Vol.2009-ARC-186 No.2
Vol.2009-HPC-123 No.2
2009/11/30

Cray XT5 における Hybrid 並列数値流体アプリケーションの性能評価

西 條 晶 彦[†1]　松 澤 照 男[†1]

SMP Cluster 型の並列計算機で MPI を用いて領域分割による並列化を行った場合，高並列時には並列化効率が落ちる場合がある．そこで MPI による領域分割と OpenMP によるデータ分割の手法をを組み合わせた Hybrid 並列の技術が注目されている．しかし，Hybrid 並列の性能はアプリケーションの特性やアーキテクチャの通信特性・メモリ特性に依存する．ここでは流体計算に特有の疎行列を扱うアプリケーションを，2009 年に北陸先端大に導入された SMP Cluster である Cray XT5 上で動作させそれらの特性を調査する．

Performance Evalutation of Hybrid MPI CFD Applications on Cray XT5

Akihiko SAIJO [†1] and Teruo Matsuzawa[†1]

When SMP Cluster type MPP is used for domain docompotiosin in massive core numbers, it tends to occur that parallel efficiecy drops. So Now Hybrid Parallelism, mixing MPI parallelism and OpenMP data decomposition parallelism are desirable in the furture. But the methodorody of Hybrid parallelism are highly dependent on the properties of the application and architecture communcation and memory characterists. This report shows that the evaluation of the hybrid paralle sparse matrix solver, using the Cray XT5 SMP Cluster introduced in Japan Advanced Institute of Science and Technology since 2009.

†1 北陸先端科学技術大学院大学
Japan Advanced Institute of Science and Technology

1. はじめに

大規模な数値シミュレーションを高速に行うために並列計算機の効率的な需要が高まっている．並列計算機は大きく 3 つにわけて，巨大メモリを多くのプロセッサで共有する共有メモリ型 (SMP)，1 プロセッサ計算機をネットワークで多数繋いだ分散メモリ型 (Cluster)，この二つの組み合わせた共有メモリ型計算機をネットワークで多数繋いだ分散共有メモリ型 (SMP Cluster) がある．現在の超並列計算機はほとんどすべて分散共有メモリ型のアーキテクチャをとっている．

このような分散共有メモリ型のアーキテクチャで高い性能を出すために Hybrid 並列の手法がある．Hybrid 並列とは，並列計算において分散並列 (MPI) と共有メモリ並列（スレッド）を組み合わせて高速化をはかるプログラミングモデルである．分散共有メモリ型のマシンにおいてはノード内を共有メモリで，ノード間をメッセージパッシングで並列化する，というのがアーキテクチャの観点から見て自然である．

しかしながら，Hybrid 並列は実装が煩雑になることや，問題の特性によって並列度が小さい場合にはかえって性能を落としてしまうことがあることもあって，あまり普及しなかった．しかし，2 コア，4 コアといったメニーコアの形態の CPU が普及し，1 万コアを越える SMP Cluster が登場してきたことにより，Hybrid 並列の手法が再び注目されるようになってきた．

ここでは 2009 年に北陸先端大学に導入された Cray Inc. 大規模並列計算機である Cray XT5 を用い，NAS Parallel Benchmarks、並列共役勾配法ソルバにハイブリッド並列化をおこなってその性能を調べた．

2. Hybrid 並列

Hybrid 並列とは、MPI などによるメッセージ通信並列手法と、OpenMP などによる共有メモリの並列手法を組み合わせる並列プログラミングモデルのことである。

2.1 共有メモリ並列と分散並列

共有メモリ並列の手法と分散並列の手法は、それぞれ特性はあるものの並列アルゴリズムとして本質的な違いはない。しかしながら、一般的な数値計算では大規模・高並列になると計算時間に対して通信の時間の割合が多くなる。このため、共有メモリ並列の分だけ分散並列の並列度を下げてやることにより通信時間を減らすことができる。また、共有メモリ並列はコンパイラによる自動並列化が行いやすい。

Vol.2009-ARC-186 No.2
Vol.2009-HPC-123 No.2
2009/11/30

これらの共有メモリ並列の優位性を分散並列に持ち込む、というのが Hybrid 並列の考え方である。

もちろん、分散並列に用いる MPI の実装では、内部においてすでに共有メモリ並列を用いてられているため，もしも MPI 実装が十分に賢ければ Hybrid 並列は行う必要がなくなる。さらに，OpenMP による共有メモリ並列はスレッド生成によるオーバヘッドやメモリ同期の負荷が高いため，コードの特性や OpenMP の実装の仕方によって，Pure MPI のほうが性能が出る場合もある．Hybrid 並列の効果が高くなりやすいのは、並列度が高く（Hybrid 化でメッセージ通信のレイテンシの影響を抑えられる)、コアあたりの問題規模が小さい（メモリ負荷が少ない）場合である。

長　　所

- コア通信間のオーバーヘッドが少ない
- スレッドによる処理はプロセスによる処理よりも軽い
- 均等な負荷分散が行いやすい

短　　所

- 2 つのパラダイムを混ぜるためプログラムが複雑になりやすい
- スレッド生成を頻繁に行ったり，変数の同期が多い場合は分散並列よりも遅くなる

3. Cray XT5

本報告のターゲットとなる Cray XT5 の概略を述べる．Cray XT5 は Cray Research Inc. の開発した大規模並列計算機である．

北陸先端大における構成では，1 ノードあたりに動作周波数 2.4 GHz の AMD Opeteron Quad Core (Shanghai) プロセッサ を 2 ソケット，トータルで 8 PE (Processing Element) を 16GB のメモリで用いることができる．内部ネットワークは Cray SeaStar2+ チップにより各ノードを 3 次元トーラス形状に接続している（双方向理論バンド幅 9.6 GB/sec）．合計で 256 ノードを利用して最大 2048 並列，総メモリ 4TB，理論性能 19.6 TFlops に及ぶ計算が可能である．

4. 通 信 性 能

SMP Cluster である Cray XT5 ではコア間，ソケット間，ノード間での通信性能がそれぞれ特性を持っている．MPI による分散並列を行う場合，この階層的な通信特性が重要な意味を持つ．

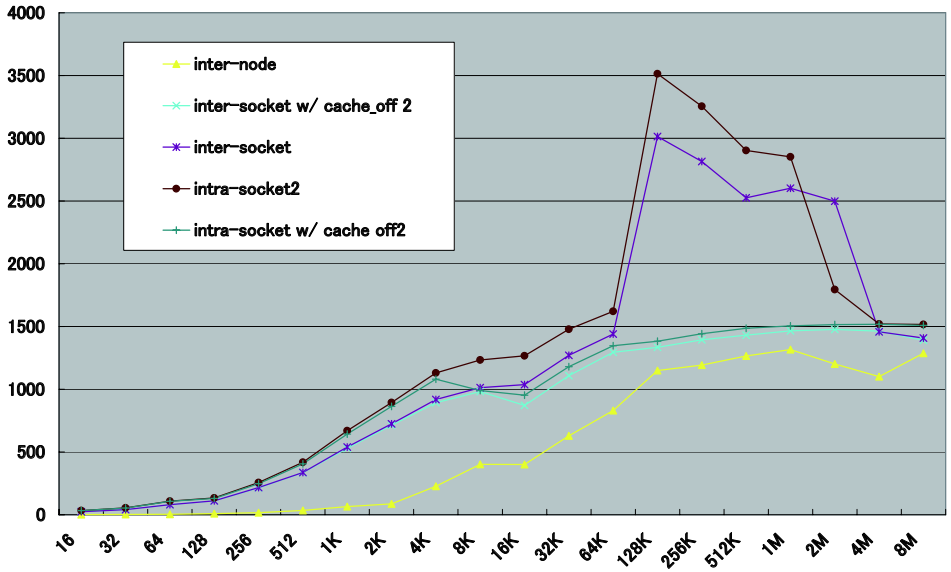

図 1　MPI スループット
Fig. 1　MPI throughput

4.1 Intel MPI Benchmarks

Cray XT5 における MPI の通信性能を IMB (Intel MPI Benchmarks) を使って計測した．図**??**に示すのはメッセージサイズを変更しながらプロセス間で MPI_Send と MPI_Recv を行う MPI Pingpong プログラムである．

5. ハイブリッド並列アプリケーション性能

流体などの領域型の解析では，計算領域を小領域に分割し，境界部分において MPI による通信を行って，同期をとることにより並列計算を行うのが一般的である．そのような領域分割の MPI コードに OpenMP のディレクティブを挿入することでアプリケーションを Hybrid 並列化することができる．このような数値流体アプリケーションを 2 つとりあげ，Cray XT5 上での Pure MPI と性能を比較する．

Vol.2009-ARC-186 No.2
Vol.2009-HPC-123 No.2
2009/11/30

5.1 NAS Parallel Benchmark, Multi-Zone Versions

NAS Parallel Benchmark (NPB)[1] は NASA Ames Research Center の NASA Advanced Supercomputing (NAS) 部門が開発した科学技術計算のベンチマークパッケージである．NPB は実装において MPI，Java, High Performance Fortran, OpenMP などいくつか派生がある．本報告では，そのうちハイブリッド並列化を施したバージョンである Multi-Zone Version (NPB MZ)[2] を用いる．

5.2 解析モデル

NPB MZ には Fortran90 で記述された 3 つのベンチマークアプリケーション LU, SP, BT がある．全て 3 次元空間における非定常圧縮性 Navier-Stokes 方程式を ADI 法によって解くものである．計算サイズは小さい順から S, W, A, B, C, D, E, F の 8 つのクラスが用意されている．以下にそれぞれのアプリケーションの概略を示す．

LU 上下三角行列を対称 SOR 法を用いて解く．最大並列度が 16 までなのでハイブリッド並列の効果が現れにくく，本報告では用いない．

SP スカラー 5 重対角行列を ADI 法を用いて解く．

BT ブロック 3 重対角行列を ADI 法を用いて解く．メッシュの分割サイズが一様ではないので負荷分散が難しくなる．

Multi-Zone Version はオリジナルの NPB と比べて並列度の条件が緩やかである．これは並列化の手法が異なっているからである．NPB MZ では計算対象となる 3 次元直方体を x 方向と y 方向に荒く分割し，分割単位を Zone とする．時間ステップごとに各 Zone に計算を割り当て，得られた境界データを隣接する Zone 間で交換することにより並列化を行っている．このような手法の結果，オリジナルの NPB に比べてメッシュのアスペクト比は異なるが，総メッシュ数はクラス S を除いてほぼ同じである．

5.3 実行結果と考察

NPB MZ を Cray XT5 上で実行した結果を述べる．NPB MZ のバージョンは 3.2 を用い，コンパイラは Portland Group Inc. (PGI) のものを，コンパイラオプションは過度の最適化の影響を避けるため “-tp shanghai-64 -r8” とし、，ハイブリッド並列を有効にする場合は “-mp=nonuma” を付加して OpenMP の宣言を有効にした．

計算サイズはクラス D とし，ハイブリッド並列を行わない Pure MPI，1 プロセスあたり 4 つスレッドを用いる Hybrid (x4), 1 プロセスあたり 8 つのスレッドを用いる Hybrid (x8) をそれぞれ測定した．グラフの縦軸は 1 PE あたりの性能 (Mop/s/PE)，横軸は PE 数である．

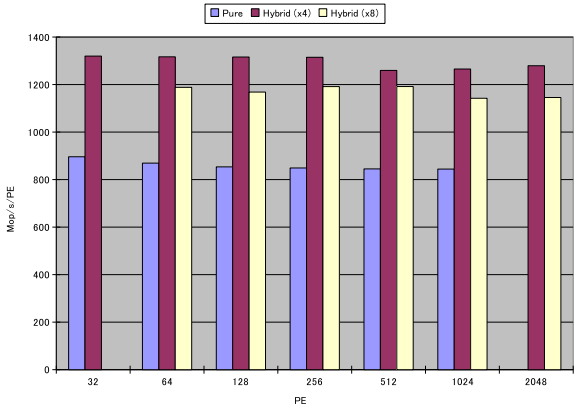

図 2 SP (CLASS D) における PE あたりの性能

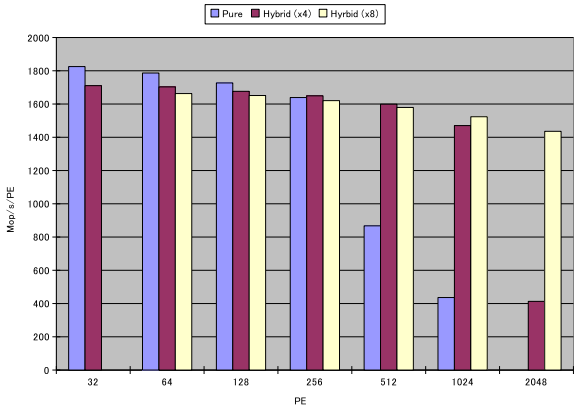

図 3 BT (CLASS D) における PE あたりの性能

アプリケーション SP のハイブリッド並列化効果を示すグラフ（図 2）においては，PE の数に関わらず，ハイブリッド並列は常にピュア並列よりも高い性能を出している．並列度が上がってもこの傾向は変わらない．

アプリケーション BT のハイブリッド並列化効果を示すグラフ（図 3）においては，全ての並列化において並列度が上がるとともに性能は落ちていく．128 並列まではピュアとハイ

Vol.2009-ARC-186 No.2
Vol.2009-HPC-123 No.2
2009/11/30

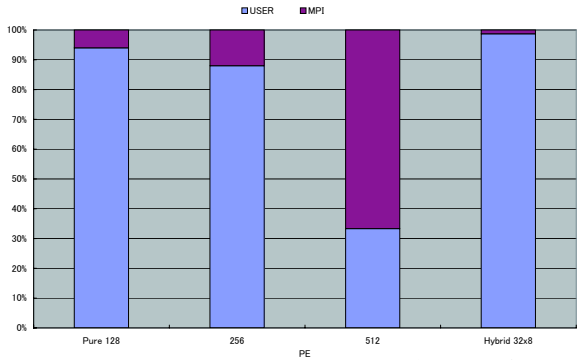

図 4 BT (CLASS D) における MPI 通信時間の割合

ブリッドの性能差はほとんど見られず，わずかにピュア並列のほうがよい．しかし，512 並列以上になるとピュア並列は大きく性能を下げるのに対し，ハイブリッド並列は 1024 並列でも高い並列化性能を保つことがわかる．

図 4 に BT の全実行時間における MPI 通信が占める時間の割合を示す．512 並列においては MPI 通信は全体の 65%にも及び，性能低下の原因となっている．

6. 並列 CG 法ソルバ

ハイブリッド並列を施した共役勾配法ソルバの性能評価として，2 次元ラプラス問題の有限要素法解析を行う．

原点を左下にもつ，辺の長さが 1.0 の正方形の領域において，ラプラス方程式を考える．

$$\triangle\phi(x,y) = 0, (x,y) \in \mathbf{R}^2$$

境界条件として，底辺に $\phi = \sin(\pi x)$ を，それ以外の境界面には 0 をそれぞれ与える．

6.1 領 域 分 割

領域を 3 角形要素に切りわけ，これをミネソタ大学で開発されているメッシュ分割ライブラリである，Metis によって領域分割する．Metis はグラフ理論に基づいた分割アルゴリズムを用いており，特に，複雑な形状の領域分割を行うのに適している．

図 6 は正方領域を 16,384 $(= 128^2)$ 節点，32,258 要素のメッシュに切り，Metis によって 8 分割して色付けしたものである．ここでは multilevel k-way partitioning scheme を利用した分割で，各領域を構成する辺が最小になるように分割が行われている．

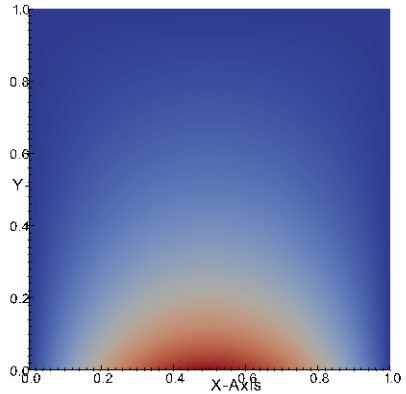

図 5 正方領域上の Laplace 問題の解析結果
Fig. 5 Laplace' Problem on square

6.2 CG 法のハイブリッド並列化

偏微分方程式を有限要素法で離散化し，線形方程式を共役勾配法で解く．有限要素法による離散化についてはここでは詳しく触れない．

CG 法のハイブリッド化にあたっては，まずループの前に OMP parallel 節を置き，do ループの前で OMP do ディレクティブを挿入する．ただし，MPI による同期や，逐次計算が必要な場所では OMP single を指定して 1 スレッドのみによる逐次計算を行った．

6.2.1 結果と考察

グラフ 7 に示すのが Hybrid 並列を行った並列 CG 法ソルバの結果である．2 次元の正方領域上に 1,048,576 節点の構造格子を取り，MPI によるハイブリッド並列ソルバと，Hybrid 並列ソルバとの経過時間を比較した．どちらも収束には同じ反復回数だけかかっている．

実行したコア数のすべてにおいて Hybrid 並列は Pure 並列に比べてわずかに性能が低い．128 コアにおいて，2870 回の反復に費やした時間は Pure 並列が 11.489s Hybr 並列が 18.740s となっている．

Vol.2009-ARC-186 No.2
Vol.2009-HPC-123 No.2
2009/11/30

図 6 Metis によるメッシュの分割
Fig. 6 Mesh decomposition by Metis

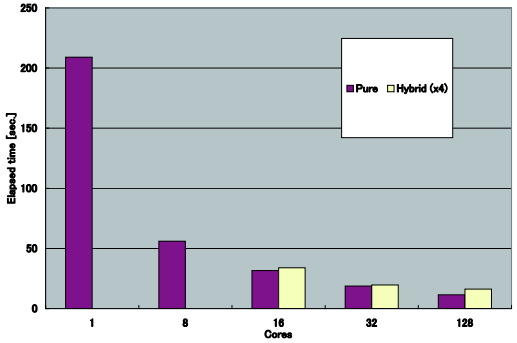

図 7 CG 法の Hybrid 並列と Pure 並列の比較
Fig. 7 Comparison of Hybridized and Pure MPI CG method

7. 結　　論

Cray XT5 上でメモリ性能と通信性能を調べ，二つの数値流体アプリケーションの性能を評価した．NPB においては Hybrid 並列アプリケーションのほうが高並列時に性能が大きく向上した．しかしながら，CG 法ソルバの Hybrid コードは性能がわずかに低かった．このことは Hybrid 並列はどのように行うかという手法によって性能が大きく異なることを意味している．

参 考 文 献

1) D. Bailey, E.Barscz, J. Barton, D.Browning, R. Carter, L. Dagum, R. Fatoohi, S. Fineberg, P. Frederikson, T. Lasinski, R. Schreiber, H. Simon, V. Venkatakrishnan, S. Weeratunga. “The NAS Parallel Benchmarks.” NAS Technical Report RNR-94-007, NASA Ames Research Center, Moffett Field, CA, 1994
2) R.F. Van Der Wijngaart, H. Jin. “NAS parallel benchmarks, multizone versions”, NAS Technical Report NAS-03-010, NASA Ames Research Center, Moffett Field, CA, 2003.