

OpenMP/MPI ハイブリッド並列プログラミングモデルの多重格子法への適用

中島研吾^{†,††}

OpenMP/MPI ハイブリッド並列プログラミングモデルを、並列多重格子前処理付き反復法を使用した、三次元有限体積法に基づく不均質多孔質媒体中における地下水流れ問題シミュレーションに適用した。開発したプログラムの性能と安定性を T2K オープンスパコン（東大）、Cray-XT4 の 1,024 コアまでを使用して評価した。First Touch Data Placement, 連続メモリアクセスのためのデータ再配置, 適切な NUMA control の組み合わせにより、OpenMP/MPI ハイブリッド並列プログラミングモデルが Flat MPI と同等かそれを上回る性能を発揮することがわかった。

Parallel Multigrid Solvers using OpenMP/MPI Hybrid Programming Models

Kengo Nakajima^{†,††}

OpenMP/MPI hybrid parallel programming models were implemented to 3D finite-volume based simulation code for groundwater flow problems through heterogeneous porous media using parallel iterative solvers with multigrid preconditioning. Performance and robustness of the developed code has been evaluated on the “T2K Open Supercomputer (Tokyo)” and “Cray-XT4” using up to 1,024 cores. OpenMP/MPI hybrid parallel programming model demonstrated better performance and robustness than flat MPI with large number of cores for ill-conditioned problems with appropriate command lines for NUMA control, first touch data placement and reordering of the mesh data for contiguous access to memory.

1. はじめに

近年、マルチコアプロセッサの普及、大規模システムにおけるコア数の増加を背景として、ハイブリッド (Hybrid) 並列プログラミングモデルが脚光を浴びるようにな

り、Flat MPI (または Pure MPI) との優劣に関する議論が盛んとなっている。Hybrid 並列プログラミングモデルはメッセージパッシングによる「coarse-grain parallelism」と、ディレクティブによる「fine-grain parallelism」の融合であり、一般的には MPI と OpenMP を組み合わせたスタイルである。

著者は [1] において、有限要素法に基づく三次元弾性静力学問題向けシミュレーションで使用されている前処理付き反復法に OpenMP/MPI ハイブリッド並列プログラミングモデルを適用し、T2K オープンスパコン（東大）（以下「T2K（東大）」）[2] の 512 コアを使用した評価を実施した。OpenMP/MPI ハイブリッド並列プログラミングモデルは適切な NUMA control の組み合わせにより、OpenMP/MPI ハイブリッド並列プログラミングモデルが Flat MPI と同等かそれを上回る性能を発揮することがわかった。更に、First Touch Data Placement, 連続メモリアクセスのためのデータ再配置を適用することにより、特にコア当たり問題規模が小さい場合の性能が改善されることが明らかとなった。

本研究では、OpenMP/MPI ハイブリッド並列プログラミングモデルを、[3] で開発された、並列多重格子前処理付き反復法を使用した、三次元有限体積法に基づく不均質場における地下水流れ問題シミュレーションに適用した。多重格子法 (multigrid) は、大規模問題におけるスケーラブルな手法として注目されているが、T2K（東大）のようなマルチコア・マルチソケットクラスタにおいて、OpenMP/MPI ハイブリッド並列プログラミングを適用し、評価した例は無い。本研究では T2K（東大）の他、米国ローレンスバークレイ国立研究所 National Energy Research Scientific Computing Center (NERSC) の有する「Cray-XT4」[4] の 1,024 コアまでを使用して、Flat MPI と OpenMP/MPI ハイブリッド並列プログラミングモデルの評価を実施した。

2. 計算環境

本研究では、T2K オープンスパコン（東大）（T2K（東大））[2]、および Cray-XT4 (NERSC, National Energy Research Scientific Computing Center) 1,024 コアまでを使用して評価した。

T2K（東大）は筑波大、東大、京大の 3 大学で定められた「T2K オープンスパコン仕様」に基づき日立製作所が製作した 952 ノード、約 15,000 コア、ピーク性能 140 TFLOPS のクラスタ型コンピュータシステムである [5]。各ノードは cc-NUMA (Cache-Coherent NUMA) アーキテクチャに基づき AMD quad-core Opteron (2.3GHz) 4 ソケット、合計 16 コアから構成されている (図 1)、ノードあたりの記憶容量は 32GB (一部 128GB) である。Cray XT4 (Franklin) システムは Fig.1 に示した AMD quad-core Opteron (2.3GHz) 1 ソケットを 1 ノードとしたクラスタ型コンピュータシステムであり、9,572 ノード、約 38,000 コア、ピーク性能 352TFLOPS である。表 1 に両システム

[†] 東京大学情報基盤センター

Information Technology Center, The University of Tokyo

^{††} 科学技術振興機構 戦略的創造研究推進事業 (CREST)

CREST, Japan Science and Technology Agency (JST)

のネットワーク諸元を示す。ネットワークトポロジは T2K（東大）が多段クロスバーに対して、Cray XT4 は 3D トラス構造である。

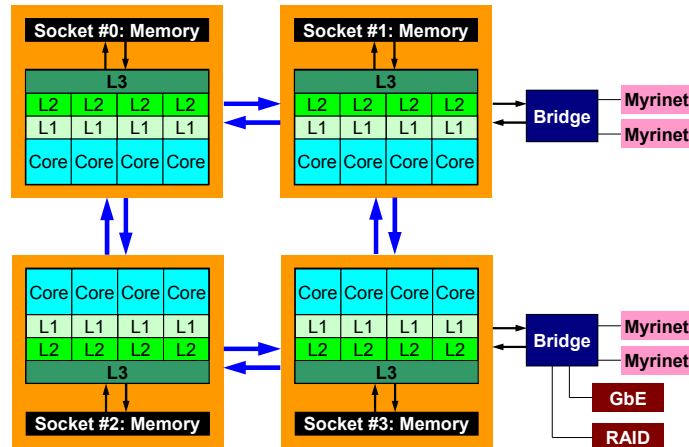


図1 T2K（東大）1 ノード，各ノードは AMD Quad-core Opteron（2.3GHz）4 つ搭載

表1 T2K（東大），Cray-XT4 のノード概要

	T2K（東大）	Cray-XT4
Interconnect	Myrinet-10G×4	Cray SeaStar2
Network Topology	Multistage Crossbar	3D Truss
Comm. Bandwidth (GB/sec)	5.0	7.6
Comm. Latency (μsec)	2.0	5.0 (nearest neighbor) 6.0 (far-away nodes)

3. アプリケーション，実装

3.1 三次元有限体積法に基づく不均質場における地下水流れ問題シミュレーション

本研究では，図2に示すような不均質な多孔質媒体中の三次元地下水流れを並列有限体積法（Finite Volume Method, FVM）によって解くアプリケーションを扱う。対象とする問題は以下に示すような，ポアソン方程式および境界条件である：

$$\nabla \cdot (\lambda(x, y, z) \nabla \phi) = q, \phi = 0 \text{ at } z = z_{\max}$$

ここで， ϕ は水頭ポテンシャル， $\lambda(x, y, z)$ は透水係数で位置座標の関数であり，セル（cell）ごとに異なっている。透水係数は，地質統計学の分野で使用される Sequential Gaussian アルゴリズム [6] により発生させた値を使用した（図2（a））。 q は体積フラックスであり，本研究では一様（=1.0）に設定されている。

透水係数の最小値，最大値，平均値はそれぞれ 10^{-5} ， 10^{+5} ， 10^0 となるように設定されている。有限体積セルは一辺長さ 1.0 の立方体である。このような問題設定では，条件数が 10^{10} のオーダーとなるような対称，正定な悪条件マトリクスを係数とする線形方程式を解く必要がある。本研究で対象とするモデルは，各々 128^3 セルから構成される同じ不均質場に基づく部分モデルの集合である。したがって， x ， y ， z 各方向に周期的に同じ不均質パターンが繰り返される。

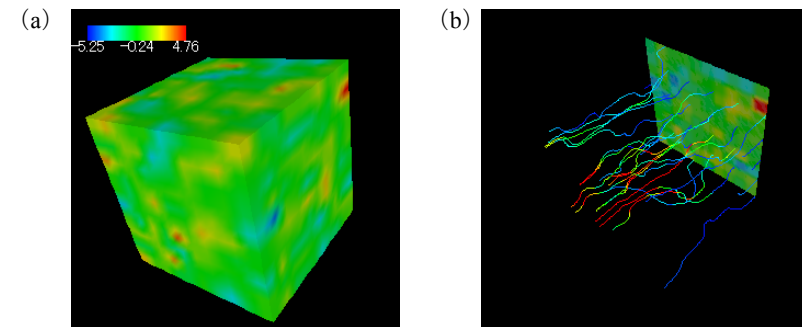


図2 不均質多孔質媒体中の地下水流れの例，(a) 透水係数分布，(b) 流線

3.2 多重格子法による前処理付き反復法

本研究では，ポアソン方程式を有限体積法によって離散化して得られる対称，正定（Symmetric Positive Definite, SPD）な行列を係数行列とする連立一次方程式を，多重格子法（Multigrid）による前処理を施した共役勾配法（Conjugate Gradient Method, CG）によって解く。このような前処理付き共役勾配法を MGCG 法 [3] と呼ぶ。残差ノルム $\|b\| - [A]\{x\} / \|b\|$ が 10^{-12} 未満となるまで反復が繰り返される。

多重格子法は大規模問題向けのスケラブルな解法として注目されている。Gauss-Seidel 法などの古典的反復法はセルサイズに相当する波長をもった誤差成分の減衰には適しているが，誤差の成分のうち，長い波長の成分は緩和を繰り返しても中々

収束しない。多重格子法は、長い波長の成分が粗い格子上で効率的に減衰するという考え方に基いている [7]。多重格子法は、細かい格子において対象とする線形方程式の残差を計算し、修正方程式を粗い格子へ補間（制限補間, **restriction**）して解き、その結果を細かい格子に補間（延長補間, **prolongation**）して誤差を補正するというプロセスを、再帰的に多段階に適用することによって構築可能である。各レベルの計算が適切に実施されれば、誤差のあらゆる長さの波長をもった成分を一樣に減衰させることができるため、計算時間が問題規模に比例するいわゆる「scalable」な手法の実現が可能である。本研究では、図 3 に示すように、8 個の「子 (children)」セルから 1 個の「親 (parent)」セルが生成されるような等方的な幾何学的多重格子法に基づき、格子間のオペレーションとしては、最密格子と最疎格子の間を直線的に動く V サイクル [7] を採用した。本研究では、各レベルにおける多重格子法のオペレーションは並列に実施されるが、最も粗い格子レベル (図 3 における Level=k) では 1 コアに集めて計算を実施する。

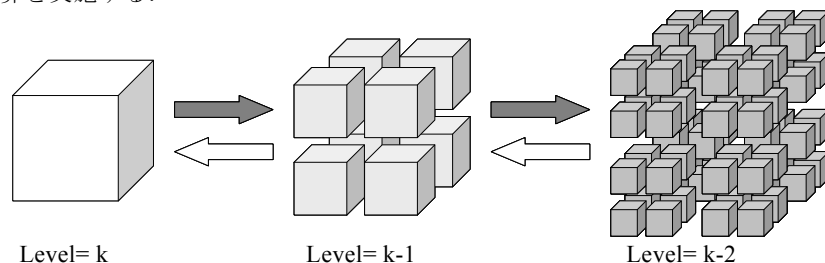


図 3 幾何学的多重格子法のプロセス (8 children=1 parent)

多重格子法では、各レベルにおける線形方程式を緩和的に計算するための演算子を緩和演算子 (smoothing operator, smoother) と呼んでいる。緩和演算子として代表的なものは Gauss-Seidel 法であり多くの研究で使用されているが、悪条件問題向けには不完全 LU 分解、不完全コレスキー分解が有効である [3,7,8]。本研究では、フィルインを生じない不完全コレスキー分解 (IC(0)) を緩和演算子として採用した。IC(0)のプロセス (分解, 前進後退代入) は大域的な処理を含むため、並列化は本来困難である。各領域において独立に IC(0)処理を実施するような、ブロック Jacobi 型の局所処理によって並列化は可能であるが、特に悪条件問題の場合、領域数が増えると収束が悪化する。ここで、加法シュワルツ法 (Additive Schwarz Domain Decomposition, 以下 ASDD) [9] を組み合わせることにより、並列計算においても安定した解を得ることが可能となる。ASDD 法のアルゴリズムは以下の通りである：

① M を全体前処理行列, r と z をベクトルとして, $Mz=r$ を前進後退代入によっ

て解くものとする。

② 全体領域を図 4 (a) に示すような 2 領域, すなわち, Ω_1 および Ω_2 に分割したと仮定し、各領域で独立に局所前処理を実施する：

$$z_{\Omega_1} = M_{\Omega_1}^{-1} r_{\Omega_1}, \quad z_{\Omega_2} = M_{\Omega_2}^{-1} r_{\Omega_2}$$

③ 各領域間のオーバーラップ領域 Γ_1 および Γ_2 の効果を次式によって導入する (図 4 (b))。ここで n は ASDD のサイクル数である：

$$z_{\Omega_1}^n = z_{\Omega_1}^{n-1} + M_{\Omega_1}^{-1}(r_{\Omega_1} - M_{\Omega_1} z_{\Omega_1}^{n-1} - M_{\Gamma_1} z_{\Gamma_1}^{n-1}), \quad z_{\Omega_2}^n = z_{\Omega_2}^{n-1} + M_{\Omega_2}^{-1}(r_{\Omega_2} - M_{\Omega_2} z_{\Omega_2}^{n-1} - M_{\Gamma_2} z_{\Gamma_2}^{n-1})$$

④ ②, ③を繰り返す。

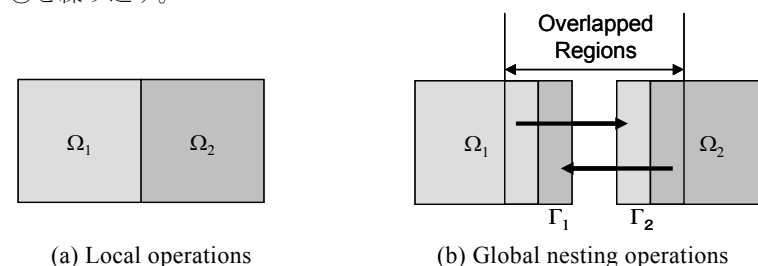


図 4 加法シュワルツ法 (Additive Schwarz Domain Decomposition, ASDD)

3.3 リオーダリング, 最適化手法

OpenMP/MPI ハイブリッド並列プログラミングモデルで、FVM によるアプリケーションを並列化する場合、領域分割された各領域に MPI のプロセスが割り当てられ、各領域内で OpenMP による並列化が行われる。各領域においては、不完全コレスキー分解のように大域的な依存性を含むプロセスについては、各要素の並び替え (reordering) により依存性を排除し、並列性を抽出する手法が広く使用されている [1]。本研究では、並列性が高く悪条件問題に対して安定な CM-RCM 法による並び替えを適用している [1]。本手法は、Reverse Cuthill-McKee (RCM) 法と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) を組み合わせたものである。CM-RCM 法では各「色」内の要素は独立で、並列に計算を実行することが可能である。CM-RCM 法の色数の最大値は RCM におけるレベル数の最大値である。本研究では多重格子法の各レベルにおいて CM-RCM 法を適用している。

T2K (東大) において OpenMP/MPI ハイブリッド並列プログラミングモデルを使用する場合、NUMA アーキテクチャの特性を利用するための実行時制御コマンド (NUMA control) を使用して、コア (またはソケット) 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既に明らかとなっている [1]。本研究で

は、様々な実行時制御コマンドの組み合わせの中で最適のものを選択して適用した。

更に、①First Touch Data Placement [10] の適用、②連続データアクセスのためのデータ再配置によって性能の改善を実施する [1]。NUMA アーキテクチャでは、プログラムにおいて変数や配列を宣言した時点では、物理的メモリ上に記憶領域は確保されず、ある変数を最初にアクセスしたコア（の属するソケット）のローカルメモリ上に、その変数の記憶領域が確保される。これを First Touch Data Placement [10] と呼び、配列の初期化手順により大幅な性能の向上が達成できる場合もある。具体的には、実際の計算の手順にしたがって配列を初期化することによって実現できる。CM-RCM 法による並べ替えでは、①同一の色（またはレベル）に属する要素は独立であり、並列に計算可能、②「色」の順番に番号付け、③色内の要素を各スレッドに振り分ける、という方式 [1] を採用しているが、同じスレッド（すなわち同じコア）に属する要素は連続の番号では無いため、効率が低下している可能性がある。図 5 に示すように同じスレッドで処理するデータをなるべく連続に配置するように更に並び替え、更に First Touch Data Placement を適用することによって性能向上を図る [1]。

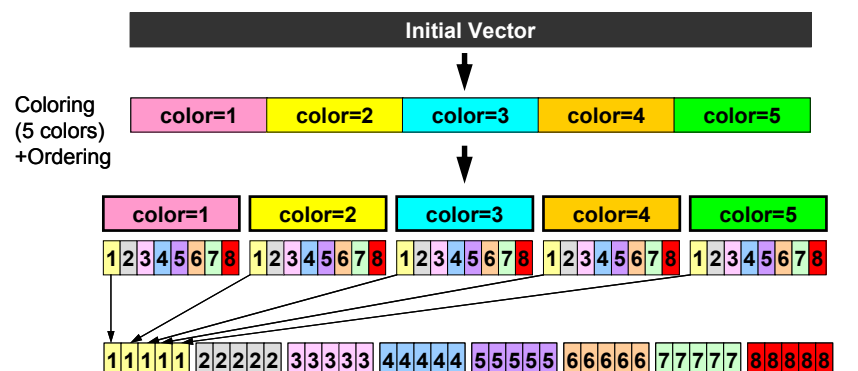


図 5 連続データアクセスのためのデータ再配置 (5 色, 8 スレッドの場合)

4. 計算結果

4.1 最適化の効果

提案手法の安定性と効率について、T2K（東大）、Cray-XT4 を使用して評価した。MGCG 法の多重格子法部分の緩和演算子としては IC(0)を適用し、V サイクルの各レベルにおいて 2 回の反復、また各反復において ASDD を 1 回適用した。CG 法の各反復において V サイクル 1 回を適用した。

以下に示す、3 種類の OpenMP/MPI ハイブリッド並列プログラミングモデルを適用し、全コアに独立に MPI プロセスを発生させる Flat MPI と比較した：

- **Hybrid 4×4 (HB 4×4)**：スレッド数 4 の MPI プロセスを 4 つ起動する、各ソケットに OpenMP スレッド×4、ノード当たり 4 つの MPI プロセス
- **Hybrid 8×2 (HB 8×2)**：スレッド数 8 の MPI プロセスを 2 つ起動する、2 ソケットに OpenMP スレッド×8、ノード当たり 2 つの MPI プロセス (T2K (東大) のみ)
- **Hybrid 16×1 (HB 16×1)**：1 ノード全体に 16 の OpenMP スレッド、1 ノード当たりの MPI プロセスは 1 つ (T2K (東大) のみ)

Cray-XT4 は 1 ノードあたり 1 ソケット (4 コア) であるため、上記のうち HB 4×4 のみを適用した。

まず、最初に 3.3 で述べたリオーダーリング、最適化の効果を評価するため、64 コア (T2K (東大)：4 ノード、Cray-XT4：16 ノード) を使用した評価を実施した。各コアにおける問題サイズ (セル数) は 262,144 (=64³)、全問題サイズは 16,777,216 である。図 6 は、各並列プログラミングモデルにおける、収束までの反復回数と CM-RCM の色数の関係である。

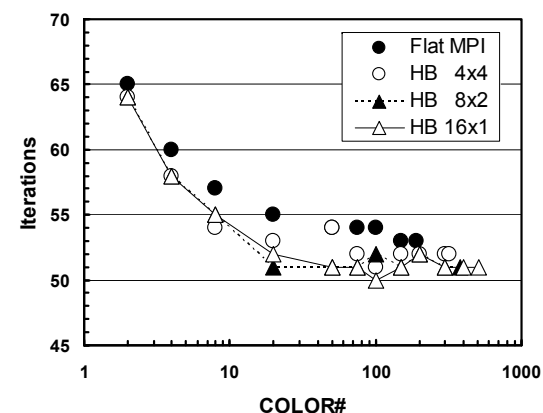


図 6 MGCG 法の反復回数と色数の関係 (16,777,216 セル, 64 コア)
(不均質多孔質媒体中の三次元地下水流れ)

一般に色数が増加すると Incompatible Nodes の数が減少するため反復回数は減少する [12]. 本研究で対象するような不均質な問題の場合、必ずしもその通りでは無いが、図 6 に示すように、一般的傾向としては色数の現象とともに、反復回数は減少している. 図 7 は、3.3 で示した最適化 (NUMA Control, First Touch Data Placement, 連続データアクセスのための再データ配置) を適用した後の、各プログラミングモデルの各色における計算性能である.

図 6 に示したように、色数が増えると反復回数が減少しているにもかかわらず、図 7 (a) に示すように、各並列プログラミングモデルにおいて、色数=2 の場合 (CM-RCM(2)) の計算時間 (MGCG ソルバーの計算時間) が最も短い. 反復あたりの計算時間についても、図 7 (b) に示したように、CM-RCM(2)が他と比べて小さく、性能が高いことがわかる.

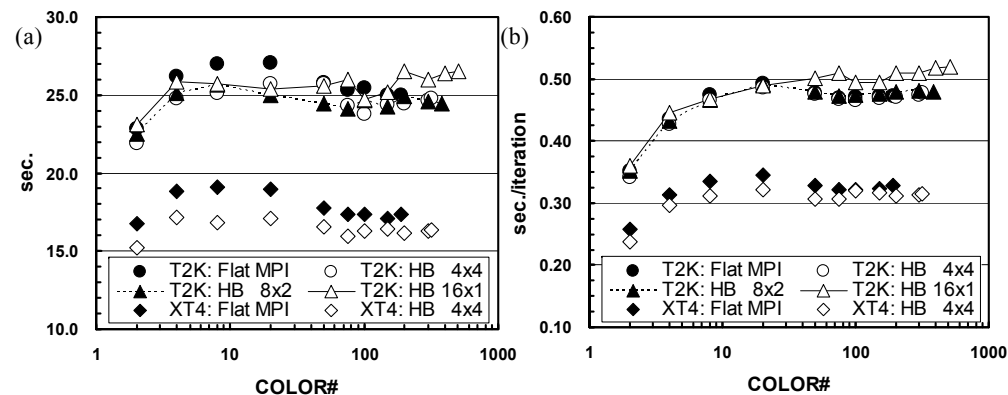


図 7 MGCG 法の計算性能と CM-RCM 法の色数の関係 ((a) MGCG ソルバー計算時間, (b) 1 反復あたり計算時間) (16,777,216 セル, 64 コア) (不均質多孔質媒体中の三次元地下水流れ) (3.3 に示した最適化 (NUMA Control, First Touch Data Placement, 連続データアクセスのための再データ配置 (図 5) 適用後))

表 2 T2K (東大), Cray-XT4 の性能概要

	T2K (東大)	Cray-XT4
STREAM: Triadd [12] (GB/sec)	19.7	24.6
GeoFEM Benchmark [13] (Relative Performance)	1.00	1.67

Cray-XT4 の性能は T2K (東大) よりも 40%~50% 高い. 表 2 に示すように, STREAM ベンチマーク [12] で測定したメモリの性能は 25% 程度 Cray-XT4 が高く, 更に Cray-XT4 では, cache coherency の影響が無いため, 特に粗いレベルにおいてキャッシュがより有効に利用されていると考えられる.

図 8 は T2K (東大) において, 3.3 で述べた最適化の効果を各並列プログラミングモデルについて CM-RCM(2)の場合について比較したものである. 実行時に NUMA control を適用することで, 特に OpenMP/MPI ハイブリッド並列プログラミングモデルは 2 倍以上の高速化が可能である. 更に, First Touch Data Placement, 連続データアクセスのための再データ配置 (図 5) を適用することにより, Flat MPI と OpenMP/MPI ハイブリッド並列プログラミングの性能はほぼ同等となる. HB 4x4 では, もともとデータが各ソケットのローカルメモリに配置されているため, First Touch Data Placement と再データ配置の効果はわずかである.

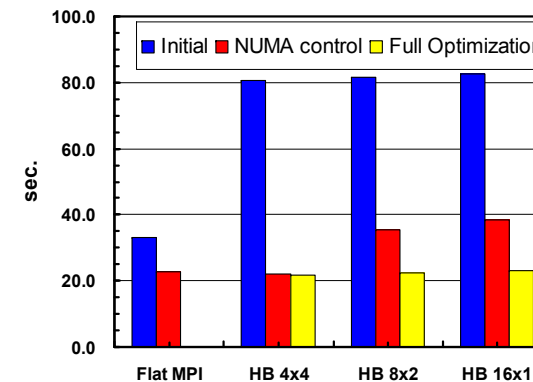


図 8 MGCG 法の計算性能と最適化の効果 (MGCG ソルバー計算時間) (CM-RCM(2), 16,777,216 セル, T2K (東大), 64 コア) (不均質多孔質媒体中の三次元地下水流れ) (■ Initial : 初期状態, ■ NUMA control : 実行時の最適な NUMA control の適用, ■ Full Optimization : NUMA Control, First Touch Data Placement, 連続データアクセスのための再データ配置を全て適用した場合)

4.2 大規模問題

T2K (東大), Cray-XT4 の 16~1,024 コアを使用して, 大規模問題に対する性能と安定性を評価した. コアあたりの問題規模を固定した Weak Scaling によって評価を実施した. コアあたり問題サイズ (セル数) は 262,144 ($=64^3$) であり, 最大問題規模は 268,435,456 セルである. 問題設定は 4.1 と同じであり, 図 8 に示した最適化されたソ

ルバーを使用し、CM-RCM(2)を適用した。

図9は、16コア～1,024コアを利用した場合のMGCG法の計算性能である。図9(a)は収束までの反復回数である。完全にスケラブルな場合は問題規模によって、反復回数は変化しないが、本研究で扱っているような悪条件問題では、問題規模が大きくなるに従って、反復回数が若干増加している。その傾向はFlat MPIにおいて特に顕著である。コア数が256(問題規模としては約 10^8 セル)を超えると、Flat MPIの反復回数の増加が顕著になる。それと比較すると、OpenMP/MPIハイブリッド並列プログラミングモデルの場合は若干の増加は見られるもののFlat MPIほど顕著では無い。本研究では、基本的にはブロックJacobi型の局所IC(0)前処理を使用しており、ASDDによる安定化が図られているものの、一般に領域数が増加するほど反復回数が増加する傾向にある。したがって、1つのMPIプロセスあたりのスレッド数を増やすほど反復回数の増加を抑制することが可能であると考えられる。実際、図9(a)に見られるように、HB 16×1は他の並列プログラミングモデルと比較して、反復回数の増加は低く抑えられている。また、ここでは全ケースについて、64コアの場合に計算時間が最も少ないCM-RCM(2)を適用しているが、特にコア数が増加した場合、CM-RCMの色数を増やしたり、RCMなど収束性の良い方法を適用することが有効となる場合もあると考えられる。

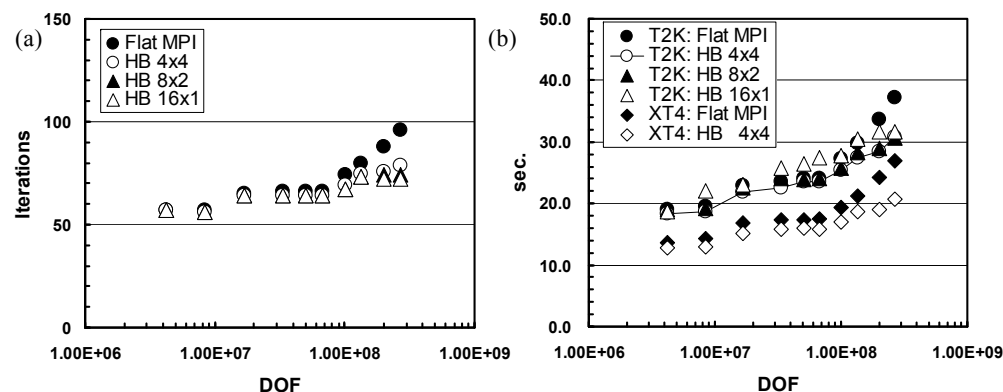


図9 MGCG法の計算性能((a)反復回数, (b)MGCGソルバー計算時間)(最適化されたソルバーを適用, CM-RCM(2), 16～1,024コア)(Weak Scaling, 262,144セル/コア, 最大問題規模: 268,435,456セル)(不均質多孔質媒体中の三次元地下水流れ)

5. まとめ

OpenMP/MPIハイブリッド並列プログラミングモデルを、並列多重格子前処理付き反復法を使用した、三次元有限体積法に基づく不均質多孔質媒体中における地下水流れ問題シミュレーションに適用した。多重格子法の緩和演算子としては、IC(0)を適用した。開発したプログラムの性能と安定性をT2K オープンスパコン(東大)、Cray-XT4の1,024コアまでを使用して評価した。First Touch Data Placement, 連続メモリアクセスのためのデータ再配置, 適切なNUMA controlの組み合わせにより、OpenMP/MPIハイブリッド並列プログラミングモデルがFlat MPIと同等かそれを上回る性能を発揮することがわかった。特に問題規模が大きくなり、コア数が増加すると、OpenMP/MPIハイブリッド並列プログラミングモデルの優位性は顕著である。

参考文献

- [1] Nakajima, K.: Flat MPI vs. Hybrid: Evaluation of Parallel Programming Models for Preconditioned Iterative Solvers on "T2K Open Supercomputer", IEEE Proceedings of the 38th International Conference on Parallel Processing (ICPP-09), pp.73-80 (2009)
- [2] Information Technology Center, The University of Tokyo: <http://www.cc.u-tokyo.ac.jp/>
- [3] 中島研吾, 不均質場におけるマルチレベル解法, ハイパフォーマンスコンピューティングと計算科学シンポジウム HPCS2006 論文集, pp.95-102 (2006)
- [4] NERSC, Lawrence Berkeley National Laboratory: <http://www.nersc.gov/>
- [5] The T2K Open Supercomputer Alliance: <http://www.open-supercomputer.org/>
- [6] Deutsch, C.V., Journel, A.G.: GSLIB Geostatistical Software Library and User's Guide, Second Edition. Oxford University Press (1998)
- [7] Tottemberg, U., Oosterlee, C. and Schuller, A.: Multigrid, Academic Press (2001)
- [8] Nakajima, K.: Parallel Multilevel Iterative Linear Solvers with Unstructured Adaptive Grids for Simulations in Earth Science, Concurrency and Computation: Practice and Experience 14-6/7, pp.484-498 (2002)
- [9] Smith, B., Bjørstad, P. and Gropp, W.: Domain Decomposition, Parallel Multilevel Methods for Elliptic Partial Differential Equations, Cambridge Press (1996)
- [10] Mattson, T.G., Sanders, B.A., Massingill, B.L.: Patterns for Parallel Programming, Software Patterns Series (SPS), Addison-Wesley (2005)
- [11] Washio, T., Maruyama, K., Osoda, T., Shimizu, F., Doi, S.: Efficient implementations of block sparse matrix operations on shared memory vector machines. Proceedings of The 4th International Conference on Supercomputing in Nuclear Applications (SNA2000) (2000)
- [12] STREAM (Sustainable Memory Bandwidth in High Performance Computers): <http://www.cs.virginia.edu/stream/>
- [13] 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反復法の性能, 情報処理学会研究報告 (HPC-120-6) (2009)