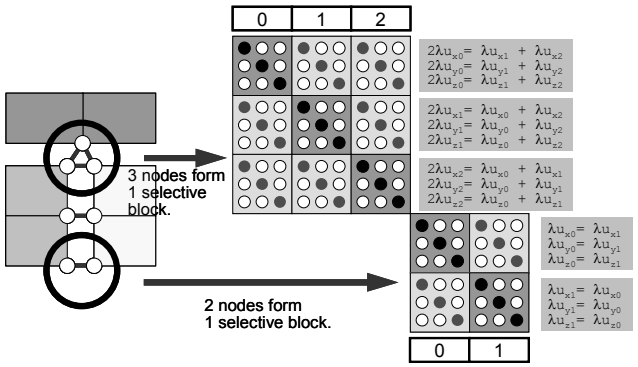
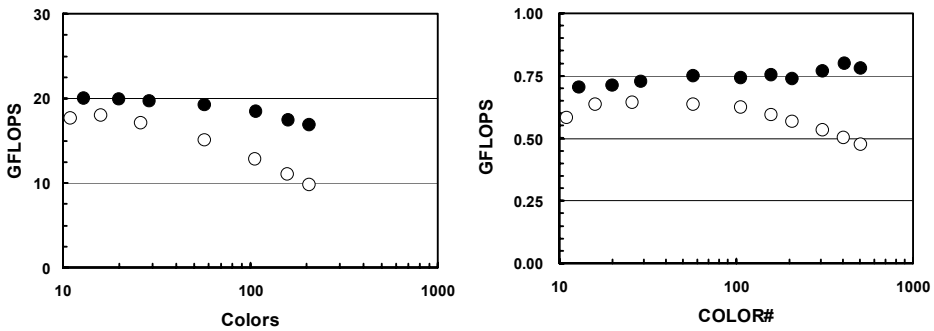


図 14　選択的ブロッキング（selective blocking）前処理の概要：同じ「接触グループ」に属する節点群をブロック化して解く



（a）Earth Simulator　　（b）IBM SP3

図 15　接触問題（西南日本モデル）の計算結果（色数と性能の関係），並列プログラミングモデルの影響（●：MPI/DJDS，○：OpenMP/DJDS）

（4）二重球殻間領域三次元ポアソン方程式

マントル対流，海洋大循環モデルなどで使用される，二重球殻間の領域における非圧縮性流体のシミュレーションにおいて得られるポアソン方程式を，Gauss-Seidel 法を緩和演算子とするマルチグリッド前処理付き CG 法（MGCG）で解く。空間は図 16 に示すように正二十面体を分割して得られる三角形を底面とする三角柱メッシュによって離散化されており，メッシュ分割のための階層構造をマルチグリッドに使用する。問題規模は，6,144,000 要素に固定されてお

り，マルチカラーの色数の効果，OpenMP と MPI（8 領域）の比較検討を実施する。マトリクス格納法は DJDS のみである。

図 17 は ES, SP3 を使用した場合の色数と GFLOPS 値の関係である〔4.6〕。色数を増やすと，MPI ではほとんど計算性能の変化は無いが，OpenMP では色数の増加とともに計算時間が増加する。色数が 12 色と 2000 色の場合を比較すると計算性能の比は 6.02（SR8000/G1），3.90（SP3），となる。これは（2），図 12 で述べた前進後退代入処理における OpenMP の同期オーバーヘッドによるものと考えられる。マルチグリッド法では，Gauss-Seidel 法による緩和計算において図 12 に示すのと同様な前進後退代入処理が発生するが，粗い格子上では計算量そのものが減るため，同期オーバーヘッドの影響を受けやすくなる。したがって，マルチカラーオーダリングによる ICCG 法と比較して，色数によるオーバーヘッドの増加はより顕著である。

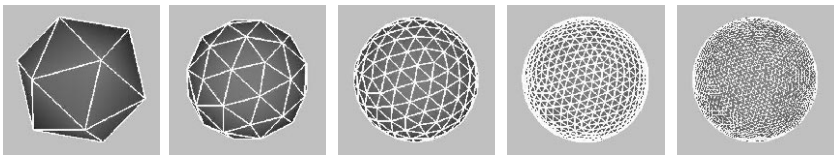
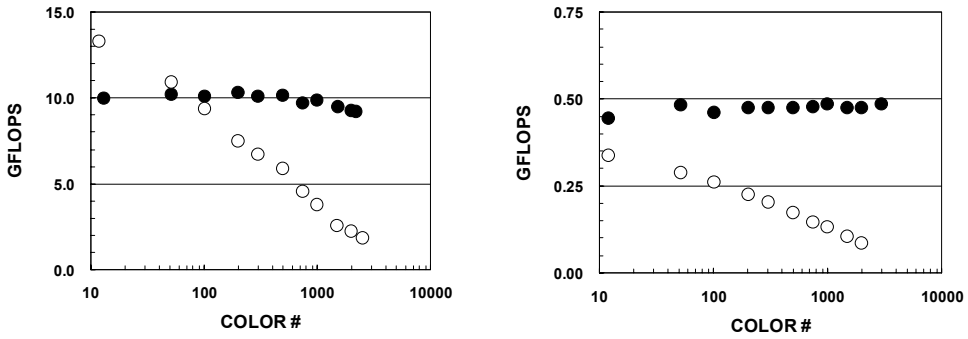


図 16　正二十面体の分割によって生成した球面上の三角形メッシュ



（a）Earth Simulator　　（b）IBM SP3

図 17　二重球殻モデル（多重格子法）の計算結果（色数と性能の関係），並列プログラミングモデルの影響（●：MPI/DJDS，○：OpenMP/DJDS）

（5）三次元弾性問題（複数ノード）

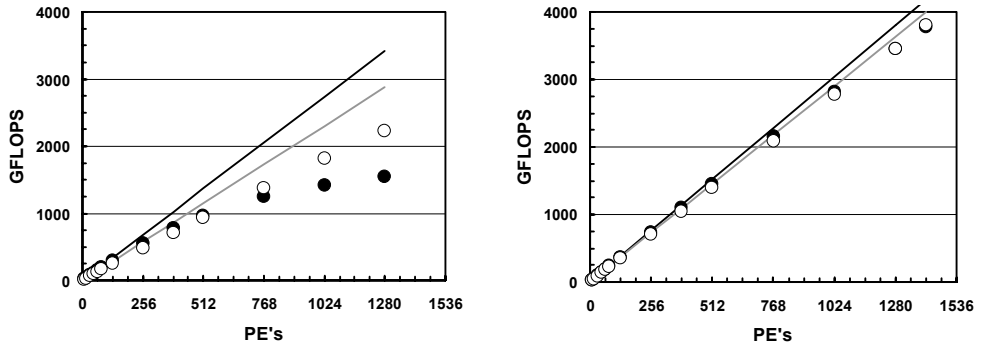
最後に図 8 に示した Cube モデルにおける三次元弾性解析を 100 SMP ノード以上を使用して計算した場合の結果について示す。Hybrid と Flat MPI それぞれの性能を評価した。PE（または SMP ノード）あたりの問題規模を固定し，ノード数を変化させた場合の計算（Weak Scaling）を実施した。ES の場合は 1 ノード（8 PE）～176 ノード（1,408 PE），IBM SP-3 の場合は 1 ノード（8 PE）～128 ノード（1,024 PE）を使用した。

ES では DJDS を使用した場合，最大問題サイズは 2.21×10^9 DOF，最高性能は 3.80 TFLOPS であり，これは ES の 176 ノード（1,408 PE）のピーク性能（10.24 TFLOPS）の 33.7%に相当する（図 18）。直線は，1 ノード（8 PE）における性能を基準とした，Flat MPI，Hybrid の場合の理想値である。Hybrid と Flat MPI はほぼ同じ性能であるが，SMP ノード数が増加し，かつ各

PE あたりの問題規模が比較的小さいとき，Hybrid の性能が卓越する（図 18）。これは文献〔13〕で紹介されているように，ES の MPI latency の値が比較的高いため，MPI プロセス数が増加すると通信の遅延効果が大きくなるためと予想される。Flat MPI は Hybrid と比較して MPI プロセスの数が 8 倍となるため，SMP ノード数が大きく，PE 当たりの問題規模が比較的小さいときはこの効果は顕著になる。

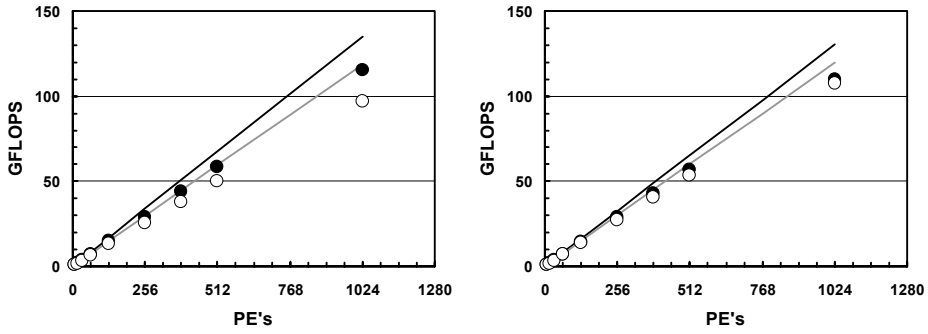
図 19 は IBM SP3（DCRS）による結果である。最大問題サイズは 3.84×10^8 DOF，最高性能は 110 GFLOPS であり，これは 128 ノード（1,024 PE）のピーク性能の 7.16 %に相当する。高い並列性能が得られているが，ES の場合に見られたような，SMP ノード数が増加した場合の Flat MPI の性能低下は観察されなかった。128 ノードにおける結果は 1 ノードの性能の外挿に近い値である。ES，SP3 いずれの場合も，PE あたりの問題規模が大きい場合は，通信のオーバーヘッドの影響が少なく，SMP ノード数が増えても理想値に近い性能が得られている。

図 20 は，更にわかりやすいように，ES および SP3 について，PE あたりの問題規模が大きいケースについて，1 ノード（8 PE）からの計算性能増加率（Speed-Up）として表した図である。Hybrid, Flat MPI ともにほぼ理想値に近い性能の増加傾向を見せているが, ES, SP3 ともに Hybrid の方が若干増加率が高い。



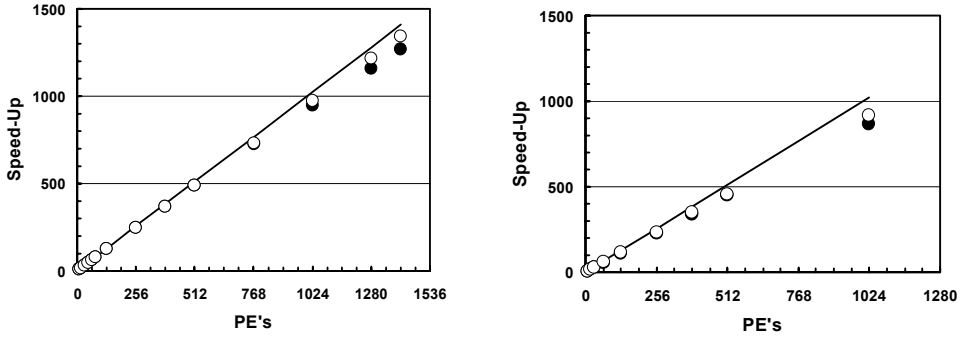
（a）3×32^3=98,304 DOF/PE　　（b）3×64×64×128=1,572,864 DOF/PE

図 18　「地球シミュレータ（ES）」における Cube モデルの計算結果（Weak Scaling）（●：Flat MPI/DCRS，○：Hybrid/DCRS，ー：Flat MPI/DCRS（ideal），—：Hybrid/DCRS （ideal））



（a）3×32^3=98,304 DOF/PE　　（b）3×50^3=375,000 DOF/PE

図 19　「IBM SP3」における Cube モデルの計算結果（Weak Scaling）（●：Flat MPI/DCRS，○：Hybrid/DCRS，ー：Flat MPI/DCRS（ideal），—：Hybrid/DCRS （ideal））



（a）ES：3×64×64×128=1,572,864 DOF/PE　　（b）IBM SP3：3×50^3=375,000 DOF/PE

図 20　「地球シミュレータ」，「IBM SP3」における Cube モデルの計算結果，1 ノードからの計算性能増加率（Weak Scaling）（●：Flat MPI/DCRS，○：Hybrid/DCRS，－：ideal）

表 2 からもわかるように，Cube モデルの計算を ES で実行した場合 IBM SP3 と比較して約 25 倍程度の GFLOPS 値が得られている。しかしながら，表 1 からも明らかなように，MPI latency の値はそれほど大きく変わらない。

三次元固体力学に対して反復法を適用した場合，行列ベクトル積の計算（Matrix-Vector Multiplication：mat-vec）においては以下に示すように，1 つの非対角成分に対して 18 回の浮動小数点演算が必要になる〔12〕。

```
do j= 1, NLU
do i= 1, N
  k= ITEM(i,j)
  Y(3*i-2)= Y(3*i-2) + AMAT(1,i,j)*X(3*k-2) +                    &
                       AMAT(2,i,j)*X(3*k-1) + AMAT(3,i,j)*X(3*k)
  Y(3*i-1)= Y(3*i-1) + AMAT(4,i,j)*X(3*k-2) +                    &
                       AMAT(5,i,j)*X(3*k-1) + AMAT(6,i,j)*X(3*k)
  Y(3*i  )= Y(3*i  ) + AMAT(7,i,j)*X(3*k-2) +                    &
                       AMAT(8,i,j)*X(3*k-1) + AMAT(9,i,j)*X(3*k)
enddo
enddo
```

非対角成分の数を 30 とすると，一回の mat-vec の計算における浮動小数点演算の回数は，FEM における節点数を N とすると 540N 程度となる。例えば 1 PE において $N=32^3$ の場合（= 98,304 DOF/PE= 786,432 DOF/SMP-node，図 18（a），図 19（b）の小規模問題ケースに相当する），ES の性能が 2.80 GFLOPS（ピークの 35%）とすると，計算時間は約 6 msec である。並列有限要素法では mat-vec を一回実行するたびに，領域境界のデータ交換が必要になり，通信が発生する（図 5，図 6 参照）〔2〕。通信時間は 50 μsec のオーダーであり，ほとんど無視できる。また，また，図 6 に示すように，領域間のデータ交換には，送信・受信バッファへのコピーが必要となるが，ES では表 1 に示すように非常に高いメモリバンド幅を実現しているので，これもほとんど無視できる。表 1 によると ES の MPI latency が 7 μsec 程度であるので，もし 1,000 PE 以上を使用した場合，計算時間と遅延時間が同じオーダーになり，影響は深刻となる可能性がある。

実際，図 21 に示すように，ES では，1 反復あたりの通信オーバーヘッド（1 ノード＝8PE の計算時間を基準として算出）は問題サイズ（すなわち通信量）に依存せず，特に SMP ノード数

が増加した場合は，Flat MPI と Hybrid の差のみが顕著に現れており，MPI latency の影響が大きいであることがわかる〔6〕。IBM SP3 の場合はこれほど顕著では無いが，図 20 に示すように，問題規模が大きい場合でも，SMP ノード数が増加した場合には，Hybrid が Flat MPI よりも性能増加率が高くなる傾向にある。

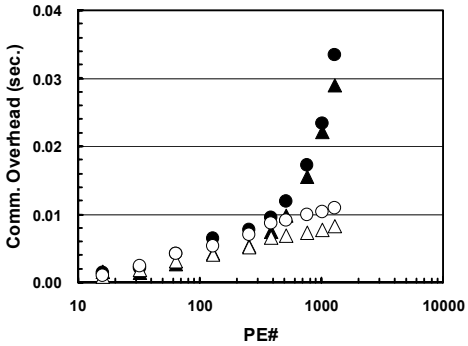


図 21　「地球シミュレータ」における 1 反復あたりの通信オーバーヘッド（1 ノードの計算時間との差から算出，DJDS）（●：Flat MPI: 3×50^3=375,000 DOF/PE，○：Hybrid: 3×50^3=375,000 DOF/PE，▲：Flat MPI: 3×32^3=98,304 DOF/PE，△：Hybrid: 3×32^3=98,304 DOF/PE）

4. Hitachi 11000 における実行例（1 ノード）

（1）概要

Hitachi SR11000/J2（SR11000/J2）[12]の 1 ノードを使用して，GeoFEM ベンチマークを実施した。2007 年 3 月以前は, 本学情報基盤センターにおいては, IBM POWER5 プロセッサ（1.90 GHz）に基づく Hitachi SR11000/J1（SR11000/J1）が導入されていた。Hitachi SR11000 は 1 ノード 16 コアから構成される SMP クラスタ型アーキテクチュアを採用しているが〔14〕，SR11000/J1 ではこの 1 ノードを 8 コアずつから構成される SMP ノード 2 つに更に分割して, 使用した〔7〕。比較のため，ここでは SR11000/J1，SR11000/J2 ともに，1 ノードあたり 8 コアを使用した場合についての検討を中心に実施する。プログラムのコンパイルにあたっては，推奨オプションである「-Oss -64 -looptiling (-noparallel または -omp)」を適用した。

Hitachi SR11000/J2 のハードウェア諸元は表 1 に既に示した通りである。2 つの IBM POWER5+ コア（2.3 GHz，ピーク性能：9.2 GFLOPS）によって，POWER5+チップが構成されている。各コアは 32KB の L1 キャッシュを持ち，L2 キャッシュ（1.875 MB），L3 キャッシュ（36 MB）は各チップ内で 2 つのコアに共有されている。チップ内にはメモリコントローラが内蔵されており，高速で信頼性の高いメモリへのアクセスが可能である。4 つのチップ，すなわち 8 つのコアからモジュール（Multi Chip Mudule：MCM）が構成され，2 つの MCM，すなわち 16 個のコアが 1 つの SMP ノードを形成している。POWER5+は大容量のキャッシュを搭載しているが，広範囲なアプリケーションで高性能を実現するためには，メモリ上の大規模データへのアクセス機能を高める必要がある。Hitachi SR11000/J2 ではこのために擬似ベクトル処理（Pseudo Vector Processing：PVP），コンパイラによるソフトウェアアシストプリフェッチがサポートされており，安定した高いメモリアクセス性能が実現されている〔14〕。各 SMP ノードは 128 GB

[12] http://www.cc.u-tokyo.ac.jp/

のメモリを搭載しており，全体システムでは，128 ノード（2,048 コア），18.8 TFLOPS のピーク性能，16.4 TB の主記憶容量である。各 SMP ノードは三次元クロスバーにより接続されており，双方向の転送速度は 12.0 GB/sec.である。

（2）三次元弾性問題（Cube モデル）

図 22 は SR11000/J1， SR11000/J2 の 8 コアを使用して様々な規模で三次元弾性解析（Cube モデル（図 8），色数＝100 または最内ループ長＞256）を実施した場合の結果である。表 3 は，表 2 と同様に，1 コアあたりの性能を予測した値と実測値とを比較したものである。図 10（b）で示したスカラープロセッサの典型的な挙動を示しており，問題規模が小さい場合はキャッシュを有効利用できているが，規模が大きくなると若干性能が低下する。しかしながら，その低下は SP3 の場合と比較して顕著ではない。表 3 に示すように，メモリバンド幅（実測）をピーク性能で割った BYTE/FLOP の値は，POWER3 と比較して POWER5，POWER5+は約 1.5 倍になっている。また，SR11000/J1，SR11000/J2 ともに，実測性能は予測値を大きく上回っている。予測値算定の際には，〔12〕の手法に基づき，キャッシュの効果を無視している。しかしながら，各 SMP ノードの 8 コアを使用した場合には，各チップ（2 コア）に装着された L2 キャッシュ（1.875 MB），L3 キャッシュ（36 MB）を 1 コアで利用できるため，問題規模が大きくなっても，キャッシュが効果的に機能しているものと考えられる。

図 23 は Hitachi SR8000/MPP（東京大学情報基盤センター）を使用した場合の結果である〔4,5,6〕。SR8000 では擬似ベクトル処理（PVP）の効果が高く，図 10（a）に示した「地球シミュレータ」の挙動にむしろ近くなっている。それと比較すると，SR11000/J1，SR11000/J2 では擬似ベクトル処理の効果は小さい。SR11000/J2 と SR8000 を比較すると，約 4 倍の速度向上が達成されている。

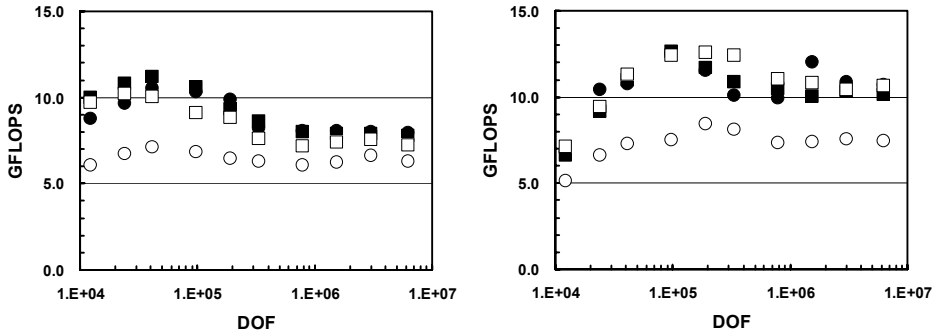
MPI と（●■）OpenMP（○□）の性能は変わらないが，OpenMP/DJDS（○）の性能は，他の場合と比較して低い。各チップに装着されたキャッシュを効率的に利用できないため，特に小規模問題では性能が悪い。

図 24 は，行列格納方法として DCRS を採用した場合に：

① SR11000/J1　8 コア（図 22 と同じ）●
② SR11000/J2　8 コア（図 22 と同じ）●
③ SR11000/J2　16 コア（全体問題規模が図 22 の場合と同じ）□
④ SR11000/J2　16 コア（コアあたり問題規模が図 22 の場合と同じ）△

について問題規模と計算性能（対ピーク性能比）の関係を示したものである。表 3，図 24 からわかるように，8 コアで比較すると，SR11000/J2（●）は SR11000/J1（●）よりも対ピーク性能比は高く，図 24（b）からもわかるように特に OpenMP の場合に顕著である。SR11000/J2 において，16 コアを使用した場合，コアあたりの問題規模による性能の差はほとんどなく（□，△），8 コアの場合からの性能低下はそれぞれ約 15%（MPI），約 21%（OpenMP）であり，OpenMP の場合の方が影響が大きい。また，OpenMP と MPI を比較すると，MPI の方が若干性能が良い。図 25 は SR11000/J2 において，コアあたりの問題規模を固定して，8 コア，16 コアを使用した場合の比較である。図 25（b）に見られるように，16 コアを使用した場合，

OpenMP/DJDS（○）と他の手法との差は 8 コアの場合と比較して更に顕著である。また 8 コア使用の場合（図 25（a））では，IBM SP3 で見られたような問題規模増加による性能低下（図 10 (b)）はほとんど認められなかったが，16 コアの場合はより明瞭に認められる（図 25 (b)）。8 コア使用の場合には，各チップに装着された L2 キャッシュ，L3 キャッシュを全て 1 コアで使用することが可能であるが，16 コアの場合は 2 コアでの共有となるため，問題規模増加による性能低下の効果がより顕著に認められる。



（a）Hitachi SR11000/J1　　（b）Hitachi SR11000/J2

図 22　Cube モデルの計算結果（問題規模と性能の関係），並列プログラミングモデル，行列格納方法の影響（●：MPI/DJDS，○：OpenMP/DJDS，■：MPI/DCRS，□：OpenMP/DCRS）

表 3　Hitachi SR11000/J1，Hitachi SR11000/J2 のハードウェア諸元と Cube モデルの性能〔4,6,7〕

	Hitachi SR11000/J1	Hitachi SR11000/J2	ES	IBM SP-3
Peak Performance/core (GFLOPS)	7.60	9.20	8.00	1.50
Measured Memory BW (GB/sec/core)	4.62	6.40	26.6	.623
Estimated Performance (GFLOPS (% of peak))	0.643-0.703 (8.45-9.24)	0.880-0.973 (9.56-10.6)	2.31-3.24 (28.8-40.5)	0.072-0.076 (4.80-5.05)
Measured Performance (GFLOPS (% of peak))	0.998 (13.1) (DCRS)	1.34 (14.5) (DCRS)	2.93 (36.6) (DJDS)	0.122 (8.11) (DCRS)
BYTE/FLOP	0.608	0.696	3.325	0.413

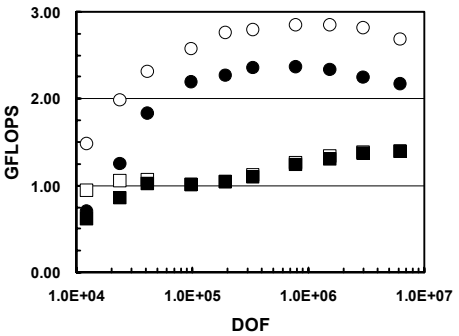
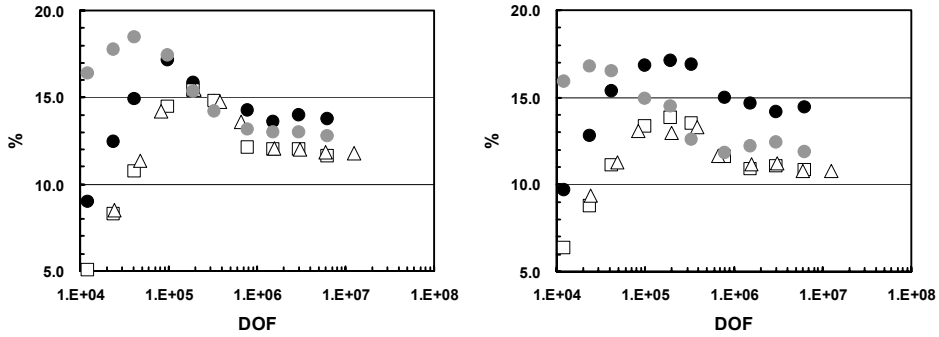
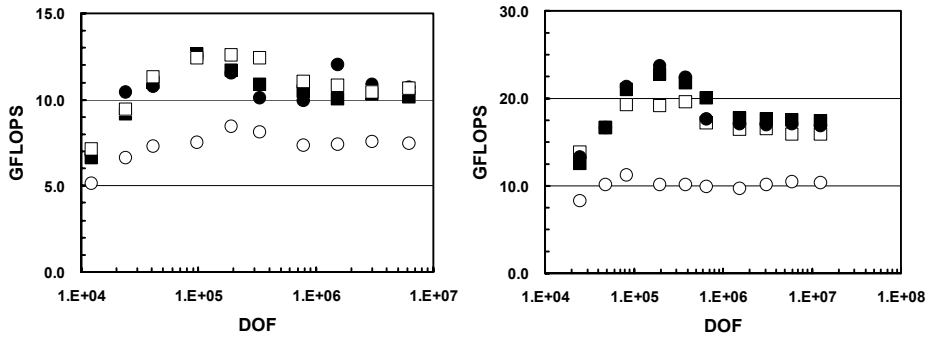


図 23　「Hitachi SR8000/MPP」による Cube モデルの計算結果（問題規模と性能の関係），並列プログラミングモデル, 行列格納方法の影響（●:MPI/DJDS, ○:OpenMP/DJDS, ■:MPI/DCRS, □：OpenMP/DCRS）〔4,5,6〕



（a）MPI　　（b）OpenMP

図 24　Cube モデルの計算結果（問題規模と性能（対ピーク性能比）の関係），並列プログラミングモデル，プロセッサ，SMP ノードあたりコア数の影響（● : SR11000/J1/DCRS 8 cores, ●:SR11000/J2/DCRS 8 cores, □:SR11000/J2/DCRS 16 cores（全体問題規模が●の場合と同じ），△ : SR11000/J2/DCRS 16 cores（コアあたり問題規模が●の場合と同じ））

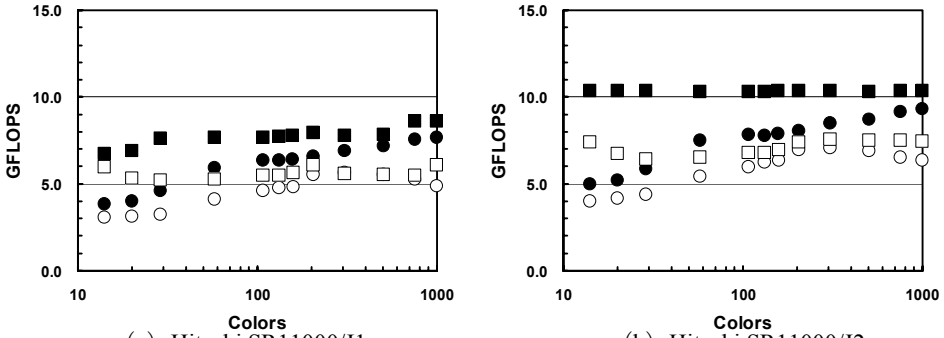


（a）8 cores/node　　（b）16 cores/node

図 25　「Hitachi SR11000/J2」における Cube モデルの計算結果（問題規模と性能の関係），並列プログラミングモデル，行列格納方法の影響，1 ノード 8 コアまたは 16 コアを使用した場合の比較，コアあたり問題規模は同じ（● : MPI/DJDS，○ : OpenMP/DJDS，■ : MPI/DCRS，□ : OpenMP/DCRS）

（3）三次元弾性問題（PGA モデル）

図 26 は様々な色数で三次元弾性解析（PGA モデル（図 9））を実施した場合の結果である。ここでは，スカラープロセッサに特有な DCRS（■□）＞DJDS（●○），MPI（●■）＞OpenMP（○□）という傾向がより顕著である（全体としては，■＞●＞□＞○）。SR11000/J1 と J2 の性能比は，図 22，表 3 の割合と対応している。DCRS では色数の性能に対する影響は小さいが，DJDS では色数が増加すると，3.（2）で述べたようにキャッシュがより有効に利用できるため，性能が高くなる。MPI/DJDS（●）の性能は色数の増加によって向上し，OpenMP/DJDS（○）の場合も 300 色程度までは色数の増加によって性能が向上する。色数が 300 以上では，OpenMP のオーバーヘッドによる性能の低下が見られるが，図 11 に示す IBM SP3 の OpenMP/DJDS（○）の場合と比較すると低下の度合いは小さい。これも BYTE/FLOP 値の増加，チップ内にメモリコントローラ内蔵，および L2 キャッシュ，L3 キャッシュの効果によるものと考えられる。

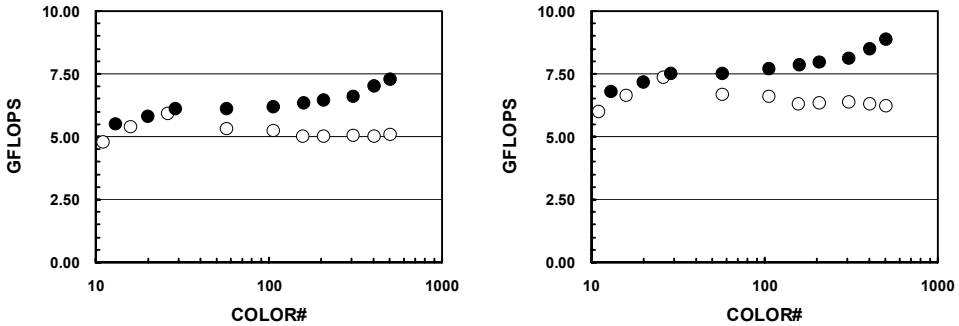


（a）Hitachi SR11000/J1　　（b）Hitachi SR11000/J2

図 26　PGA モデルの計算結果（色数と性能の関係），並列プログラミングモデル，行列格納方法の影響（●：MPI/DJDS，○：OpenMP/DJDS，■：MPI/DCRS，□：OpenMP/DCRS）

（4）三次元接触問題

図 27 は様々な色数で三次元接触問題（西南日本モデル（図 13））の計算を実施した場合の結果である。係数行列格納法としては DJDS のみ考慮した。色数の性能に対する効果は PGA の場合と同様である。

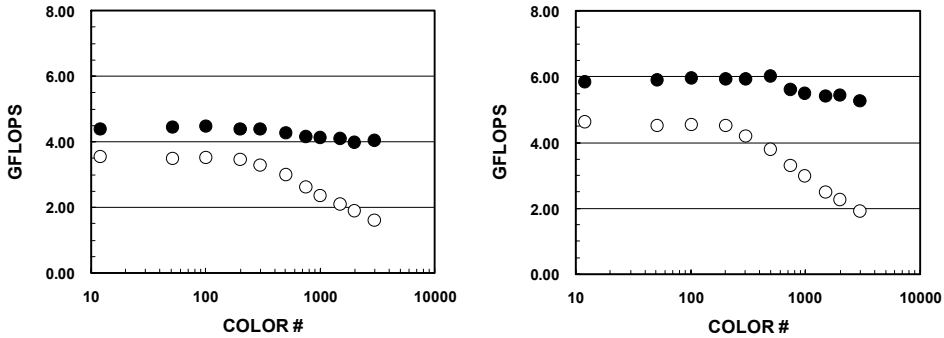


（a）Hitachi SR11000/J1　　（b）Hitachi SR11000/J2

図 27　接触問題（西南日本モデル）の計算結果（色数と性能の関係），並列プログラミングモデルの影響（●：MPI/DJDS，○：OpenMP/DJDS）

（5）二重球殻間領域三次元ポアソン方程式

図 28 は，様々な色数について，図 16 に示す二重球殻間領域におけるポアソン方程式をマルチグリッド前処理付き CG 法（MGCG）で解いた場合の計算性能である。係数行列格納法としては DJDS のみ考慮した。色数の性能に対する効果は PGA，接触問題の場合と同様である。また，図 17 で示した ES，IBM SP3 と比較すると，OpenMP を適用した場合の色数の増加による性能低下の割合は少ない。色数が 12 色と 2000 色の場合を比較すると計算性能の比はそれぞれ 1.86（SR11000/J1），2.07（SR11000/J2）となっている。これは図 17 に示した SP3（3.90）の場合と比較すると大幅な改善である。これも BYTE/FLOP 値の増加，チップ内にメモリコントローラ内蔵，および L2 キャッシュ，L3 キャッシュの効果によるものと考えられる。



（a）Hitachi SR11000/J1　　（b）Hitachi SR11000/J2

図 28　二重球殻モデル（多重格子法）の計算結果（色数と性能の関係），並列プログラミングモデルの影響（●：MPI/DJDS，○：OpenMP/DJDS）

5. Hitachi 11000 における実行例（複数ノード）

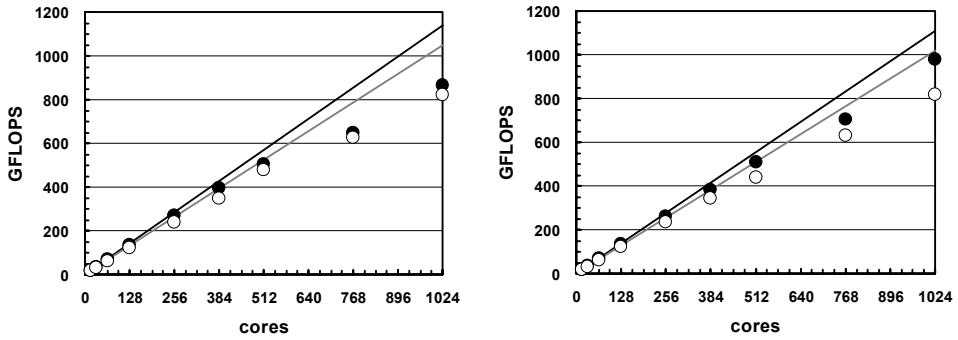
図 8 に示す三次元弾性問題（Cube モデル）について，SR11000/J2 において，コアあたりの問題規模を固定し，SMP ノード数を変化させた場合の計算（Weak Scaling）を実施した。係数行列格納法としては DCRS を適用した。各 SMP ノード 16 コアを使用して，1 コアあたり 98,304 DOF，786,432 DOF の場合について，64 ノード（1,024 コア）まで計算を実施した。1 ノードあたりの問題規模は，1,572,864 DOF（$=16\times3\times32^3$）および 12,582,912 DOF（$=16\times3\times64^3$）であり，最大問題規模は，100,663,296 および 805,306,368 DOF である。最高性能は 977 GFLOPS であり，これは SR11000/J2 の 64 ノード（1,024 コア）のピーク性能（9.42 TFLOPS）の 10.4%に相当する（図 29）。直線は，1 ノード（16 コア）における性能を基準とした，Flat MPI，Hybrid の場合の理想値である。図 30 は更にわかりやすいように 1 ノード（16 コア）からの計算性能増加率（Speed-Up）として表した図である。図 18～図 20 の場合と同様に，コアあたりの問題規模が小さい場合は，SMP ノード数が増加すると，通信のオーバーヘッドの影響が顕著となり，理想値との差が大きくなるが, 問題規模が大きい場合は, オーバーヘッドの影響は比較的少ない。

ここで注意しなければならないのは，ES，IBM SP3 の場合と異なり，コアあたりの問題規模が大きい場合，SMP ノード数が増加すると，Flat MPI の方が Hybrid よりも計算性能増加率が大きくなっていることである。3. で述べた〔2,4,5,6〕の結果によると，並列有限要素法の場合，SMP ノード数が増加すると，MPI latency の影響が顕著となり，特に MPI プロセス数の多い Flat

MPI はその影響を受けやすいため，Hybrid と比較して，相対的に性能が大幅に低下するような現象が見られた。しかしながら図 29，図 30 の例ではこれとは逆の現象が生じている。

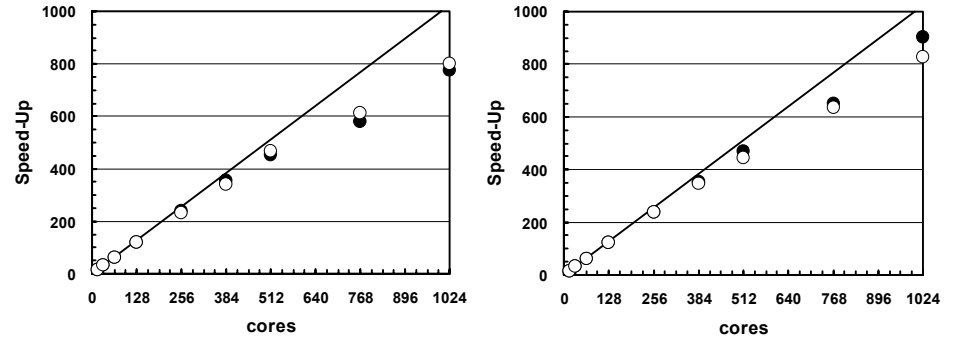
Hitachi SR11000/J2 では，SMP ノード間の通信用ポートとして各 SMP ノードに 2GB/sec のものが 6 本用意されている。Flat MPI の場合はこれが全て効率よく利用されているものと考えられる。Hybrid の場合，SMP ノードあたりの通信プロセスは図 31 に示すように一つであり，MPI_ISEND，MPI_IRECV の呼び出しの前後に送信バッファ（WS）へのコピー，受信バッファ（WR）からのコピーがあり，この部分は OpenMP によって並列化されている。

Hitachi SR11000/J2 では，Hybrid の場合は，使用するポート数は基本的に一つであるが，SMP ノード間通信量に応じて，使用ポートの数が動的に変化するような設定となっており，ポートあたりの通信量が 256KB を超える場合には，複数のポートに分割して送受信が行われる。有限要素法では領域間の通信量は比較的少ないが，コアあたり 3×64^3=786,432 DOF の場合は，各 SMP ノードの通信量は 4MB 弱，98,304 DOF（$=3\times32^3$）の場合は 1MB 弱のオーダーとなり，複数のポートが使用されている可能性がある。コアあたり（すなわち SMP ノードあたり）問題規模が大きくなり，SMP ノード間通信量が増加して，複数のポートが利用される場合の効果については，様々な問題規模，SMP ノード数における検討が必要である。



（a）3×32^3=98,304 DOF/core　　（b）3×64^3=786,432 DOF/core

図 29 「Hitachi SR11000/J2」における Cube モデルの計算結果（Weak Scaling）（●：Flat MPI/DCRS，○：Hybrid/DCRS，ー：Flat MPI/DCRS（ideal），—：Hybrid/DCRS （ideal））



（a）3×32^3=98,304 DOF/core　　（b）3×64^3=786,432 DOF/core

図 30 「Hitachi SR11000/J2」における Cube モデルの計算結果，1 ノードからの計算性能増加率（Weak Scaling）（●：Flat MPI/DCRS，○：Hybrid/DCRS，ー：ideal）

```
  do neib= 1, NEIBPETOT
    istart= STACK_EXPORT(neib-1)
    inum  = STACK_EXPORT(neib  ) - istart
!$omp parallel do private (k,ii)
    do k= istart+1, istart+inum
      ii   = 3*NOD_EXPORT(k)
      WS(3*k-2)= X(ii-2);WS(3*k-1)= X(ii-1);WS(3*k)= X(ii)
    enddo
!$omp end parallel do

    call MPI_ISEND (WS(3*istart+1), 3*inum, …)
  enddo

  do neib= 1, NEIBPETOT
    istart= STACK_IMPORT(neib-1)
    inum  = STACK_IMPORT(neib  ) - istart
    call MPI_IRECV (WR(3*istart+1), 3*inum, …)
  enddo

  call MPI_WAITALL (NEIBPETOT, req2, sta2, ierr)

  do neib= 1, NEIBPETOT
    istart= STACK_IMPORT(neib-1)
    inum  = STACK_IMPORT(neib  ) - istart
!$omp parallel do private (k,ii)
    do k= istart+1, istart+inum
          ii   = 3*NOD_IMPORT(k)
      X(ii-2)= WR(3*k-2);X(ii-1)= WR(3*k-1);X(ii)= WR(3*k)
    enddo
!$omp end parallel do
  enddo

  call MPI_WAITALL (NEIBPETOT, req1, sta1, ierr)
```

図 31　並列有限要素法（三次元弾性問題）の領域間通信（Hybrid 並列プログラミングモデル）（図 5，図 6 参照）

6.まとめ

本研究では，固体地球シミュレーション用並列有限要素法プラットフォーム「GeoFEM」において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群（GeoFEM ベンチマーク）を使用して，Hybrid および Flat MPI の両プログラミングモデルについて，本学情報基盤センターの Hitachi SR11000/J2 の性能評価を実施した。

GeoFEM ベンチマークでは，①三次元弾性問題（Cube・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，における並列前処理付き反復法ソルバーの実行性能（GFLOPS 値）を様々な条件下で計測する。プログラムは全て OpenMP ディレクティヴを含む FORTRAN90 および MPI で記述されており，SMP クラスタの性能評価に適している。様々なハードウェアに対応可能なように，連立一次方程式の係数マトリクスの格納法として，ベクトルプロセッサ向けの DJDS とスカラープロセッサ向けの DCRS の両方が使用可能である。

Hitachi SR11000/J2 では，IBM POWER5+　2 コアから構成されるチップ内にメモリコントローラが内蔵されており，かつ，大容量の L2 キャッシュ，L3 キャッシュを利用可能である。従来のスカラープロセッサ（IBM POWER3）を使用した IBM SP3 と比較して，高い BYTE/FLOP 値を示しており，スカラープロセッサ特有の問題規模の増加に伴う性能低下が非常に少なく，高い対ピーク性能比を示している。また，従来機種である Hitachi SR8000/MPP と比較すると，実効性能で約 4 倍の性能増加が得られている。

1 ノード内における MPI と OpenMP の差はほとんど無く，これもメモリコントローラ内蔵，大容量キャッシュの効果であると考えられる。OpenMP 使用時にマルチカラーオーダリングの色数を増加させることによる性能低下もほとんど見られず，DJDS によって係数行列を格納している場合には，逆にデータの局所性が増すことによる性能向上が顕著に見られた。特に，

MGCG 法による二重球殻間領域三次元ポアソン方程式のベンチマークの場合は，色数 2000 色まで増加させても， Hitachi SR11000/J2（OpenMP）で半分程度の性能低下に抑えられている。

複数ノードを使用した，Weak Scaling による性能評価については，これまでの「地球シミュレータ」，「IBM SP3」の性能評価から得られた知見によれば，SMP ノード数が増加した場合は Flat MPI よりも Hybrid の方が有利と考えられていた。これは，本研究で対象としている並列有限要素法においては，通信バンド幅（すなわち通信量）よりも，通信プロセス数（すなわちレイテンシ）に起因するオーバーヘッドの影響をより顕著に被るためである。

しかしながら，Hitachi SR11000/J2 の場合は，特にコアあたりの問題規模が大きい場合には，SMP ノード数が増加した場合, Flat MPI の方がむしろ Hybrid よりも性能が高いことがわかった。これは，Flat MPI と Hybrid において各 SMP ノード間の使用通信ポート数の決定法が異なり，Flat MPI では 6 本の通信ポート全てを常時使用するのに対して，Hybrid では通信量に応じて使用ポート数を調節していることと関連している可能性がある。更に詳細な検討が必要であるが，Flat MPI と Hybrid の選択にあたっては，こうしたハードウェアの特性を理解した上で，決定する必要がある。逆に，ハードウェア，特に大規模並列計算機のノード間ネットワークの設計にあたっては, アプリケーション特有の通信パターンを考慮することも重要であると考えられる。

1．でも触れたように，マルチコアプロセッサの普及によって，Flat MPI と Hybrid の優劣に関する議論は再び注目をあびつつある。現状では，いわゆるマルチコアプロセッサの能力は，本稿で紹介したハードウェアと比較して，特にメモリ関連の能力が劣るため，Hybrid，OpenMP では中々性能が出ない。ハードウェア能力の向上に期待する，というのも一つの考え方であるが，現状のハードウェアに適した，安定で効率の高い手法（特に反復法のための並列前処理手法）の開発は重要である。

謝 辞

本研究は, 東京大学21世紀COEプログラム「多圏地球システムの進化と変動の予測可能性」，および科学技術振興機構戦略的創造研究推進事業（CREST）の補助を受けている。計算機環境を提供いただいた東京大学情報基盤センター，地球シミュレータセンターおよび Lawrence Berkeley National Laboratory に謝意を表する。

参 考 文 献

[1] Rabenseifner, R. (2002) Communication Bandwidth of Parallel Programming Models on Hybrid Architectures. Lecture Notes in Computer Science 2327, 437-448

[2] Nakajima, K. (2003) Parallel Iterative Solvers of GeoFEM with Selective Blocking Pre-conditioning for Nonlinear Contact Problems on the Earth Simulator. ACM/IEEE Proceedings of SC2003

[3] 奥田洋司, 中島研吾　共編 (2004)「並列有限要素解析〔I〕クラスタコンピューティング」, 培風館

[4] Nakajima, K. (2004) Preconditioned Iterative Linear Solvers for Unstructured Grids on the Earth Simulator, IEEE Proceedings of 7th International Conference on High Performance Computing and Grid in Asia Pacific Region (HPC Asia 2004), 150-169

[5] 中島研吾（2005） SMPクラスタ型並列計算機におけるプログラミングモデル：Flat MPI vs. Hybrid，京都大学学術情報メディアセンター全国共同利用版広報5-2, 2-10

[6] Nakajima, K. (2005) Parallel programming models for finite-element method using preconditioned iterative solvers with multicolor ordering on various types of SMP cluster, IEEE Proceedings of 8th International Conference on High Performance Computing and Grid in Asia Pacific Region (HPC Asia 2005), 83-90

[7] 中島研吾（2006）GeoFEMベンチマークによる Hitachi SR11000／J1およびIBM p5-595のノード性能評価, 情報処理学会研究報告 2006-HPC-105-11, 61-66

[8] Saad, Y. (2003) “Iterative Methods for Sparse Linear Systems 2nd Edition”, SIAM

[9] Doi, S. and Washio, T. (1999) Using Multicolor Ordering with Many Colors to Strike a Better Balance between Parallelism and Convergence. Proceedings of RIKEN Symposium on Lin-ear Algebra and its Applications, 19-26

[10] Simon, H.D. (1991) Partitioning of unstructured problems for parallel processing, Computing Systems in Engineering 2, 135-148

[11] Carter, J., Oliker, L., and Shalf, J. (2007) Performance Evaluation of Scientific Applications on Modern Parallel Vector Systems, Lecture Notes in Computer Science 4395, 490-503

[12] Nakajima, K. (2005) Three-Level Hybrid vs. Flat MPI on the Earth Simulator: Parallel Iterative Solvers for Finite-Element Method, Applied Numerical Mathematics 54, 237-255

[13] Kerbyson, et al. (2002) A Comparison Between the Earth Simulator and Alpha Server Systems using Predictive Application Performance Models, LA-UR-02-5222, Los Alamos National Laboratory

[14] 青木秀貴，中村友洋，助川直伸，齋藤拡二，深川正一，中川八穂子，五百木伸洋（2005）スーパーテクニカルサーバーSR11000 モデル J1 のノードアーキテクチュアと性能評価, 情報処理学会論文誌：コンピューティングシステム Vol.45 No.SIG12（ACS11），27-36